Normes de saisie et de dépouillement des textes politiques

Dominique Labbé

Université Pierre Mendès-France Grenoble II
Institut d'Etudes politiques de Grenoble

CERAT
politique - administration - territoire

Cahier n°7 Avril 1990

NORMES DE SAISIE ET DE DEPOUILLEMENT DES TEXTES POLITIQUES

Dominique Labbé

TABLE DES MATIERES

INTRODUCTION

Les sciences sociales, malgré leur grande diversité, affrontent un problème commun : la forme des informations, des données sur lesquelles elles travaillent. Ces données sont en grande partie des *communications* orales - discours, entretiens, témoignages - ou écrites : livres, journaux, revues, correspondances, archives, c'est-à-dire des *mots*. Or que font les chercheurs avec ces messages ? Au mieux, ils les traitent avec les instruments que nous a légués la critique littéraire. Quelques-uns se risquent à une analyse de contenu. Combien sont prêts à aller chercher dans la science du langage les instruments indispensables à une analyse approfondie des données qu'ils manipulent ? Il est vrai que leur démarche se heurte à des obstacles de taille. Ainsi l'éclatement de la linguistique générale en de multiples courants et chapelles. Ou encore la lourdeur des moyens nécessaires pour traiter un volume raisonnable de textes.

La question posée est simple en apparence : on dispose d'une série de messages et, aujourd'hui grâce à la micro-informatique, de moyens d'archivage et de tri. Comment exploiter ces ressources au profit de nos recherches ? Il n'existe malheureusement pas de réponse claire à cette question. Les solutions adoptées ont beaucoup varié, notamment au niveau de la saisie des textes, ce qui rend les enregistrements sur support informatique incomparables voire incompatibles[1]. De plus, l'évolution technique rapide a ouvert des possibilités qui paraissaient chimériques il y a peu et déclassé les fichiers anciens… Quant aux dépouillements, ils aboutissent à des résultats fort différents suivant qu'on décide de traiter les mots sensiblement tels qu'ils existent dans le fichier en machine ou qu'on opère de nouvelles codifications sur ceux-ci. Dans le premier cas, on travaille sur des "types" (ou "formes graphiques") : c'est l'option retenue à Nancy pour la confection du Trésor de la langue française ou au Laboratoire de lexicologie politique de Saint Cloud[2] ; en l'état actuel des choses, cette solution paraît difficilement évitable quand on opère sur de très vastes corpus. Dans le second cas, on convertit les mots en "vocables" à l'aide d'une "norme de dépouillement" qui s'ajoute à la norme de saisie. C'est l'option retenue par la plupart des lexicographes et par certains lexicomètres[3].

En l'état actuel des choses, chaque chercheur doit donc choisir entre plusieurs solutions en fonction des buts qu'il poursuit. C'est ce que nous avons fait depuis la fin 1983, date à laquelle nous avons entrepris d'implanter sur micro-ordinateur une chaîne raisonnée et complète de traitement des textes politiques. Au cours de ces sept ans, un journal de bord a été tenu sur lequel ont été notés, au jour le jour, les problèmes de méthode rencontrés, les arguments en

---

1. Deux normes de saisie méritent d'être signalées. Celle adoptée au début des années soixante pour la confection à Nancy du Trésor de la langue française (lire à ce sujet la préface au tome 1 du *Dictionnaire de la Langue du XIXe et du XXe siècle*, Paris, Klincksieck, 1971). Et du Laboratoire de lexicologie politique de Saint-Cloud : Pierre Lafon, Josette Lefevre, André Salem, Maurice Tournier, *Le Machinal. Principes d'enregistrement informatique des textes*, Paris, Klincksieck, 1985. Le mémoire de Majid Sekhraoui compare ces deux normes (*La saisie des textes et le traitement des mots : problèmes posés, essai de solution*, Mémoire sous la direction de Georges Th. Guilbaud, Ecole des hautes études en sciences sociales, juillet 1981).

2. Les principaux travaux du Laboratoire de lexicologie politique de Saint Cloud sont présentés dans la revue *Mots* publiée par les Presses de la Fondation nationale des sciences politiques.

3. L'un des premiers exposés d'une norme de dépouillement se trouve chez Charles Muller, *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse, 1967 (réédition : Genève-Paris, Slatkine-Champion, 1979, p 27-38). Voir également : Charles Bernet, *Le vocabulaire des tragédies de Racine (Analyse statistique)*, Genève-Paris, Slatkine-Champion, 1983, p. 27-31. Il existe beaucoup d'autres normes de dépouillement. Par exemple : Alphonse Juilland, Dorothy Brodin, Catherine Davidovitch, *Frequency Dictionnary of French Words*, La Haye, Mouton, 1970. ENGWALL Gunnel, *Vocabulaire du roman français (1962-1968) Dictionnaire des fréquences*, Stockholm, Almqvist-Wicksell International, 1984. Anthony A. Lyne, *The vocabulary of french business correspondance : word frequencies, collocations and problems of lexicometric method*, Genève-Paris, Slatkine-Champion, 1985…

présence, les solutions retenues, les opérations de codage, l'architecture des programmes et les difficultés survenues dans leur mise en oeuvre.

Le discours politique français contemporain était la "matière première" sur laquelle ont été expérimentés les outils mis au point. Ces analyses ont débouché sur plusieurs dépouillements[1], l'élaboration, en collaboration avec Pierre Hubert, d'un modèle de description du vocabulaire[2], la publication de quelques articles et d'un livre portant sur le premier septennat de François Mitterrand[3].

Le document qu'on va lire est issu de ce journal de bord rapidement remis en forme et allégé. Il s'agit d'un document de travail dont la formulation n'est pas définitive et qui ne prétend pas résoudre tous les problèmes. Le lecteur n'y trouvera pas un système achevé mais une série de procédés, des listes de cas aussi complètes que possible et quelques pistes de réflexion. Nous le prions de ne pas nous tenir rigueur des fautes, maladresses et redites qui nous auraient échappé au cours de cette remise en forme.

Les principaux programmes utilisés pour la saisie et les dépouillements des textes sont présentés dans le tableau récapitulatif ci-contre. La présente note ne porte que sur les deux premières étapes (saisie et lemmatisation). Les autres étapes et les calculs obéissent aux procédés standards (comme la constitution des index) ou ont déjà été présentés par ailleurs[4]. Nous avons placé en annexe une courte notice sur l'indice de répartition dont le calcul n'a jamais été explicité.Cet indice a été mis au point avec l'aide de Pierre Hubert. Il a été utilisé dans l'index placé à la fin de notre ouvrage sur le vocabulaire du président Mitterrand[5].

Le propos général de cette note peut se résumer ainsi : pour réaliser un traitement informatique des textes politiques, qui produise des résultats fiables et intéressants, deux conditions doivent être remplies. D'une part, la saisie de ces textes doit obéir à des règles rigoureuses et, d'autre part, il faut réaliser une lemmatisation préalable à tout traitement statistique.

