Profilage de textes : cadre de travail et expérience

B. Habert, G. Illouz, P. Lafon, S. Fleury, H. Folch, S. Heiden, S. Prévost
LIMSI & UMR 8503

{habert,illouz}@limsi.fr,
{lafon,fleury,folch,heiden,prevost}@ens-fcl.fr

Abstract

The increasing use of "huge corpora" in natural language processing and text analysis implies that lexical, morphosyntactic and syntactic homogeneity be mastered. This requires the development of text profiling tools. We have developed such tools and a related methodology within the ELRA benchmark called "Contribution to the construction of contemporary french corpora". We show the first results of this approach as applied to the speeches of De Gaulle and Mitterrand on radio and television. We present our conclusions on this experience in particular on the relevance of the features we use for text profiling.

Résumé

Le recours croissant aux « très grands corpus » en Traitement Automatique des Langues (TAL) comme en analyse textuelle suppose de maîtriser l'homogénéité lexicale, morpho-syntaxique et syntaxique des données utilisées. Cela implique en amont le développement d'outils de profilage de textes. Nous mettons en place de tels outils et la méthodologie associée dans le cadre de l'appel d'offres ELRA *Contribution à la réalisation de corpus du français contemporain*. Nous montrons sur les discours radio-télévisés de De Gaulle et de Mitterrand les premiers résultats de cette approche. Nous tirons les conséquences de cette expérience pour les traits que nous employons pour profiler les textes. .

Keywords : Typologie de textes, genres textuels, corpus annotés, linguistique de corpus

1. Profiler les textes : enjeux

Nous appelons *profilage de textes* un bilan quantitatif fondé sur des indices linguistiques d'emploi du vocabulaire, mais aussi de catégories (morpho-syntaxiques, syntaxiques, sémantiques, structurelles) et de patrons morpho-syntaxiques, etc., dans les parties d'un corpus, pour regrouper ensuite ces parties en sous-ensembles homogènes sur ces points. Ce bilan doit également permettre de positionner un nouveau texte par rapport aux regroupements déjà obtenus.

Pouvoir profiler des textes représente un enjeu important à la fois pour l'analyse textuelle et pour le TAL (Traitement Automatique des Langues). Pour l'analyse textuelle, il s'agit de connaître les proximités d'un texte ou d'un corpus, en terme de « facture linguistique », au sens large, avec d'autres textes ou d'autres corpus, pour pouvoir étendre la portée des constats effectués sur ce texte et ce corpus. Pour le TAL, alors que les données textuelles disponibles pour l'acquisition de connaissances lexicales, syntaxiques et sémantiques ont atteint des proportions volumineuses (comme les 100 millions de mots étiquetés du BNC – British National Corpus), elles rassemblent parfois des composants extrêmement hétérogènes. Il en va ainsi des données de presse, comme les CD-ROM du *Monde*, qui sont souvent mises à contribution vu leur accessibilité. (Illouz et al., 1999) montre par exemple des écarts importants entre les principales sections du *Monde*

(politique, économie, étranger, arts et spectacles, information générale, éducation / médecine / société) en ce qui concerne les catégories morpho-syntaxiques et le lexique utilisés.

Plusieurs études convergent pour rendre plausible l'hypothèse selon laquelle la fiabilité des traitements automatiques dépendrait de l'homogénéité des données en cause.

Étiquetage (Illouz, 1999) met en évidence la corrélation – lors de l'action d'évaluation d'étiqueteurs morpho-syntaxiques GRACE (Adda et al., 1999) – entre la plus ou moins grande précision des étiqueteurs et la nature des textes à catégoriser (mémoires, romans, ou essais extraits de la base Frantext de l'INaLF et fragments du *Monde*).

Parsage (Sekine, 1998) utilise 8 domaines du corpus Brown. Il examine les performances d'un analyseur syntaxique probabiliste selon que l'apprentissage de la grammaire s'effectue sur le même domaine que celui du test, sur tous les domaines confondus, sur la partie *fiction* (fiction, western, romans sentimentaux) ou sur la partie *non-fiction* (reportages, éditoriaux, passe-temps, textes érudits). Les performances vont dans l'ordre décroissant suivant : identité domaine d'apprentissage/de test, appartenance des domaines d'apprentissage/de test à la même « classe », apprentissage sur un corpus relevant de tous les domaines à la fois. Entraîner l'analyseur sur une classe (*fiction* par exemple) et l'utiliser sur l'autre classe (*non-fiction*) donne les résultats les plus mauvais.

