Analyse comparative de corpus oraux et écrits français: mots, lemmes et classes morpho-syntaxiques

V. Gendner[◇♣], *M. Adda-Decker*[◇]

[◇]LIMSI/CNRS
Bâtiment 508, 91403 Orsay cédex B.P. 133 FRANCE
Tél.: ++33 (0)1 69 85 80 06 - Fax: ++33 (0)1 69 85 80 88
Mél: {gendner,madda}@limsi.fr - http://www.limsi.fr/TLP/gendner
[♣] LATTICE, Université Paris VII

RÉSUMÉ

Corpora of oral and written French have been automatically tagged with lemma and morpho-syntactic information: radio/TV broadcast transcripts and *Le Monde* newspaper. For both corpora of 40M words each, we measure the corpus vocabulary in terms of lexical forms and lemmas. Morpho-syntactic information has then been examined using the most common Parts Of Speech (POS): noun, verb, adjective, adverb, pronoun, conjunction, determiner and preposition. Distributions of word occurrences and of vocabulary items have been computed as a function of POS. A comparison between oral and written French is carried out. Beyond a quantitative description of oral and written corpora, this study aims at using more linguistic knowledge in speech recognition systems.

1. INTRODUCTION

Dans ce travail nous nous intéressons au traitement et à l'analyse de grands corpus de textes écrits et de transcriptions d'oral. Plus particulièrement nous essayons d'en dégager des propriétés lexicales et morpho-syntaxiques exploitables par la suite dans la modélisation de la langue pour la reconnaissance automatique de la parole.

Les recherches menées en reconnaissance automatique de la parole ont permis d'accumuler au cours du temps de grands volumes de données textuelles, qui peuvent se compter en centaines de millions de mots. Leur provenance est diverse : il s'agit de sources écrites, comme les journaux tels que *Le Monde*, les dépêches d'agence de presse. Plus récemment de grandes quantités de données audio transcrites, en particulier des journaux radio et télédiffusés ont pu être collectées. Ces corpus permettent d'effectuer des comparaisons quantitatives et qualitatives entre langue orale et langue écrite dans un cadre thématique journalistique.

Les traitements classiquement effectués sur les corpus de texte pour la reconnaissance de la parole concernent essentiellement le découpage en mots et en phrases et des étapes de normalisation des formes lexicales observées. Parmi ces étapes les plus importantes concernent le traitement des sigles et des nombres et le traitement des majuscules en début de phrase. Ces étapes visent à réduire le nombre de formes distinctes possibles, afin de garantir une meilleure couverture lexicale et un meilleur partage de l'information contextuelle dans les modèles de langage de type N-gramme de mots.

La lemmatisation paraît un traitement tout à fait approprié pour le français, où un grand nombre de formes fléchies sont possibles pour les noms, les adjectifs et surtout les verbes. Nous pouvons citer les travaux fait chez IBM France [1], où les lemmes ont été considérés pour la reconnaissance automatique de la parole. Cependant la lemmatisation, n'est généralement pas effectuée pour la reconnaissance de la parole proprement dite, elle est introduite après une transcription automatique dans un but d'indexation automatique de corpus audio ou de recherche d'information dans de tels corpus. Or nous savons qu'à taille de vocabulaire identique, la couverture lexicale est plus faible en français qu'en anglais, notamment à cause de variations flexionnelles des verbes, où des dizaines de formes distinctes sont couramment observées.

L'étiquetage morpho-syntaxique et l'estimation de modèles de langage N-gramme de classe est exploré par de nombreuses équipes [1, 2, 3] afin d'estimer des modèles de langage N-gramme de classe. Ceci permet d'augmenter les connaissances linguistiques a priori du système de reconnaissance, notamment pour les problèmes d'accord en genre et en nombre pour les homophones hétérographes en français. Avec le même objectif nous avons effectué un étiquetage morpho-syntaxique sur une variété de corpus de français oral et écrit.

Dans cet article nous étudions la répartition des parties du discours dans de grands corpus oraux et écrits et la répartition du vocabulaire des corpus sur ces parties du discours. Ces mesures seront faites sur l'ensemble de chaque corpus et sur le sous-ensemble des mots apparaissant plus d'une fois (afin de minimiser l'impact des formes erronées). Dans un deuxième temps nous allons regarder le même type d'information, mais uniquement sur la partie du corpus couverte par un vocabulaire de systèmes de reconnaissance de taille fixe (65k mots). Ce type d'analyse permettra d'étudier des corrélations éventuelles entre certains problèmes de modélisation de la langue pour la reconnaissance et les parties du discours (POS). En particulier on pourra donner la distribution des mots hors vocabulaire (MHV) en fonction des POS.