En ce qui concerne le premier point, le principe de base veut qu'on fasse peser le moins de contraintes spécifiques sur la saisie. D'une part, pour le présent, *les règles de saisie ne doivent pas trop freiner l'opérateur et ne pas multiplier les risques d'erreur*. D'autre part, pour l'avenir, il faudrait pouvoir récupérer les fichiers constitués pour les usages normaux (disquettes de traitement de textes, base de données, publications, bandes des imprimeurs…) afin de les traiter après une relecture sérieuse mais n'équivalant pas cependant à une deuxième saisie. Ainsi

---

1. Outre les interventions radio-télévisées du président Mitterrand lors de son premier septennat, ont été dépouillés : les entretiens télévisés du général de Gaulle avec Michel Droit entre les deux tours de l'élection présidentielle de décembre 1965, les débats télévisés Giscard-Mitterrand (mai 1981), Chirac-Fabius (octobre 1985), Mitterrand-Chirac (avril 1988) ainsi que *La lettre à tous les Français* de Mitterrand (avril 1988). Les index sont disponibles auprès du CERAT.

2 Le "modèle de partition du vocabulaire" postule que tout locuteur dispose de plusieurs sources où puiser ses mots : un *vocabulaire général* où se trouvent les mots utilisés quelles que soient les circonstances et des lexiques *spécialisés* mobilisés en fonction du sujet traité ou en fonction de l'interlocuteur… En cas de spécialisation nulle ou très faible, chaque fragment du texte peut être analysé comme un "échantillon" de l'ensemble. En revanche, plus seront nombreux les mots tirés de lexiques spécifiques, plus on s'éloignera de cette situation idéale : chaque fragment aura une spécialisation lexicale et sa structure s'écartera de celle de l'ensemble. Le paramètre P ("de partition") mesure cet écart entre les observations et les valeurs obtenues par le calcul : il permet ainsi d'estimer le poids des vocabulaires spécialisés utilisés dans le texte. Cf Pierre HUBERT, Dominique LABBE, "Un modèle de partition du vocabulaire", in Dominique Labbé, Philippe Thoiron, Daniel Serant, *Etudes sur la richesse et la structure lexicales*, Paris-Genève, Slatkine-Champion, 1988, p 92-114.

3. Dominique Labbé, *Le vocabulaire de François Mitterrand*, Paris, Presses de la Fondation nationale des sciences politiques, 1990.

4. Voir notamment nos deux articles rédigés en collaboration avec Pierre Hubert et présentés dans l'ouvrage collectif sur la richesse du vocabulaire

5 . Voir Dominique Labbé, *op cit*, p 44-55. Egalement : Pierre Hubert, Dominique Labbé, "La répartition des mots dans le vocabulaire présidentiel", *Mots*, 22, mars 1990, p 80-89.

pourrions-nous anticiper sur les futurs programmes de reconnaissance des formes en définissant les quelques contraintes supplémentaires que pourraient exiger les lexicographes.

Sur le deuxième point, une pluralité de normes de dépouillement est inévitable. Après plus de vingt-cinq ans de polémiques à ce propos, il est devenu évident qu'une norme unique de dépouillement ne sera jamais admise parce que les philosophies et les buts qui motivent les dépouillements lexicographiques sont trop divers. En fait, tout milite pour une pluralité consciemment organisée entre deux pôles extrêmes correspondant aux deux grandes étapes dans l'analyse, qui elles-mêmes, se décomposent en plusieurs moments.

Premièrement, une "norme formelle" se place aussi près que possible du texte saisi suivant les règles courantes de la typographie du français. Elle traite des "formes" et aboutit à des listes qui sont la première étape de tout traitement lexicographique. C'est le laboratoire de lexicologie politique de Saint-Cloud qui a insisté sur l'importance du traitement des "formes graphiques" et qui a été le plus loin dans la codification de cette norme. Nous consacrons la première partie de cette note à la *norme de Saint Cloud*.

Deuxièmement, une "norme de lemmatisation" qui se place aussi près que possible des habitudes lexicographiques courantes (encore que celles-ci ne soient pas fixées bien rigoureusement[1]). En effet, le premier objectif des dépouillements est de parvenir à constituer des sortes de dictionnaires - des index, associés pour certains mots à des concordances - à partir desquels il est possible d'analyser à loisir le vocabulaire. Par là, le lexicographe veut livrer à un public de non-spécialistes des outils de connaissance sur un auteur, un groupe… mais aussi sur la langue. Charles Muller a été, en ce domaine, un pionnier. C'est pourquoi nous proposons de baptiser cette norme "lexicographique" : *norme Charles Muller*.

Naturellement, si C. Muller et le Laboratoire de Saint Cloud ont été des guides essentiels, la responsabilité des pages qui suivent est nôtre.

Nous consacrerons une partie de ce document à chacune de ces deux normes. Afin de ne pas surcharger le texte, nous avons renvoyé en annexe, une série de tableaux et de listes de mots. Pour faciliter la tâche du lecteur nous avons adopté la solution peu élégante consistant à numéroter les paragraphes du texte. Nous donnons à la fin du livre une liste des principales difficultés avec le numéro du paragraphe où figure la solution proposée. Si le mot ou l'expression recherchés ne figurent pas dans la liste placée à la fin de l'ouvrage, la table des matières détaillée permettra de retrouver aisément le passage où le cas est traité.

---

1. Cf par exemple la comparaison entre les dictionnaires dans Alain Rey, *Le lexique : images et modèles du dictionnaire à la lexicologie*, Paris, A Colin, 1977.

PREMIERE PARTIE.
LA NORME SAINT-CLOUD

> "L'enregistrement d'un texte n'est pas fait pour interpréter, coder les sens, les contenus ou les thèmes, analyser les liens et fonctions grammaticales et réunir les flexions sous des lemmes, mais pour fournir simplement au chercheur un matériel identique ou presque à l'édition de référence."
> (*Le Machinal*, p 6-7)

La "norme de Saint-Cloud" désigne les règles régissant la saisie des textes sur support informatique et le traitement des fichiers qui résultent de cette première opération.

Comme l'indique notre titre, nous nous appuyons ici principalement sur la norme élaborée par le Laboratoire de lexicologie politique de Saint-Cloud telle qu'elle est codifiée dans le *Machinal*[1] après avoir été ébauchée dans les années soixante[2]. A l'usage, il nous est apparu que cette norme pouvait recevoir certaines adaptations en fonction du matériau un peu particulier que nous traitons :

- d'une part, elle doit être complétée par quelques règles visant à normaliser les transcriptions des prestations orales sur lesquelles nous avons travaillé ;
- d'autre part, elle peut être allégée de la plupart des clefs — que l'équipe de Saint-Cloud appelle "péri-textuelles" et qui visent principalement à mémoriser les caractéristiques typographiques du texte — puisque nous travaillons essentiellement sur des transcriptions de l'oral où ces préoccupations n'ont pas lieu d'être.

Quelques aménagements secondaires portent également sur la saisie des textes (chapitre I).

La reconnaissance des formes a été reprise plus à fond notamment quant aux problèmes posés par les mots composés et les locutions, problèmes qui sont trop brièvement traités dans le *Machinal* (chapitre II).

Enfin, l'ensemble de ces opérations étant réalisées avec l'assistance de l'ordinateur, nous décrivons succinctement les programmes qui ont été élaborés pour la circonstance, le format des fichiers et les sorties obtenues à ce stade du traitement (chapitre III).