Recherche d'information (Karlgren, 1999) utilise la portion du *Wall Street Journal* provenant du corpus TIPSTER [1] et les requêtes d'interrogation 202 à 300 de la campagne d'évaluation TREC (*Text Retrieval Conference*) assorties des jugements de pertinence sur les 74 516 articles en question [2], c'est-à-dire de l'indication que l'article est ou non une réponse correcte à une requête donnée. Elle mesure un certain nombre de caractéristiques stylistiques de chaque article : longueur moyenne des mots, proportion de mots longs, fréquence moyenne des mots, fréquence moyenne de mots capitalisés, proportion de nombres, pronoms personnels… À cette aune, il s'avère que les textes jugés pertinents diffèrent significativement des textes jugés non pertinents, et surtout que les textes les plus fréquemment retenus par les systèmes en compétition à TREC (qu'ils soient pertinents ou non) s'écartent également significativement des textes pour lesquels il n'y a pas de jugement de pertinence.

2. Une démarche typologique inductive

Dès lors que la fiabilité des traitements dépend de la nature des textes traités, il importe de savoir classer ces derniers. L'optique, inductive, dans laquelle nous nous inscrivons consiste à faire émerger *a posteriori* les types de textes – considérés comme des agglomérats fonctionnellement cohérents de traits linguistiques – grâce à un traitement statistique multidimensionnel de textes annotés. Cette optique constitue la ligne directrice des travaux de D. Biber (Biber, 1988; Biber, 1995). Biber examine les cooccurrences entre 67 traits linguistiques dans les 1 000 premiers mots de 481 textes d'anglais contemporain écrit et oral. Ces textes proviennent de LOB et London-Lund, complétés par des lettres personnelles et professionnelles et relèvent d'une quinzaine de « genres » divers : articles de recherche, reportages, conversations, nouvelles ra-

[1] http ://www.tipster.org

[2] Ces articles proviennent des années 1990 à 1992. 2 039 sont pertinents pour au moins une requête. 35 289 ne sont pertinents pour aucune requête. Il reste 37 188 articles non jugés.

diophoniques... Les traits étudiés ressortissent à 16 catégories distinctes (marqueurs de temps et d'aspect, questions, passifs, modaux...). Ils sont identifiés automatiquement. La statistique multidimensionnelle permet d'obtenir des pôles multiples, positifs et négatifs, correspondant à des constellations de traits linguistiques corrélés. Ces pôles constituent deux à deux des dimensions textuelles. Chaque texte, par son emploi des traits linguistiques retenus, se situe en un point déterminé de l'espace à n dimensions issu de l'analyse. Les techniques de classification automatique permettent de regrouper les textes en fonction de leurs coordonnées sur ces dimensions. Les types de textes qui en résultent ne recoupent directement ni les « genres » des données de départ ni les registres intuitivement distingués.

Nous mettons en place, dans le cadre du projet `TyPTex` (*Typage et Profilage de Textes*) commun au LIMSI et à l'UMR 8503 et soutenu financièrement par ELRA (*European Language Resources Association*) dans le cadre de l'appel d'offres *Contribution à la réalisation de corpus du français contemporain*, une méthodologie permettant de tester et d'étendre les propositions de Biber, en utilisant en particulier les acquis pour l'analyse du français de (Sueur, 1982) et de (Bronckart et al., 1985).

3. Architecture de profilage

Comme le montre la figure 1, p. 4, on dispose au départ d'une base de textes. Chacun comprend un en-tête documentaire ou « cartouche » (*header*) suivant les recommandations de la TEI (Dunlop, 1995). Les critères d'une requête ou d'une sélection aboutissent à un corpus, c'est-à-dire un ensemble de textes rassemblés en fonction d'une hypothèse déterminée. Chacun de ces textes est soumis à un étiquetage morpho-syntaxique, qui permet d'associer à chaque mot ou unité polylexicale un lemme, une partie du discours et des indications morpho-syntaxiques plus fines. Le marquage typologique se fonde sur l'ensemble de ces informations et opère un transfert (par regroupements, dégroupements, transformations, complémentations ou même omissions), vers de nouvelles catégories correspondant aux traits linguistiques dont on veut étudier la distribution. Le corpus marqué (et éventuellement corrigé par le biais de `CorTecs` [3] (Heiden et al., 1998)) est alors soumis à des logiciels de comptage. En particulier, on construit la matrice des fréquences de chaque trait dans chaque texte. Cette matrice sert tant à la recherche optimale de traits pertinents à une opposition, qu'à la classification inductive ou supervisée [4].