Nous regardons ensuite la lemmatisation, qui permet de ramener les variations flexionnelles à des formes canoniques. L'effet de la lemmatisation sur la couverture 'lexicale' d'un système de reconnaissance est mesurée.

Par ces travaux comparatifs sur grands corpus, nous espérons contribuer à l'analyse descriptive du français oral et écrit, et à dégager des orientations de travaux pour la reconnaissance automatique du français parlé.

2. DESCRIPTION DES CORPUS

Dans cette étude nous avons utilisé deux corpus d'environ 40 millions de mots chacun. Les deux corpus couvrent le domaine journalistique, le premier correspondant à des journaux écrits, le deuxième à des émissions journalistiques radio- et télédiffusées transcrites.

2.1. Corpus de français écrit

Le premier corpus provient majoritairement du quotidien *Le Monde*, des années 1987 à 1996 et une proportion faible (environ 10 %) est extraite de l'hebdomadaire *Le Monde Diplomatique* (années 90-95). Afin de totaliser 40M de mots une quantité limitée de mots a été utilisée de chaque année. Un extrait est montré ci-après :

```
par exemple , s' agissant d' un sujet aussi
sensible que la maintenance des centrales
nucléaires , et donc de leur sécurité , les
syndicats découvrent avec angoisse que
celle-ci , au nom d' impératifs
gestionnaires , peut être confiée à des
sous-traitants .
```

2.2. Corpus de français oral transcrit

Le deuxième corpus de 40 millions de mots correspond à des transcriptions d'émissions journalistiques radio- et télédiffusées. Ces transcriptions ont été acquises dans le cadre du projet européen IST-1999-10354-ALERT. Les transcriptions couvrent certaines périodes entre 1997 et 2000. Voici un exemple de transcriptions :

```
globalement , notre sentiment c' est que si
les syndicats signent ça c' est la fin du
syndicalisme confédéré en France , ça veut
dire que les syndicats n' auront plus aucune
crédibilité comme défendant les intérêts des
travailleuses et des travailleurs en attente ,
ça veut dire que ça sera une trahison .
```

2.3. Normalisation des textes

Les textes et transcriptions ont été nettoyés et segmentés en "mots". La définition du mot dépend des contraintes imposées par l'étiqueteur morpho-syntaxique, le lemmatiseur et le système de reconnaissance.

Dans tous les cas nous appliquons un certain nombre de traitements, afin d'enlever des textes des balises structurantes et afin de minimiser le nombre de 'mots' correspondant à des erreurs typographiques ou de formatage. Un certain nombre d'opérations de normalisation sont toujours effectuées : séparation en phrases et une première segmentation en mots. Pour le système de reconnaissance les textes sont segmentés afin d'optimiser la couverture lexicale avec un vocabulaire système de taille fixe (65k mots) tout en définissant des mots non ambigus [6, 5]. Pour optimiser la couverture lexicale on a tendance par exemple à éclater les sigles non acronymes, et à transformer les nombres sous forme de chiffres en suites de mots. Ces normalisation ont un impact important sur la couverture lexicale [6]. Pour l'étiquetage morpho-syntaxique ces derniers choix ne sont pas adaptés. Des versions de textes intermédiaires ont donc dues être générées, gardant les nombres et sigles dans leur forme originale.

2.4. Vocabulaires

Dans la table 1 les tailles de corpus et de vocabulaires sont précisées. Trois types de vocabulaires sont considérés : le vocabulaire TOTAL, ie. le nombre de formes distinctes sur corpus; ensuite la partie du vocabulaire dont les formes apparaissent au moins 2 fois dans le corpus (NB.OCC>1); finalement le vocabulaire comprenant les 65k mots les plus fréquents appelé SYSTÈME. En comparant l'oral à l'écrit on peut constater, qu'à taille de corpus pratiquement identique, le corpus oral contient significativement moins de formes distinctes (156k) que le corpus écrit (281k). Ceci se traduit par une couverture nettement meilleure sur l'oral : 99.6%, à comparer à 98.6% pour l'écrit. L'oral apparaît donc comme une expression de langue lexicalement plus redondante que l'écrit. On peut se poser la question si le contenu plus riche du corpus écrit, avec des extraits des années 87-96, n'est pas lié à son étalement plus grand dans le temps? Nous avons vérifié ce point sur des corpus de 20M de mots, couvrant tous les deux la même époque (1997). Nous avons obtenu 127k formes distinctes pour l'oral et 215k pour l'écrit. La différence très importante en variété lexicale est donc bien due au style employé à l'écrit plutôt qu'au paramètre temps.