---

1. Pierre Lafon, Josette Lefevre, André Salem, Maurice Tournier, *Le Machinal. Principes d'enregistrement informatique des textes*, Paris, Klincksieck, 1985. La philosophie d'ensemble qui fonde cette norme n'a jamais été totalement explicitée. On en trouvera un bon résumé dans Michel Demonet, Annie Geffroy et Al., *Des tracts en mai 68*, Paris, Presses de la Fondation nationale des sciences politiques, 1975, p 19-28.

2. Annie Geffroy, Pierre Lafon, Maurice Tournier, *Enregistrement et traitement lexicométrique des textes*, Paris, CNRS, 1975. Cf également le mémoire de Majid Sekhraoui, *La saisie des textes et le traitement des mots : problèmes posés, essai de solution*, Mémoire sous la direction de Georges Th. Guilbaud, Ecole des hautes études en sciences sociales, juillet 1981.

CHAPITRE I.
LES REGLES DE SAISIE.

L'objectif premier est ici de limiter, autant que possible, les opérations sur le corpus pour des raisons de temps - la saisie et les corrections sont déjà des opérations longues et fastidieuses - et de maîtrise des problèmes de variations graphiques. De ce fait, nous nous écartons du *Machinal* qui montre un certain culte pour la forme imprimée et qui risque de donner trop d'importance à ce qui n'est que de la typographie ou des conventions passagères[1]… Donc, si l'étude des variantes typographiques n'est pas incluse au départ dans les objectifs de la recherche, il nous semble que l'on peut s'épargner certaines de clefs conventionnelles proposées dans le *Machinal*…

D'autre part, nous travaillons sur la transcription de bandes sonores ou audio-visuelles. En effet, au cours de son septennat, F. Mitterrand, voulant garder la possibilité de changer son discours jusqu'au moment de le prononcer, ne communique pas de texte écrit à ses services. Nous disposons donc de deux sources : la bande sonore et la transcription qui en est effectuée - a posteriori - par les services de presse de l'Elysée. Notre propre transcription a suivi au plus près les techniques de la sténographie. Nous ne nous en sommes écartés que pour la ponctuation et pour rétablir quelques redites ou hésitations coupées par le secrétariat[2] .

1.1. LA PONCTUATION

La transcription informatique de la ponctuation d'un texte écrit suit quelques règles simples. En revanche, la ponctuation d'une transcription de l'oral pose de redoutables problèmes.

1.11. Les règles de codification de la ponctuation

Avec un texte écrit il suffit de respecter la ponctuation de l'auteur en éliminant les signes de "fantaisie". Les seuls signes reconnus par nos programmes sont : ( ) " " - , ; : ? ! … Les crochets sont convertis en parenthèses. Les différents styles de guillemets sont fondus en un seul, les barres (/) sont converties en virgules.

Nous avons suivi les règles normales de la typographie. Celles-ci ne sont pas toujours uniformes. A noter :

- Le tiret doit être impérativement précédé et suivi d'un blanc pour le démarquer du tiret agglutinant des mots composés ;
- Nous plaçons également un blanc de séparation devant et derrière les signes suivants ; : ? !
- la virgule, le point, les parenthèses et les guillemets sont collés au mot.
- les trois points (…) forment un seul caractère afin d'éviter la confusion avec le signe de la ponctuation majeure (qui marque la fin de la phrase).

Enfin, les virgules, les guillemets, voire les points-virgules peuvent jouer le rôle de séparateur de chaînes de caractères dans les compilateurs informatiques et ne pas être reconnus comme du texte. D'où la nécessité d'un programme de transcodage préalable (ce programme est entièrement automatique et, pour éviter les erreurs, l'opérateur n'intervient pas à ce stade).

---

1. Ce respect de la forme graphique s'explique aussi par la nature du matériau traité : textes imprimés et journaux où la typographie, les italiques, les majuscules, la mise en page… peuvent jouer un rôle important comme l'expliquent les premières pages du *Machinal*.

2. Après le début de notre travail, est paru le livre de Blanche-Benveniste et Jeanjean sur le Français écrit. Cet ouvrage comporte de considérations très pertinentes dont certaines dépassent d'ailleurs le cadre de notre étude…Claire Blanche-Benveniste et Colette Jeanjean, *Le français parlé*, Paris, Didier, 1987. Voir également le compte rendu de Françoise Gadet dans *Mots*, 18, mars 1989, p 118-122.

1.12. La transcription de la ponctuation d'un discours oral

Nous nous sommes trouvé face à un problème redoutable : la transcription des prestations orales. Comme nous l'indiquions ci-dessus, les interventions télévisées du président Mitterrand ne nous sont connues que par leur enregistrements et leurs éventuelles transcriptions par son secrétariat. Outre la question, secondaire pour nous, des hésitations, des "euh" (fort peu nombreux) et de la normalisation des interjections, la difficulté essentielle réside dans la ponctuation.

1.121. La ponctuation selon les conventions sténographiques

Les sténographes ponctuent en fonction de la longueur de la pose dans le débit oral et ne tiennent manifestement peu ou pas du tout compte de la cohérence syntaxique de la phrase. On a :

- la virgule marque une pause légère ou sépare des groupes nominaux non coordonnés dans les énumérations qui ne sont pas interrompues par des pauses ;
- le point indique une pause marquée précédée d'une descente dans l'intonation ;
- le point d'exclamation remplace le point quand la pause est précédée d'un maintien ou d'une montée de l'intonation ;
- le point d'interrogation remplace le point à la fin d'une période quand l'intonation ou la construction indique une nuance interrogative ;
- les points de suspension indiquent un silence marqué, une période interrompue ou une reprise qui interrompt le déroulement normal de la période.

En confrontant les transcriptions réalisées par les services de l'Elysée avec les enregistrements des émissions correspondantes, nous avons pu constater que le travail de sténographie était généralement de bonne qualité mais que le résultat était souvent incohérent, du point de vue sémantique, car le style de F. Mitterrand s'écarte assez sensiblement des conventions qui viennent d'être présentées…

1.122. La ponctuation à bases syntaxiques et sémantiques

Le cas de François Mitterrand montre combien les conventions formelles de la sténographie peuvent parfois se révéler incompatibles avec une transcription sémantiquement cohérente. En effet, le président marque des pauses importantes au milieu de ses périodes oratoires et comme, de plus, il place souvent de multiples parenthèses entre les éléments essentiels de la phrase - tel le verbe et son sujet - il arrive fréquemment que, dans les transcriptions diffusées par ses propres services, ces éléments syntaxiquement liés se trouvent séparés par deux ou trois points. Ces difficultés peuvent rendre assez obscurs les propos présidentiels. Il est donc indispensable de recourir à des critères syntaxiques et sémantiques pour rétablir une ponctuation cohérente et, subsidiairement, pour limiter l'usage du point de suspension qui a tendance à proliférer dans ces transcriptions !