Nous avons utilisé pour l'étiquetage de départ `Sylex-Base` (Ingenia, 1995), étiqueteur/analyseur basé sur le travail de P. Constant (Constant, 1991). La grille des traits retenus et comptés pour le marquage typologique et issue de cet étiquetage comporte 229 éléments. Elle a un caractère expérimental mais elle ne saurait de toute façon prétendre ni à l'exhaustivité, ni à l'universalité, ni même à l'homogénéïté [5]. Elle contient ici les catégories traditionnelles de la grammaire (`nom, adjectif, adverbe, préposition`, etc.), des combinaisons de catégories, de sous-catégories et de flexions (adverbe négatif, déterminant possessif première personne singu-

[3] Il s'agit d'un programme sous `Unix` d'aide à la correction de textes catégorisés. Chaque mot peut porter plusieurs étiquettes (dont la nature est libre : lemme, catégorie morpho-syntaxique, catégorie sémantique...). Le travail par concordances permet de regrouper des contextes similaires et de propager aisément des corrections. On obtient une plus grande cohérence de correction. Une version de `CorTecs` sous `Windows` est en préparation.

[4] Les indications typologiques obtenues seront réintroduites dans les cartouches des textes et fourniront ainsi de nouveaux critères de sélection.

[5] Tandis que l'étiquetage morpho-syntaxique est relativement standardisé, donnant lieu d'ailleurs à des comparaisons et à des évaluations, les traits retenus par le marquage typologique peuvent varier en fonction des textes soumis à l'analyse et des hypothèses typologiques faites sur ces textes.

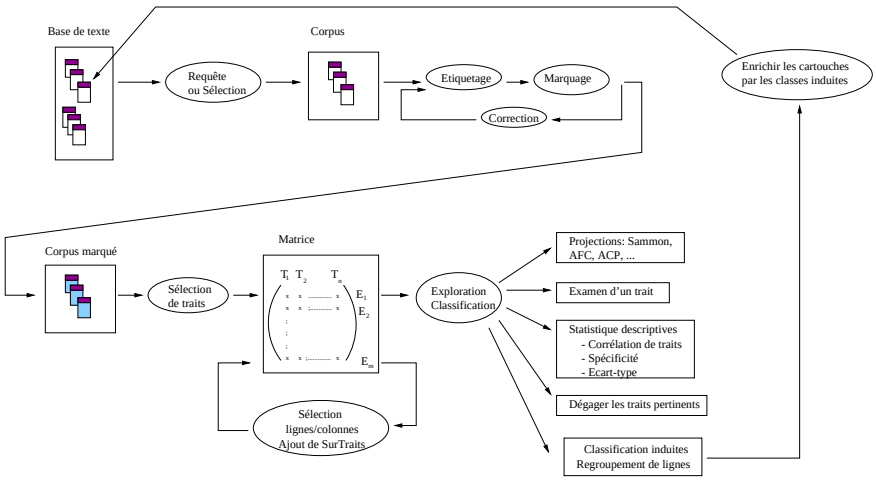

FIG. 1: Architecture de profilage de textes.

lier, article défini, etc.), des marques de temps (présent, futur, passé simple, participe présent, etc.), des modes (indicatif, infinitif, conditionnel, etc.), quelques outils textuels (présentatifs : *il y a*, *c'est*, expression de la nécessité : *il faut*, etc.), ainsi que des marques de l'écrit (virgule, point virgule, guillemet, etc.). Tous les comptages sont exclusifs, c'est-à-dire que la fréquence du trait `detpos` (déterminant possessif) exclut celle de `detpos1p` (déterminant possessif première personne du pluriel). La plupart du temps, une occurrence du texte correspond à une occurrence de trait. Parfois à plusieurs : une forme verbale peut incrémenter plusieurs traits, par exemple : verbe, indicatif et présent. La grille comporte quelques traits résiduels associés à des ambigüités non levées, par exemple `nom|verbe` ou `indicatif présent|impératif`[6].