TAB. 1 – *Pour les deux corpus* **écrit** *et* **oral** *sont indiqués la taille du vocabulaire total, le nombre de formes distinctes d'occurrence supérieure strictement à 1. La colonne* SYSTÈME *correspond au 65k mots les plus fréquents dans le corpus. Pour ces 3 vocabulaires le nombre d'occurrences dans le corpus est indiqué dans la deuxième ligne. La troisième ligne donne la couverture lexicale sur les mêmes corpus.*

écrit			
VOCABULAIRE :	TOTAL	NB.OCC>1	SYSTÈME
taille voc.:	281k	175k	65k
taille corpus:	42.36M	42.23M	41.77M
couverture	100%	99.7%	98.6
oral			
VOCABULAIRE:	TOTAL	NB.OCC>1	SYSTÈME
taille voc:	156k	109k	65k
taille corpus:	40.80M	40.74M	40.62M
couverture	100%	99.8%	99.6

3. ÉTIQUETAGE MORPHO-SYNTAXIQUE

3.1. Choix d'étiquettes

Nous nous sommes intéressées à deux jeux d'étiquettes. Le premier comprend simplement les parties du discours (POS) principales (Nom, Verbe, Adjectif, Adverbe, Pronom, Conjonction, Déterminant, Préposition, Numéral, Interjection) plus une étiquette de ponctuation. À l'aide d'un tel jeu d'étiquettes il est facile d'étiqueter de grands corpus de manière robuste, ce qui peut être un atout pour l'estimation de modèles de langages pour la reconnaissance de la parole. En effet, le but ultime de cet étiquetage est de fournir des informations supplémentaires à un décodeur de parole. Or une analyse des erreurs de reconnaissance dans de bonnes conditions (parole lue) montre qu'une erreur sur quatre est une confusion entre deux homophones morphologiques. Il s'agit de confusions de genre (ex: *médiatisé* / *médiatisée*) ou de nombre (ex: *illicite* / *illicites*),

mais aussi de confusions de temps et de mode (ex: *encaisser* / *encaissé*, *chantez* / *chanté*). De ce fait, nous avons ajouté les informations de genre et de nombre, ainsi que la distinction de temps et de mode dans un deuxième jeu qui comprend ainsi 90 étiquettes (Verbe: 40 étiquettes; Nom: 9 ; Pron: 9 ; Dét: 9 ; Adj: 9 ; Adv: 1 ; Conj: 1 ; Prép: 1 ; Num: 5 ; Ponct: 5 ;Interj: 1).

3.2. *Choix du tagger*

En nous référant aux conclusions de Valli&Veronis [4] qui indiquent que la langue orale peut être facilement étiquetée avec des outils conçus pour l'écrit, nous avons testé différents taggers de ce type sur une partie représentative du corpus. Nos expériences confirment leurs conclusions. Les différents taggers produisant des résultats comparables, nous avons choisi pour le traitement de l'ensemble des données, l'analyseur inclus dans le correcteur orthographique Cordial qui nous a semblé particulièrement robuste, rapide et disponible facilement. Un post-traitement basé sur le tagger de Brill (version entraînée pour le français à l'INaLF lors de la campagne d'évaluation GRACE) ajoute l'information de nombre lorsque l'analyseur de Cordial produit une sortie sous-spécifiée (env. 4% d'étiquettes corrigées).

3.3. *Résultats d'étiquetage*

Dans la figure 1 nous montrons les proportions des 8 principales POS dans les corpus. Pour l'oral et l'écrit nous observons des distributions relativement similaires : les 4 POS les plus fréquentes sont : NOM, DET, PREP, VERB avec chacune plus de 10%. ADJ, ADV et CONJ restent nettement en-dessous de 10%. Pour la catégorie PRON l'écrit est également nettement en-dessous des 10% alors que l'oral les dépasse clairement.

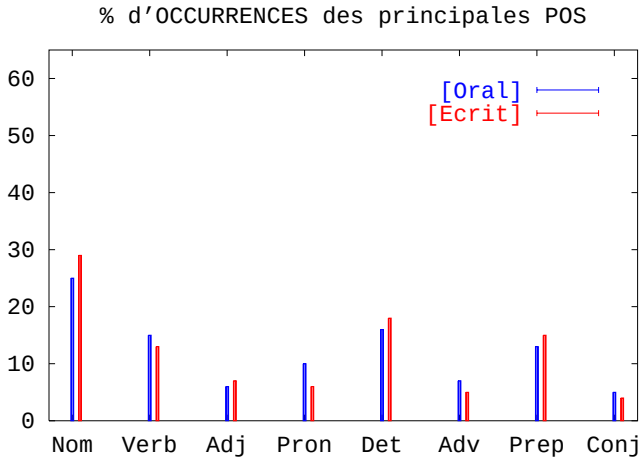

FIG. 1 – *Proportions dans les corpus des 8 POS principales (en excluant les ponctuations, les interjections et les numéraux des corpus taggés).*
oral: à gauche; écrit : à droite.