L'appel à ces critères permet également de compléter les transcriptions où certains signes sont peu ou pas employés :

• Les deux points se placent entre deux termes d'une phrase quand le second est présenté comme le développement logique du premier :

- le second terme est une conséquence du premier ("Le rôle principal du président de la république : veiller au respect de la constitution") ;
- le second terme est une énumération annoncée par le premier ("Le président a trois fonctions : garant de la constitution, chef des armées…") ;
- le second terme est une parole rapportée ("Vous dites : «Je dissoudrais l'Assemblée…»") ;

• Les guillemets :

- encadrent les parole rapportées (cf ci-dessus) ;

- encadrent un ou plusieurs mots que l'auteur ne veut pas prendre à son compte de manière explicite ("c'est ce que vous appelez la «cohabitation»") ;

• les parenthèses :

Elles encadrent dans la phrase un élément isolé et non coordonné avec le reste du propos. Il est parfois difficile de choisir entre la parenthèse et les virgules. La parenthèse est l'exception et correspond à un groupe nominal ou une brève proposition non coordonnée avec ce qui précède et ce qui suit ; une inflexion prononcée de la voix peut également conduire à arbitrer en faveur des parenthèses ;

• Les tirets assurent deux fonctions :

- ils placent sur le même plan deux propositions enchâssées l'une dans l'autre sans élément de coordination : "Le rôle de la France - je le répète encore une fois - c'est…" ;

- ils jouent un rôle proche de la parenthèse, mais marquent que dans le débit oratoire a été interrompu par une pause plus importante. Ils assurent la mise en relief de l'élément entre tirets : "Le président des Etats-Unis - nouvellement élu - m'a indiqué…") ;

Le choix entre ces différents éléments n'est pas toujours aisé. Il serait certainement souhaitable que des études plus approfondies soient menées sur ce point. En effet, la phrase est l'élément naturel de l'étude du contexte étroit des mots. De plus, sa construction est un trait essentiel du style d'un auteur[1].

1.2. LES MOTS A MAJUSCULE

En principe seuls les "noms propres" ont leur première lettre en capitale. Cependant, il peut y avoir quelques autres "mots à majuscule".

1.21. Les noms propres

La catégorie des noms propres doit être conçue restrictivement : noms de personnes, de pays, de peuples et de lieux (mers, fleuves, villes, régions, départements, monuments…) Le nom propre sera simplement identifié par sa première lettre en majuscule (la France, les Français…) Cette dernière règle est impérative pour distinguer l'adjectif du nom (les *Américains*, "les citoyens *américains*"). Cette convention impose que la majuscule soit réservée aux noms propres (cf. plus bas…)

Normalisation des transcriptions :

- pour les noms propres composés, le premier membre comporte une majuscule afin de signaler où commence le nom propre[2], par exemple : De Gaulle, La Rochelle, De la Palice, Grande Bretagne, Le Perreux…

- s'il est mentionné, le prénom est écrit en toutes lettres : "R. Barre" devient "Raymond Barre"[3]. Ce choix pose le problème du M. (M. Rocard : Monsieur ou Michel ?) que la bande son nous permet de résoudre ;

---

1. Nous renvoyons sur ce point à Conrad Bureau, *Linguistique fonctionnelle et stylistique objective*, Paris, PUF, 1976. Nous avons tenté une application de ces instruments sur les interventions radio-télévisées du président. Nous en rendons compte dans le dernier chapitre de notre livre sur *Le vocabulaire de François Mitterrand*.

2. Cette convention un peu gênante a été adoptée pour uniformiser les graphies et simplifier les programmes informatiques ou la constitution des index. A l'usage, il apparaît en effet que les tables de noms propres, que l'on peut élaborer, deviennent vite gigantesques et qu'elles ne couvrent jamais qu'une petite partie du champ potentiel. De ce fait, l'opérateur est souvent sollicité.

3. Nous avons constaté que, dans les documents à notre disposition, il n'y avait pas de normalisation sur ce point. Par exemple, quand le président dit : "Laurent Fabius", le prénom est transcrit parfois par L. et parfois en toutes lettres…

- Le prénom est une forme distincte du nom. Les raisons de ce choix tiennent à la recherche d'une norme synthétique et à la nécessité d'un index aussi aisé à manipuler que possible : si {prénom-nom} est une forme à côté de {nom} : il faudra aller chercher une éventuelle mention à Raymond Barre sous Raymond et sous Barre avec une difficulté supplémentaire pour les noms propres en plusieurs mots (Giscard, Giscard d'Estaing et Valéry Giscard d'Estaing…) Au contraire, le fait de détacher le prénom conduit à une seule entrée par personne nommée. En contrepartie, on perd une information : qui a droit au prénom qui est appelé par son nom seul ? Le programme de concordance permet de retrouver aisément ce renseignement…

- Les noms de bateaux donnent en une seule forme et commencent toujours par une majuscule. Ainsi par exemple, pour le 14 juillet 1982, le Président Mitterrand visite l'escorteur "Georges Leygues" (un seul mot)…

1.22. Les autres mots à majuscules

Dans le français contemporain, le nom commun à majuscule est une catégorie en expansion rapide. Ainsi écrit-on aujourd'hui le *P*remier ministre, le président de la *R*épublique, l'*A*dministration ou l'*U*niversité… Il en est d'ailleurs de même dans les sciences ou l'on écrit volontiers "la *S*cience", "la *P*hysique", "l'*H*istoire" ou "la *M*atière"… L'index des noms propres risque d'être encombré par ces intrus et le lecteur sera certainement dérouté de ne pas trouver "administration", "république" ou "constitution" dans l'index des noms communs où, en homme raisonnable, il s'attend à les trouver. Aussi écrirons-nous - contre les conventions dominantes - *r*épublique, *h*istoire,*a*ssemblée, *p*arlement, *a*dministration…

Outre qu'il serait dangereux de donner à la catégorie des noms propres une extension excessive et des contours flous, deux difficultés supplémentaires militent pour cette solution :

- d'une part, la longueur excessive de certaines appellations : on retrouve le même problème que pour les chiffres (cf ci-dessous, § 1.31) ;

- d'autre part, la variabilité de certaines formes : on dit aussi bien "Fonds monétaire" que "FMI" ou "Fonds monétaire international", "l'Assemblée" que "l'Assemblée nationale" voire "la Chambre"… La lexicalisation n'est donc pas totale contrairement par exemple à Moyen-Orient (le trait d'union n'est pas une preuve absolue puisqu'il est là principalement pour une raison euphonique…)

Au total : "Airbus" devient "airbus" ; "Assemblée", "assemblée", "Premier ministre", "premier ministre" ; "République", "république" ou "Sénat", "sénat"…

On peut convenir de conserver la majuscule dans certains cas et à condition que son usage soit de règle ou, pour le moins, assez fréquent :

- pour distinguer des homographes affectant des formes à haute fréquence. Par exemple, "Est" : point cardinal et verbe être à la troisième personne du singulier ;

- pour distinguer, au sein d'une même catégorie grammaticale des mots homographes dont le sens est différent : Mirage (l'avion) et mirage (l'illusion), Etat et état (civil) ; Constitution (texte), constitution (physique) ; Communauté (européenne). Si nous sommes finalement résolus à voir dans la "Communauté" (européenne), un nom de pays au même titre de Europe, les autres mots dont les majuscules ont été conservées au dépouillement sont comptés dans l'index terminal comme noms communs et non comme noms propres et ils sont transcrits en minuscules avec un code indiquant éventuellement une homographie au sein d'une même catégorie et d'un même genre grammatical (cf. le § 6.122).