4. Façons de dire : De Gaulle et François Mitterrand

Nous avons appliqué cette démarche à un corpus[7] d'interventions radio-télévisées faites par De Gaulle et Mitterrand pendant leur mandat présidentiel.

L'étude des spécificités de la distribution des traits linguistiques est faite à travers le regroupement par année des interventions télévisées de De Gaulle (12 ans de 1958 à 1969) et de Mitterrand (8 ans de 1981 à 1988). Le tableau ainsi soumis à l'analyse a pour dimension $229 * 20$.

Si l'on s'en tient à une démarche *a priori*, comment caractériser notre corpus ? Les interventions réunies ressortissent à des genres divers (conférence de presse, interview, discours à la nation, vœux de nouvel an, etc.), mais la diversité des genres se trouve ici neutralisée par le regroupement annuel que nous avons opéré. D'autre part, nos textes présentent un haut degré d'homogénéïté. Ils ont des conditions de production tout à fait semblables (même domaine de discours, émetteur occupant la même fonction, réception identique, même canal). De plus, en étudiant les contrastes de répartition des traits seuls à l'exclusion des signifiants du lexique,

[6]Ces ambiguïtés auraient pu être éliminées à l'aide de CORTECS, mais ce programme n'a pas été mis en œuvre ici.

[7]Il nous a été fourni par Dominique Labbé (CERAT), déjà catégorisé et lemmatisé. Nous n'avons pu utiliser cette annotation fiable (vérifiée manuellement). Notre travail de typologie porte sur des corpus diversifiés et très vastes (20 millions de mots), c'est pourquoi nous avons dû recourir à un étiqueteur, `Sylex-Base`, certes moins précis, mais adapté aux textes « tout venant ».

nous visons à éliminer le contexte situationnel et historique des textes (la variation diachronique), et la majeure partie de la thématique, précisément celle qui se manifeste dans le lexique. Malgré cette épuration discursive drastique (neutralisation des genres, élimination de la majeure partie de la thématique, de la variation diachronique et du contexte historique), des contrastes frappants subsistent dans le mode d'expression des deux présidents.

Nous nous limitons ici à la répartition des cinquante traits les plus fréquents. Le résultat du calcul des spécificités (Lafon, 1980) montre très clairement que beaucoup de traits opposent les deux émetteurs en ce qu'ils sont dominants (sur-employé ou banal) chez l'un, et récessifs (sous-employé ou banal) chez l'autre, ou l'inverse évidemment. Dans le tableau fourni en annexe, `D` ou `M` en colonne, suivis du millésime, renvoient respectivement à De Gaulle et à Mitterrand. Un plus (`+`) renvoie à un sur-emploi, un moins (`-`) à un sous-emploi, le vide à un emploi banal. Le nombre qui suit `+` ou `-` indique l'ordre de grandeur de la probabilité de ce sur- ou sous-emploi. Plus il est élevé, plus le phénomène est significatif.

Traits dominants chez Mitterrand et récessifs chez De Gaulle adverbe négatif, pronom personnel première personne singulier, indicatif présent, article indéfini, passé composé, tiret, pronom démonstratif, *c'est* (à l'indicatif présent), deux points, pronom personnel, nombre cardinal, point d'interrogation, *il y a* (à l'indicatif présent), *il faut* (à l'indicatif présent).

Traits dominants chez De Gaulle et récessifs chez Mitterrand virgule, coordonnant, pronom personnel, déterminant possessif, déterminant possessif première personne pluriel, participe passé, participe présent, subjonctif présent.

On remarque également que des exceptions nombreuses aux régularités précédentes se situent toutes soit dans les dernières interventions de De Gaulle (1969), soit dans les premières interventions de Mitterrand (1981, 1982)[8]. Si donc on exclut ces trois années qui apparaissent singulières, d'autres régularités importantes viennent s'ajouter aux précédentes.

Traits complémentaires dominants chez Mitterrand et récessifs chez De Gaulle pronom personnel *on*, *pouvoir* à l'indicatif présent.

Traits complémentaires dominants chez De Gaulle et récessifs chez Mitterrand nom, préposition, article défini, adjectif, pronom personnel première personne pluriel, conditionnel.