Les 3 étiquettes NOM, VERB, ADJ représentent environ 45% du corpus oral, 51% du corpus écrit. Si on regarde maintenant la répartition des POS dans le vocabulaire dans la figure 2, on peut remarquer que ces 3 POS totalisent la quasi-totalité des entrées lexicales. Les répartitions des POS avec les deux vocabulaires NB.OCC>1 et SYSTÈME

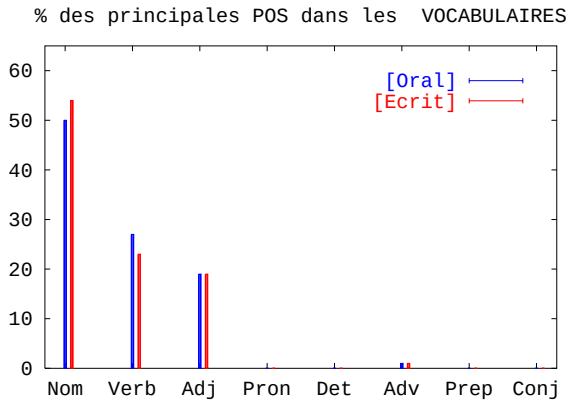

FIG. 2 – *Proportions dans le vocabulaire* TOTAL *des 8 POS principales. Les noms représentent au minimum 50 % des entrées. La proportion des verbes est plus faible pour le corpus écrit, pourtant, les chiffres absolus sont très semblables. La faible proportion de verbes est probablement due au nombre comparativement élevé de noms dans le vocabulaire des textes écrits (32000 vs 19000).*
oral: à gauche; écrit : à droite.

seront rajoutés dans cette figure (dans la version définitive). Les NOMS à eux seuls représentent plus de la moitié du vocabulaire TOTAL. La figure 3 compare la répartition des noms triés par rang de fréquence décroissant entre oral et écrit. On peut voir que les noms les plus fréquents sont davantage répétés à l'oral, et que par ailleurs il y a moins de noms rares. La distribution pour le langage écrit est plus étalée. Ceci rejoint l'observation faite de manière globale sur la taille des vocabulaires plus haut.

Pour la couverture lexicale les problèmes proviennent donc essentiellement des trois POS :NOM, VERB, ADJ. Ce sont ces trois catégories qui admettent le plus de variations dues aux flexions. L'impact de la lemmatisation sur le vocabulaire est donc intéressant à mesurer.

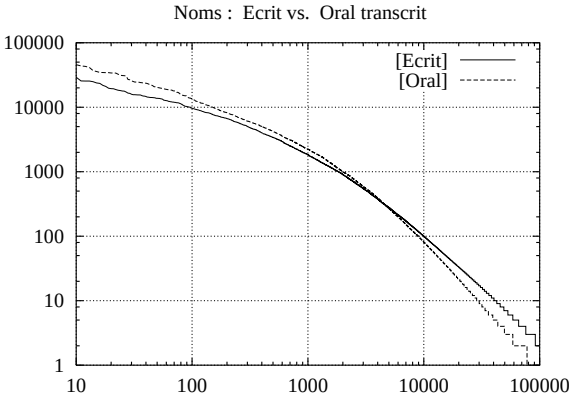

FIG. 3 – *Occurrences des mots de la catégorie Nom par rang de fréquence décroissant. Comparaison du corpus écrit versus corpus oral transcrit)*

4. LEMMATISATION

Lors de l'étiquetage morpho-syntaxique chaque mot a été associé à son lemme correspondant; un exemple est donné dans la table 2. Dans la table 3 nous donnons le nombre de

TAB. 2 – *Exemple de texte lemmatisé et taggé. Le jeu d'étiquettes utilisé ici inclut l'information de genre (m|p) et nombre (s|p, en finale), mode et temps pour les verbes.*

forme fléchie	lemme	POS
notre	notre	D.s
sentiment	sentiment	Nms
c'	ce	P..
est	être	Vmip-s
que	que	C–
si	si	C–
les	le	D.p
syndicats	syndicat	Nmp
signent	signer	Vmip-p
ça	ça	P.s