1.23. Les abréviations et sigles…

En face des sigles, une première question se pose : faut-il rétablir leur signification intégrale. Ainsi écrire "parti socialiste" au lieu de "PS" comme nous le faisons pour "R. Barre" qui devient "Raymond Barre". Cette solution ne peut-être retenue en lexicographie pour deux raisons :

- ici le locuteur dit vraiment : "PS" ;

- on voit bien qu'il y a une nuance entre les deux emplois et que le choix de l'une contre l'autre est porteur de sens. Le respect de la forme s'impose ici…

Les sigles sont reproduits en lettres majuscules et sans point ni entre chaque lettre ni à la fin du sigle car le point est réservé à la ponctuation et il est toujours considéré comme un séparateur de phrase : "CFDT" ou "PS" et non "C.F.D.T.", "P.S.").

Les abréviations sont également proscrites. Par exemple M. devient monsieur, Mme devient madame et F. devient franc(s), etc.

1.3. LES NOMBRES, CHIFFRES ET DATES

1.31. La codification des cardinaux

En dehors de "un" qui est indissolublement adjectif cardinal et article - mais aussi cardinal et pronom - l'ensemble des cardinaux peuvent être écrits en chiffres ou en lettres. On rencontre même des transcriptions complexes, tantôt en chiffres, tantôt en lettres qui respectent au plus près la formulation effectivement employée par le locuteur (par exemple : "2 milliards 850 millions de francs").

On pourrait songer à écrire le nombre en lettres et le compter comme une seule forme. Cette solution est cependant difficilement praticable : elle multiplierait les formes différentes. Par exemple, dans notre corpus il y a "mille neuf quatre-vingt un" ou "dix neuf cent quatre-vingt un" (pour 1981) De plus, la dimension de certains chiffres est un obstacle insurmontable : la gestion de la mémoire de l'ordinateur est possible mais la typographie de formes aussi longues dans un index papier génère des problèmes insolubles de mise en page.

Quand à scinder le chiffre en autant de mots qu'il contient de nombres, cette solution se heurte à deux objections :

- il pourrait être intéressant d'analyser les chiffres employés par tel ou tel homme politique. Or cette réalité du chiffre sera découpée en unités non pertinentes et donc perdue ;

- la convention graphique ne peut être respectée pour certains nombres. En effet, pour des raisons essentiellement euphoniques, on écrit : "dix-huit", "vingt-deux" ou "quatre-vingt" mais "vingt et un" ; ou encore : "soixante et onze" et "quatre-vingt-onze". Sans compter l'accord compliqué de "quatre-vingts" ou "huit cents"…

En face de ces inconvénients, une double solution a été retenue :

- écriture en chiffres et une seule forme pour le nombre quelle que soit sa longueur : 8250000 (pour 8 millions 250 mille)… L'étude des chiffres devient ainsi possible (densité, fréquence, répartition dans le corpus, etc…) Seule la virgule décimale est admise et l'on veillera spécialement à ce qu'il n'y ait pas de blanc ou de point dans les chiffres, ce qui exclut les séparateurs de milliers. Ce chiffre "original" est conservé entre < > et peut faire l'objet d'une étude spécifique, tout en ne faisant pas partie du texte lemmatisé ;

- conversion des chiffres en lettres au cours de la lemmatisation. Ici on renonce aux traits d'union dans un souci d'harmonisation et de simplification : "90" s'écrira "quatre vingt dix" en trois formes… Ceci explique que, dans l'index du vocabulaire de F. Mitterrand, on ne trouve aucun des chiffres composés de plusieurs nombres unis par un trait d'union.

1.32. La codification des ordinaux

En revanche, la codification des ordinaux se fait suivant les règles exposées à propos des mots composés : "dix-huitième" est une seule forme comme "seizième" (siècle, arrondissement…) ou "vingt et unième" (siècle) que nous avons rencontré à plusieurs reprises (chez Giscard mais non lors du septennat Mitterrand).

Ici également, les notations fantaisistes doivent être proscrites : non pas"1er", "2nd", "5ème" mais *premier*, *second*, *cinquième* puisqu'il s'agit d'ordinaux et non pas de cardinaux (donc pas de chiffres mais des lettres).

1.4. L'UNIFORMISATION DE LA GRAPHIE DES FORMES

Le problème se pose un peu différemment pour les verbes et pour les autres formes.

1.41. Les verbes à graphies multiples

Pour certains verbes, les conjugaisons anciennes se sont maintenues ; par exemple, les verbes du type "balayer", "payer". La double graphie n'a pas été réduite. Le premier principe énoncé dans notre avertissement nous a dicté cette solution (dans l'index les formes sont rattachées au même infinitif). A la réflexion, nous ne sommes pas aussi sûr du bien-fondé de cette décision. Deux arguments militent en faveur d'une graphie unique : d'une part, cette graphie unique s'impose pour les substantifs dérivés de ces verbes et d'autre part une difficulté supplémentaire survient lors de la lemmatisation puisqu'on se trouve face à des homographies du type "paie-paie" (substantif féminin et verbe payer) ou du type (puis-peux, verbe pouvoir à la première personne de l'indicatif ou conjonction)…

1.42. Les autres formes à graphies multiples

En français, la graphie de certaines formes est instable. Par exemple : "grosso modo" pour le Robert mais "grosso-modo" pour Grévisse. Les incohérences de ce genre sont fort nombreuses et le Conseil international de la langue française (CLIF) les analyse depuis plus de vingt ans[1]. Nous nous sommes rallié à ses propositions. Quand il indique une tolérance nous utilisons la forme qu'il donne en premier. Quand deux formes sont indiquées concurremment ou que le CLIF est silencieux, nous avons retenu la graphie du petit Robert. Ceci a permis l'uniformisation des substantifs à graphie double "paie et non pas paye", "paiement" et non pas "payement", etc.

La graphie est également instable pour les noms propres notamment les noms étrangers dont la transcription n'est pas codifiée ou se trouve contestée (ainsi les noms chinois). Par exemple, nous avons rencontré : Hissein, Hissen, Hissenne (pour le prénom du président tchadien Habré) ; Canaques, Kanaks, Kanacks ; Chah, shah et schah pour l'empereur d'Iran…

Pour les noms composés, la graphie n'est pas toujours unifiée. Il faut se méfier des formes comme "entr'ouvrir" qui s'écrit de plus en plus couramment : "entrouvrir" ; entre-temps (entre temps, entretemps), entre-jambes (et entrejambes). Ce problème est discuté dans le chapitre suivant.

---

1. Voir à ce sujet : Conseil international de la langue française, *Pour l'harmonisation orthographique des dictionnaires*, Paris, CLIF, 1988.

*

* *

*

En conclusion, nous voudrions souligner que ces règles ne prétendent pas être exhaustives. L'essentiel réside dans leur stabilité au cours du traitement. Si l'une d'entre elles se trouve modifiée en cours de route, il faut revenir en arrière et l'appliquer sur tout le corpus. Nous voudrions également plaider pour une certaine simplicité. Même si, intellectuellement, certains partis pris peuvent être contestés, la définition d'une norme de saisie simple et stable devrait permettre des comparaisons entre des fichiers créés par différentes personnes ou plusieurs laboratoires. Elle permettrait aussi à la statistique lexicale de ne plus travailler sur du sable comme c'est trop souvent le cas actuellement…

CHAPITRE II
LES FRONTIERES ENTRE LES FORMES

Le problème des noms propres est déjà réglé. Dans ce chapitre, nous discuterons du caractère séparateur ou agglutinant des signes non alphabétiques, ce qui nous permettra de mettre en valeur des locutions et des mots composés grâce à quelques règles simples et restrictives. Le principe de base est le suivant : tout signe graphique autre que les lettres de l'alphabet marque une frontière entre deux mots, un certain nombre d'exceptions existent mais elles sont fondées sur des règles strictes et sont interprétées restrictivement.