On peut esquisser quelques convergences, qui mettent à jour des caractéristiques discursives opposant les deux présidents. Le discours gaullien s'exprime en longues périodes, fortement articulées, comme en témoigne la conjonction de `virgule` avec `coordonnant`, tandis que la phrase mitterrandienne apparaît plus hachée avec une forte présence conjuguée du tiret d'incise et de deux points. Il est encore plus intéressant de remarquer la présence plus accentuée de syntagmes nominaux définis chez De Gaulle, comme l'indique le sur-emploi conjoint de `nom`, `préposition`, `adjectif`, `article défini` et `déterminant possessif`, face à un sur-emploi par Mitterrand des présentatifs *c'est*, *il y a*, et aussi de l'expression de la nécessité *il faut*. On relève davantage de participes présent et passé chez de Gaulle, mais aussi de subjonctifs et de conditionnels, tandis que Mitterrand utilise de préférence le présent. Ces caractéristiques

[8] Toutes sortes d'interprétations politiques pourraient être avancées pour expliquer ces exceptions de début et de fin de règne. Mais ce n'est pas ce qui importe ici pour nous.

découlent probablement de ce que Mitterrand intervient plus massivement à la télévision sur le mode de la conversation avec des journalistes, tandis que De Gaulle intervient seul avec un texte préparé. Enfin, l'usage des embrayeurs de discours différencie fortement les deux présidents : Mitterrand, comme l'avait déjà remarqué Dominique Labbé (Labbé, 1990), se sert (abuse ?) du *je*, mais aussi du *vous* et du *on*, tandis que De Gaulle préfère le *nous* et les déterminants possessifs.

Nous avons ici seulement vérifié l'opposition entre deux émetteurs comparables. Mais il nous semble possible d'élaborer des expériences de corpus qui permettent d'envisager une typologie quantifiée des textes. C'est-à-dire postuler qu'il existe une variable (qui peut avoir deux modalités ou plus) définissant plusieurs types ou genres de textes par rapport à laquelle on pourra situer chaque fragment du corpus. Cette variable résulte d'un ensemble d'indices linguistiques en même temps qu'elle les régit. Le marquage typologique et la classification ont pour fonction d'organiser et de faire émerger les sous-ensembles convergents d'indices linguistiques.

5. Perspectives : manipuler des traits « à géométrie variable »

Cette expérience de marquage typologique permet de revenir de manière critique sur les traits utilisés. Ils peuvent d'abord être trop « fins » et déboucher sur un éparpillement d'occurrences rendant impalpables les contrastes. C'est le cas dans la grille utilisée actuellement pour les temps des verbes : la catégorie verbale est « éclatée » en une cinquantaine de traits, dont la plupart totalisent un nombre limité d'occurrences. On ne dispose ainsi d'aucune prise sur le verbe dans son ensemble ni sur la manière dont cette catégorie est sollicitée. À l'inverse, certains traits sont trop grossiers et cachent probablement des oppositions effectives. Il en va ainsi de *nombres cardinaux* qui regroupe les indications de quantité, mais aussi les dates, que l'on gagnerait probablement à distinguer. On souhaiterait en fait manipuler des traits structurées de manière à pouvoir utiliser tout ou partie des informations correspondantes [9]. Ainsi, disposer de l'étiquette {catégorie=nom, type=commun, genre=masculin, nombre=singulier…} permet de garder des sous-ensembles comme {catégorie=nom}, {catégorie=nom, type=commun}, voire {genre=masculin} [10]. Il faut donc pouvoir regrouper des traits pour un contraste, en éclater d'autres, voire recommencer sur certains points l'étiquetage et le marquage.

Références

Adda G., Mariani J., Paroubek P., and Lecomte J. (1999). Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français. In Amsili P. editor, *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pages 15–24, Cargèse. ATALA.

Biber D. (1988). *Variation accross speech and writing*. Cambridge University Press, Cambridge.

Biber D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge University Press, Cambridge.

Bronckart J.-P., Bain D., Schneuwly B., Davaud C., and Pasquier A. (1985). *Le fonctionnement*

[9] C'est l'approche de (Habert and Salem, 1995).

[10] Utiliser des structures de traits du type de celles employées dans les grammaires d'unification permettrait de modéliser plus strictement les informations issues du marquage, dans l'esprit par exemple de (Gazdar et al., 1990) ainsi que les opérations dont elles sont passibles.

des discours : un modèle psychologique et une méthode d'analyse. Delachaux & Niestlé, Lausanne.