TAB. 3 – *Pour les catégories* NOM, VERB, *et* ADJ *nous indiquons le nombre de formes fléchies distinctes et le nombre de lemmes distincts ayant chacun un nombre d'occurrences à 1 strictement.*

	NOM	VERB	ADJ
	écrit		
formes fléchies	123k	36.5k	25.5k
lemmes	109k	5.5k	12.7k
%réduction	11%	85%	50%
	oral		
formes fléchies	66.6k	21.6k	14.7k
lemmes	58.8k	3.9k	7.8k
%réduction	12%	82%	47%

formes fléchies (mots) et le nombre de lemmes distincts observés plus d'une fois dans le corpus, pour les deux corpus écrit et oral. On peut remarquer que la lemmatisation ne permet de réduire que d'un peu plus de 10% le nombre d'entrées de catégorie NOM, pourtant la plus prolifique dans le vocabulaire. La lemmatisation donne une bonne réduction d'environ 50% pour les adjectifs. Sans surprise le taux le plus important a été mesuré pour les verbes : environ 85% où on observe environ 5.5k lemmes. La figure 4 illustre l'effet de la lemmatisation pour les verbes sur le corpus écrit. Le corpus oral montre une évolution tout à fait similaire. Les taux de réductions sont comparables entre l'oral et l'écrit.

5. CONCLUSIONS ET PERSPECTIVES

Cette étude a montré des différences de vocabulaire et de sa répartition sur les POS entre corpus oraux et écrits dans le domaine journalistique. Les transcriptions diffèrent des journaux écrits par un taux plus important de pronoms (10% vs 5%) et un taux moins important de noms (25% vs 30%). Les corpus écrits contiennent une variété lexicale significativement plus élevée que les corpus oraux. À taille de corpus égale, l'écrit produit presque le double de formes différentes. Les POS problématiques pour la couverture lexicale se réduisent à NOM, VERB et ADJ. La lemmatisation permet d'aboutir à une excellente couverture pour VERB et ADJ. Pour les noms, de loin la catégorie la plus riche, la lemmatisation ne permet de réduire le nombre d'entrées que de 10%. Le travail en cours inclut l'analyse des bigrammes et des trigrammes de POS ainsi que l'entraînement de modèles de langage sur de plus grandes quantités de données taggées (300 M de mots).

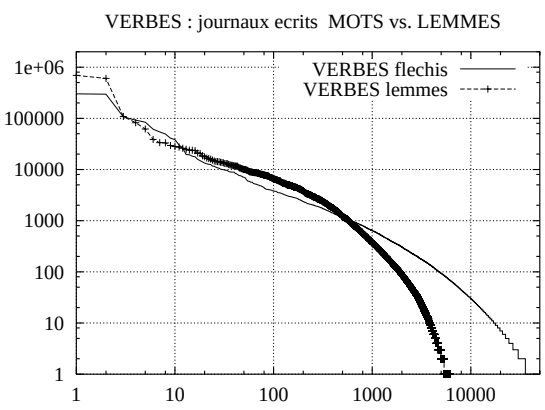

FIG. 4 – *Journaux écrits - Nombre d'occurrences des verbes en fonction de leur rang de fréquence décroissant. Comparaison formes fléchies vs. lemmes (40M mots).*

6. REMERCIEMENTS

Une partie de ce travail a été effectuée dans le cadre du projet européen IST-1999-10354-ALERT. Nous remercions toute l'équipe TLP qui participe au travail de ce projet, et tout particulièrement Gilles Adda pour la mise en place et la normalisation des corpus, mais aussi pour les discussions et critiques pertinentes.

RÉFÉRENCES

[1] M. ElBèze, *Choix d'unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance automatique de la parole*, thèse de doctorat Paris VII, 1990.

[2] F. Béchet *et al.*, *Large Span Statistical Language Models: Application to Homophone Disambiguation for Large Vocabulary Speech Recognition in French*, Eurospeech'99, Volume 4, Page 1763-1766, Budapest, septembre 1999.

[3] I. Zitouni *et al.*, *A Comparative Study Between Polyclass and Multiclass Models*, ICSLP Sydney Australia, novembre 1998,.

[4] André Valli, Jean Veronis. *Etiquetage grammatical des corpus de parole: problèmes et perspectives*, Revue Française de Linguistique Appliquée 1999.

[5] Martine Adda-Decker et al., *Large Vocabulary Speech Recognition in French*, ICASSP'99, Phoenix.

[6] G. Adda ıet al., *Text Normalization and Speech Recognition in French*, EuroSpeech'97.