2.1. LES MOTS COMPOSES SANS TRAIT D'UNION

2.11. Le problème des locutions en plusieurs mots

De manière générale, il est souhaitable de limiter le nombre de mots composés, locutions diverses. Ils manifestent la vitalité créatrice des utilisateurs du français et, par là même, ils sont instables. De plus, on finit par élaborer des listes trop longues et inutilisables. Lorsqu'on examine les index réalisés depuis trente ans, on constate de nombreux désaccords. Le Gougenheim est de loin le plus accueillant en matière de locutions mais la justification n'en est pas toujours évidente (par exemple : "tout de suite")[1]. Dans son introduction, Juilland passe très rapidement sur cette question, mais dans le corps de l'index, on trouve de nombreuses surprises Par exemple : "au revoir", "chef d'oeuvre", "moyen âge" ou les locutions conjonctives formées avec que ("aussitôt que", "bien que", "pendant que"…) sont analysées comme un seul mot ; en revanche, on ne trouve pas "c'est-à-dire", "quelqu'un" ou l'ensemble des pronoms associés à même ("moi-même", "toi-même"…) qui ont donc été découpés en deux mots[2]. En revanche G. Engwall[3] ne retient pratiquement que le trait d'union : "parce que" notamment est découpé en deux formes mais "par-là-dessus" ou "sur-le-champ" n'en forment qu'une…

2.12. Les critères généraux de délimitation des mots composés

Quels critères retenir ?

• Il nous semble nécessaire de mettre au premier plan trois principes :

- d'une part, entre plusieurs solutions égales par ailleurs, il faut préférer celle qui facilite le plus le travail de l'opérateur ou qui permet une codification automatique sans erreur ;

- d'autre part, la stabilité de la forme graphique prime le reste puisque nous sommes ici à la recherche d'unités morphologiques et non pas sémantiques[4].

- en troisième lieu, nous prendrons en compte la "lexicalisation" comme critère subsidiaire de décision.

---

1. Georges Gougenheim et al., *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier, 1964.

2. Alphonse Juilland, Dorothy Brodin et Catherine Davidovitch, *Frequency dictionnary of French Words*, Paris-La Haye, Mouton, 1970. Ce dépouillement porte sur une série d'échantillons représentatifs du français écrit du XXe siècle (500 000 mots).

3. Gunnel Engwall, *Vocabulaire du roman français (1962-1968)*, Stockholm, Almqvist et Wiksell, 1984. Dépouillement d'échantillons tirés de 25 romans français parus dans les années 1960 (soit 500 000 mots) pouvant être considérés comme représentatifs de la langue littéraire contemporaine.

4. Cf à ce sujet Alain Rey, Alain Rey, *Le lexique : images et modèles du dictionnaire à la lexicologie*, Paris, A Colin, 1977, p 25.

• Il n'est pas possible d'adopter un principe simple - du genre : "le trait d'union agglutine, l'apostrophe et le blanc séparent" - car :

- certains traits d'union ont un rôle purement "euphonique" ; ils signalent une liaison à l'oral et n'indiquent aucun lien sémantique entre les formes qu'ils relient. Ils équivalent donc à des signes de séparation ("ainsi soit-il")…

- en sens inverse, dans certaines expressions, on peut considérer que l'élément de liaison est implicite. Ainsi "de Gaulle" donne une seule forme malgré le blanc séparant les deux éléments du nom propre. On aura alors un mot composé.

2.13. Les règles de reconnaissance des mots composés et locutions

Pour détecter ces groupes de mots agglutinés, nous avons élaboré six règles simples que nous présentons ci-dessous.

2.131. La règle "a fortiori"

Le français a hérité de quelques locutions latines d'usage courant. Il faut considérer ces locutions comme une seule forme. Une seconde raison milite pour ne pas décomposer ces locutions : elles comprennent souvent "a", "de"… qui risqueraient d'être confondus avec leurs homographes du français. La liste en est donnée dans l'annexe 1.

Se rattachent à cette première catégorie des expressions vieillies ou étrangères comme "ès sciences", "ès qualité" ou "rock and roll", "show-biz", "week end"…

2.132. La règle "aujourd'hui"

Quel que soit le séparateur employé celui-ci est agglutinant quand il relie deux formes qui ne sont plus employées seules. Ainsi on ne peut employer séparément ni "aujourd'" ni "hui".

2.133. La règle "parce que"

Quel que soit le séparateur employé celui-ci est agglutinant quand une des deux formes qu'il relie n'est plus employée que dans l'expression figée : parce que, tandis que, bric à brac, d'emblée… Certes la conjonction "que" est employée dans de nombreux autres cas de figure mais "parce" ou "tandis" ne sont jamais utilisés sans "que" et rien ne peut venir s'intercaler entre les deux membres de la locution. Par exemple, la même interprétation vaut pour "bric à brac" ou "clin d'oeil" : on trouve aussi "de bric et de broc" mais pas d'autre utilisations ni de "brac" ni de "broc" ; il en est de même pour "clin" (sauf en charpenterie de marine !)

Se rattachent à cette catégorie : à reculons, à l'envi, à l'instar, peu ou prou, la plupart, d'emblée, etc… De même que les formulations particulières utilisant des lettres seules ou des sigles : anti-UV, livret A (de caisse d'épargne), système d, rayons x…

2.134. La règle "d'abord, d'accord"

Quel que soit le séparateur employé celui-ci est agglutinant quand, en pratique, une expression à forte fréquence d'emploi ne comporte plus d'homographe. Ainsi la locution adverbiale "d'abord" : en théorie on peut rencontrer le substantif "abord" employé au singulier avec l'article "de" mais, en pratique, toutes les attestations le donnent au pluriel ("des abords difficiles") ou, au singulier toujours précédé de "l'" ou de l'adjectif "prime" ("dès l'abord"). Dès lors, "d'abord" ne peut-être qu'une locution adverbiale même si elle ne remplit pas les conditions présentées en 1.2 et 1.3. Cette locution n'est pas retenue par Juilland ni par Muller ni par Engwall.

Pour un nombre limitatif de *locutions adverbiales* à haute fréquence, on analyse celles-ci comme une seule forme mais après examen des homographes potentiels et désambiguïsation de ceux-ci. Ces cas doivent être définis *a priori* et de manière définitive ; l'étude des concordances est nécessaire afin de pouvoir traiter à part les éventuels substantifs "accord" précédés de "d'".

Après coup, nous doutons du bien-fondé de cette décision. Nous l'avons prise après avoir constaté que, dans les corpus étudiés, l'association {de+accord} était toujours utilisée adverbialement. Cependant, la solution {substantif singulier précédé de l'article "d'"} est certainement attestée dans certains corpus (exemple : "il n'y a pas d'accord explicite entre eux"). Par conséquent, il n'est peut-être pas souhaitable de conserver cette locution comme une seule forme au risque d'entraîner des erreurs.