Constant P. (1991). *Analyse syntaxique par couches*. Doctorat de l'ENST, École Nationale Supérieure des Télécommunications, Paris.

Dunlop D. (1995). Practical considerations in the use of TEI headers in large corpora. *Computers and the Humanities*, (29) :85–98. Text Encoding Initiative. Background and Context, edited by Nancy Ide and Jean Véronis.

Gazdar G., Pullum G. K., Carpenter R., Klein E., Hukari T. E., and Levine R. D. (1990). Les structures de catégories. In Miller P. and Torris T. editors, *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Langue, raisonnement, calcul, chapter 6, pages 245–301. Hermès, Paris.

Habert B. and Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *TAL*, 36(1–2) :249–276. Traitements probabilistes et corpus, Benoît Habert (resp.).

Heiden S., Cuq A., Ducout D., Horlaville P., Robert J.-P., Prieur V., and Dohm B. (1998). *CorTeCs – 1.0β : Manuel de l'utilisateur*. Laboratoire de Lexicométrie et Textes Politiques – UMR 9952, CNRS – ENS Fontenay/Saint-Cloud.

Illouz G. (1999). Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In Amsili P. editor, *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pages 15–24, Cargèse. ATALA.

Illouz G., Habert B., Fleury S., Heiden S., and Lafon P. (1999). Maîtriser les déluges de données hétérogènes. In Condamines A., Fabre C., and Péry-Woodley M.-P. editors, *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, pages 37–46, Cargèse.

Ingenia (1995). *Manuel de développement Sylex-Base*. Ingenia – Langage naturel, Paris. 1.5.D.

Karlgren J. (1999). Stylistic experiments in information retrieval. In Strzalkowski T. editor, *Natural language information retrieval*, Text, speech and language technology, chapter 6, pages 147–166. Kluwer, Dordrecht.

Labbé D. (1990). *Le vocabulaire de François Mitterrand*. Presses de la Fondation Nationale des Sciences Politiques, Paris.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, (1) :128–165. Presses de la Fondation Nationale des Sciences Politiques.

Sekine S. (1998). The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington. Association for Computational Linguistics.

Sueur J.-P. (1982). Pour une grammaire du discours : élaboration d'une méthode ; exemples d'application. *MOTS*, (5) :145–185.