2.135. La règle "quelqu'un"

C. Muller considère comme une forme unique "quelqu'un" (pronom désignant "une personne"). Mais alors que faire avec "quelques uns" ("certaines personnes") qu'il faudrait différencier de "un petit nombre de gens" ou de choses ("il m'en a donné quelques uns") ? Une flexion de quelqu'un dans le premier cas et de "quelque" et de "un" dans l'autre ? Le principe n° 1 ci-dessus conduirait à considérer que tous ces cas sont deux formes distinctes : "quelque" et "un"… Cependant, trois arguments militent pour la solution inverse. Premièrement, entre deux solutions d'inconvénients égaux, il nous semble préférable de privilégier celle qui respecte la graphie normale. Deuxièmement, la plupart des index retiennent "quelqu'un" comme un seul item. Troisièmement, au cours de l'étape suivante (lemmatisation), on considérera que "quelques uns" est une flexion de "quelqu'un" (pronom). Afin d'éviter des distorsions trop fortes entre l'index des formes et l'index des lemmes, il paraît souhaitable de conserver "quelqu'un", "quelques uns" et "quelques unes" comme des formes uniques que nous pourrons ultérieurement considérer comme des flexions d'un vocable unique. En revanche "quelqu'autre", "quoiqu'il" sont deux formes…

2.136. Remarques

Les cas examinés en 4 et 5 sont restrictivement énumérés a priori. Dans leur définition, on ne retient pas la notion trop floue de "lexicalisation". Le critère essentiel sera l'impossibilité de démembrer ou de formuler différemment la locution. Par exemple, le fait que l'on puisse dire "pomme de terre", "pomme dauphine", "pommes vapeur", "pommes frites" — aussi bien que "pommes de terre frites" ou "frites" - interdit de les considérer comme une seule forme… Ceci exclut des expressions comme "chemin de fer", "trait d'union" ou "main d'oeuvre" dans lesquelles les lexicographes ont tendance à voir un seul mot[1]… Ces compositions ne comportant aucun signe de liaison sont assez nombreuses. C'est notamment le cas dans le locutions formées avec les mots : clé, mère, modèle, pilote, témoin, type… qui, sauf rares exceptions, ne prennent pas de traits d'union et seront donc rejetées par nos règles.

On pourrait songer à réintégrer ces locutions dans la liste. Cependant, il est souhaitable que la règle admette le minimum d'exceptions car elles sont dangereuses. Les multiplier c'est aussi augmenter les sources d'erreur, les pertes de temps pour l'opérateur, les sources de contestation ; c'est aussi rendre plus difficile les comparaisons entre dépouillements effectués à différentes époques… Nous avons donc résolu de les limiter très soigneusement.

Pour les même raisons, il n'est pas question de mettre un petit chien devant ou derrière la plupart des verbes en leur attachant les pronoms réfléchis, les prépositions avec lesquels ils se construisent habituellement (se, à, de, que, etc).

---

1. voir Alain Rey, *op. cit.*, p 23 sq.

Malgré ces interprétations restrictives, la norme qui vient d'être décrite génère déjà une liste de locutions fort longue quoique non exhaustive (annexe 1).

2.2. LES TIRETS AGGLUTINANTS

A priori, le trait d'union réunit deux mots pour en former un seul. Cependant l'existence du trait d'union s'explique souvent pour des raisons euphoniques ou graphiques et non pas sémantiques : dans ces cas, nous ne sommes pas forcément face à une seule unité.

2.21. Le trait d'union non-agglutinant

Les emplois grammaticaux ou d'usage n'établissant pas de lien indissoluble entre les mots ("est-ce…, n'avez-vous pas, alla-t-il, etc.) sont systématiquement remplacés par un blanc séparateur : "n' avez vous". Cette règle amène une suppression complète du -t- euphonique (par exemple lors de l'inversion du sujet et du verbe) "Alla-t-il" devient "alla il". Cette suppression est également motivée par le souci d'éviter la confusion avec le pronom de la deuxième personne du singulier (il en va de même pour le "l'" de "l'on" (on risque la confusion avec l'article ou le pronom "le").

2.22. Les compositions libres et les locutions figées

2.221. Discussion du problème

Les mots composés grâce au trait d'union ont tendance à se multiplier. Du point de vue lexical, c'est la manière la plus simple de créer des mots nouveaux. Mais le mot unique doit être distingué de la composition libre… La frontière est difficile à tracer. Ainsi avons nous relevé dans la littérature des compositions libres comme : "je-sais-tout", "bon-à-rien"… Par exemple, en 1986-87, lors d'un scandale politique, à propos d'un faux passeport délivré par l'autorité administrative, on a vu fleurir les "vrai-faux", "vrai-vrai", "faux-vrai" qui étaient autant de jeux de mots satiriques. Peu de mois après la mode est passée et il aurait été erroné de retenir ces compositions comme une seule forme. Il faut donc asseoir l'analyse sur quelques principes aussi simples que possible et ne faisant pas trop appel au "sentiment linguistique" de l'opérateur…

2.222. Les principes généraux d'analyse des mots composés

L'idée qui semble communément admise est qu'on est en présence d'un *mot unique* quand l'emploi est *lexicalisé*. Mais comment juger de cette lexicalisation ? Outre les dictionnaires qui peuvent être des témoins commodes mais parfois contradictoires, on utilise les deux critères proposés par Martinet[1] :

- la combinaison est susceptible d'entrer dans les mêmes constructions que le constituant principal (par exemple, "chaise" et "chaise-longue" ; "une longue chaise-longue" est à la limite concevable) ;
- aucun des composants ne peut-être modifié séparément des autres sans détruire le mot-composé ;

Quand le mot composé ne se trouve pas dans le dictionnaire et/ou quand il ne satisfait pas aux critères ci-dessus, il sera décomposé en plusieurs vocables.

Cette position pose une série de problèmes :

---

1. André Martinet, "Homonymes et polysèmes", *La linguistique*, 10-2, 1974, p 37-45.

- premièrement, elle rend difficile la comparaison entre des corpus émis à des époques différentes : ce qui sera considéré comme non lexicalisé à un moment donné risque de se figer à une époque ultérieure. En revanche, d'autres compositions peuvent sortir de l'usage.
- on ne peut donner la priorité aux dictionnaires car, suivant celui adopté, on aura des variations importantes : ainsi, le Robert est beaucoup plus "libéral" que le Larousse ;
- de plus, l'époque moderne semble connaître une expansion importante de ces formes, surtout dans les sciences, l'administration, la gestion… Il est possible que les contemporains préfèrent la composition à la néologie pour créer des mots nouveaux (bien que l'on rencontre assez fréquemment, dans ces compositions, des racines empruntées au grec ou au latin). Faut-il admettre tous ces mots ?
- certaines contradictions peuvent apparaître du fait de la position plus rigoureuse que nous avons adoptée concernant le blanc séparateur. Par exemple, cela revient à accepter "allocation-logement" et à rejeter "allocations familiales"…

Nous complétons donc ces principes généraux par quelques règles plus simples d'emploi.