Étiquette	F	D58	D59	D60	D61	D62	D63	D64	D65	D66	D67	D68	D69	M81	M82	M83	M84	M85	M86	M87	M88
Nombre card	2213			-03	-03		-05						-04	+02		+06	+07	+04			
Point	14123	+14	+03	+03	+02						-03			-16			-05	+02			+03
Point virg	1510	-10		-15	-04	-04	-06			-03	-04	-02	+02		+03	-11	+04	+03	+07	+23	
Point interr	1161				-05	-03	-04	-07			-02	-03		+02		+02	+04	+09			
Point excl	1104				-02	+06		-02	+07							-03	-03				-04
Deux points	2893	-17	-06	-25	-23	-20	-28	-18	-24	-23	-20	-17	-10	+20	+06	+09	+16	+21	+13	+24	
Guillemets	1093	-03		-05	-07	+04	-03					-02		+02	-04			+07			+02
Virgule	40233			+08		+04	+06	+11	+19	+02	+24	+10	+04	-04	-20	-13		-06	-11	-09	-03
Tiret	4824	-35	-13	-51	-35	-32	-36	-44	-49	-22	-45	-27	-19	+23	+04	+38	+41	+34	+21	+26	
Nom commun	75251	+05	+04	+03				+04			+03	+07	+10	+03	+02	-04	-06	-11	-03	-05	-03
Nom commun nominalisation	7615	-02			+02		+06	+04		+05	+03		+02	+05	+02			-08	-02	-13	-03
Nom propre	9416		+02	+04	+02		+02					-24	-03	-11	+03		+15	-04	-03	+02	+03
Adjectif	28542		+04	+04	+03	+06	+09	+11	+05	+03	+08	+04	+06		+02	-02	-06	-18	-11	-16	-04
Préposition	62640	+04		+07	+05		+07	+09	+04	+11	+08	+08	+06	+03	+05		-03	-45	-15	-22	-04
Sub Que	12129				+03								-03	-03	-03	+02					
Sub conditionnel	1371	+04						-03	-02									+02			
Sub relatif	8150				+03													+03	-02	-02	
Sub autre	4158	-02	+02						+04				+03			-02					
Conj coord	10980		+05	+13	+09	+10	+13	+06	+08	+07	+09	+13	+03		-03	-10	-06	-33	-11	-16	
Conj autre	5791	-03		-02	-03	-03	-02		-04	-03	-02	-03				+02	+02	+08	+06	+03	
Adv négation	12675	-09	-05	-20	-02	-16	-09	-17	-13	-14	-22	-20	-22		+03	+03		+39	+51	+38	
Adv degré	3906			-06	-03	-05			-07			-03	-03	-03	-03	+02		+22	+05	+04	+02
Adv temps	1368									+02				-02			-02				+04
Adv embrayeur	1281	+02				+04		-03	+05			-03								-02	
Adv autre	19545	-04	-03	-03			+04	+02	+03		+03			-04	-04			+03			+03
Art déf	37206	+05	+04	+09	+05	+05	+02	+03	+04	+02	+03	+03	+05	+03	+02	-04	-03	-23	-05	-12	-03
Art indéf	7627	-06		-02		-03	-02		-02	-03			-04			+03	+03	+05		+03	
Det pos 1P	1965	+02			+09	+11	+08	+05	+13	+10	+07	+06	+06	-03		-03	-09	-31	-13	-14	-03
Dét poss autre	3038	+03	+02	+10	+11	+05	+09	+07	+07	+13	+15	+05	+04	-04	-06	-06	-19	-16	-13	-13	
Dét dém	2611		-02			-03	+02	-02	-06		-03					+07					
Pro pers 1S	10322	-27	-13	-41	-36	-26	-48	-51	-51	-39	-51	-20	-22	+15	+12	+08	+16	+51	+51	+51	+09
Pro pers 1P	3585		+02		+02	+04			+11	+04			-04	+02	+04	+12		-07	-12	-25	-02
Pro pers 2P	2305	-03	-05	-09	-17	-02	-13	-14		-05	-14	-09	-09			+04	+05	+36		+13	
Pro pers On	3421	-04	-08	-07	-04	-08	-07	-04	-05	-07	-09	-09	-03		-02	+06	+08	+26	+05	+06	+04
Pro pers autre	6215	+05		+03	+03		+04	+02	+05	+03	+03			-08	-05	-03		-02		-02	-03
Pro dém	4110	-02		-03	-02	-02		-02		-02	-02					+02		+04		+04	+02
Pro autre	6259	+05	-02		+02	+05								-03	-03	-05	-02		+03	+03	
Ind prés	9031	-05	-03	-06		-04	-02	-07		-02	-10	-02	-05			+07		+18	+03	+05	
Ind prés passif	1114			+04	+02				-02	-02		-02		-03			+03				
Ind futur	2386	+08	+02	+02		-02	-04	-05	-05		-04	-03		+06	+04		+02				+02
Ind imparfait	2339		-02		-03		-03		+07	+04	+04	+03	+09		-03	-08					
Subj prés	1027			+02			+02				+02	+03		-03			-04	-03			
Conditionnel	1237		+06	+02	+12		+03	+05		+02	+10		-02	-04	-03	-05	-05	-03	-03		
Passé composé	4998		-06	-06	-05	-08	-06	-12	-08	-03	-15	-05	-05	+03	+02	+04	+04	+14	+15	+05	+05
Infinitif	13715	+04	+03			+02	+02		-07			-02					-02				
Part présent	1140	+02	+02	+05		+03	+03	+04	+04	+05	+04	+07	+02			-03	-04	-12	-08	-07	-03
Part passé	1438			+04	+03	+02		+04	+07	+05	+07	+03	+03	-02		-03		-18	-05	-06	
Il y a Ind prés	1143		-03	-04	-03	-05	-02	-05		-05	-03	-03	-04		-05	+03	+02	+10	+02	+08	+03
C'est Ind prés	3990	-08	-09	-10	-10	-13	-11	-10	-07	-09	-13	-08	-03			+02	+04	+51	+04	+20	+03
Il faut Ind prés	1284	-05	-02	-07	-05	-07	-03	-07	-04	-06	-03		-04		+02	+09	+03	+05		+09	
Pouvoir Ind prés	1182							-02	-03	-02				-02					+04	+04	+03