2.23. Les critères de reconnaissance des mots composés

Ces critères sont au nombre de trois :

2.231. La règle "C'est-à-dire"

De la même façon, le *Machinal* retient "c'est-à-dire" alors que les quasi-synonymes de "c'est-à-dire" sont assez nombreux : "en d'autres termes", "ce qui veut dire", "pour le dire autrement". Peut-on alors considérer que 4 vocables sont mobilisés pour former "c'est-à-dire" et que les tirets sont là pour des raisons uniquement phoniques ? Cette solution a l'avantage de la clarté et de la simplicité Mais ici les motifs euphoniques sont indissociables du souci de souligner l'existence d'une locution. Enfin, on voit bien que dans le cas de "c'est-à-dire", les critères proposés par Martinet s'appliquent parfaitement ; notamment la locution est insécable[1]. Dès lors il convient de considérer que "c'est-à-dire" ne forme qu'un seul mot.

2.232. La règle "franco-"

La langue permet de constructions multiples sur le modèle "franco-belge" ; "belgo-luxembourgeois" ; "latino-américain"… Quand deux adjectifs usuels sont accouplés grâce à la liaison "o-" ils constituent une forme unique.

2.233. Les verbes composés

Le cas le plus complexe à résoudre est posé par certains verbes car chacun d'entre eux peut donner naissance à de nombreuses formes. La plupart des préfixes se collent directement devant le verbe sans aucune séparation (re ou ré, pré, anti, multi…).

Les règles énoncées plus haut par Martinet peuvent être complétées par quelques considérations :

- il arrive parfois que les verbes commençant par une voyelle et composés avec un préfixe comme "pré", "post", "re" ou "anti" soient écrits avec un tiret (anti-inflamatoire). Celui-ci doit être supprimé au moment de la saisie…[2]

---

[1] Après avoir adopté cette règle, nous avons recontré chez de Gaulle : "ce n'est pas à dire que…". La locution n'est donc pas tout à fait insécable !

[2]. Voir à ce sujet, Charles Muller, "Le MOT, unité de texte et unité de lexique en statistique lexicologique", (1963), reproduit dans *Langue française et linguistique quantitative*, Paris-Genève, Slatkine-Champion, 1979, p 129.

En revanche certains préfixes sont parfois utilisés avec des traits d'union (contre-, entre-, sous-…). Pour départager ces cas des compositions libres, nous avons établi une courte liste où ne figurent que des verbes construit avec un tiret (ce qui exclut les verbes en plusieurs mots même fortement lexicalisés). Remarques :

- "contre" se compose toujours avec un tiret ;

- "entre" s'agglutine devant une consonne : entredéchirer, entredétruire, entredévorer, entrefrapper, entrehaïr, entreheurter, entrelouer, entrenuire, entreregarder, entretenir…

- "entre" devant une voyelle s'agglutine par suppression de sa dernière lettre et de l'apostrophe : entraider, entraimer, entraîner, entrégorger… sur le modèle de "entraide", "entraîneur"…

- "sous" se compose toujours avec un tiret (sous-entendre, sous-estimer, sous-évaluer, sous-exposer, sous-louer, sous-tendre, sous-traiter…) Il serait naturellement absurde de compter "sous-évaluer" pour deux formes et "surévaluer" pour une seule…

- un nombre limité de cas particuliers - arc-bouter, court-circuiter, pique-niquer, etc - sont récapitulés a priori dans la table des locutions et des mots composés.

2.3. LA TABLE DES LOCUTIONS ET MOTS COMPOSES.

Rappelons que ne sont pas examinés ici les noms propres, les sigles et abréviations. Pour les "noms communs", l'annexe 1 donne la liste que nous avons constituée pour notre propre usage. Nous insisterons sur son caractère a priori et son évolution nécessaire.

2.31. Une table a priori

La table est fixée avant le traitement afin d'éviter que ne se produise des fluctuations dans le traitement des formes.

A la rencontre de tout mot posant problème, l'opérateur consulte la table et vérifie la correction de la graphie. Au cours du traitement, l'ordinateur, lorsqu'il rencontre un blanc, un tiret, une apostrophe commence par consulter la table. S'il n'y rencontre pas la forme, il considère que le signe est un séparateur de mots…

Pour autant nous ne considérons pas cette table comme définitive comme nous l'expliquons ci-dessous.

2.32. Des évolutions possibles

Ce problème de la composition mérite une analyse plus fine et plus approfondie. Dans ce travail, on pourrait s'inspirer des propositions du CLIF[1] : "D'une manière générale, écrit Joseph Hansen, l'agglutination est privilégiée, sauf lorsque la rencontre des voyelles *a* et *i* , *a* et *u* , *e* et *u*, *i*, *o* et *u*, *o* et *eu* ferait obstacle à la prononciation et à la lecture (par exemple intra-utérin, micro-informatique), sauf aussi, tout au moins provisoirement, dans les composés de circonstance, surtout quand ils créeraient des groupes insolites comme *aeu*."

"Les noms composés du type *queue-de-poisson* ou *dos-d'âne*, formés du nom d'une partie du corps ou de vêtement, d'une préposition et du nom d'un être vivant prennent un ou des traits d'union lorsque l'ensemble du composé est employé métaphoriquement."

"Pour un certain nombre de mots formés par un redoublement de syllabe (onomatopées, formations expressives), il y a hésitation et même incohérence en ce qui concerne l'agglutination et la forme du pluriel. On a tâché d'uniformiser le traitement de ces cas en privilégiant l'agglutination avec l'accord du pluriel"[2] (cricri, cricris ; grigri, grigris ; flafla, flaflas).

*
* *
*

1. L'ouvrage déjà cité sur l'harmonisation ortograhique est malheureusement paru après l'élaboration de notre table et le début de nos traitements : nous n'avons pas pu en tenir compte comme il aurait fallu.

2. Conseil international de la langue française, *op. cit.*, p 7-8.

En conclusion, les conventions exposées dans ce chapitre nous amènent à exclure des formes composées - ayant une entrée dans la plupart des dictionnaires contemporains - comme : "aide de camp" (mais nous admettons aide-mémoire !), "apprenti sorcier", "ayant droit", "bon marché", "bon vivant", "chemin de fer", "compte rendu", "congé payé", "écart type" (mais "contrat-type" !), "état civil", "fondé de pouvoir", "libre pensée", "main d'oeuvre", "maître chanteur" (mais "maître-hôtel" !), "mère patrie", "mot clef", "nature morte", "pomme de terre", "poids lourd", "porte à porte", "repris de justice"…[1]

Comme on peut le voir, l'application de nos règles entraîne d'apparentes inconséquences qui tiennent aux incohérences de la graphie du français. Par exemple elles conduisent à compter "au-dehors" ou "au-delà" pour une forme et "en dehors" ou "en delà" pour deux (sous réserve de ne pas leur appliquer la règle "d'accord")… Ou encore, on trouve en une forme : "chef-d'oeuvre", "hors-d'oeuvre", "libre-échange", "libre-échangiste" mais en deux : "main d'oeuvre", "libre pensée", "libre penseur" (qui ont une entrée dans le Robert par exemple). On remarquera que "main d'oeuvre" semble être retenu par Saint-Cloud comme une seule forme (*machinal* p 40). Voir aussi tous les composés de "assurance" mais exclusion de "assurances sociales" (qui s'écrit sans trait d'union)… Tous ces exemples trouvent leur origine dans des inconséquences de la langue française. Or nous ne pouvons trop nous affranchir de la forme graphique ni multiplier les sources d'erreur : l'opérateur doit pouvoir recourir à des règles simples et consulter des listes point trop longues.

---

1. Nous avons parfois rencontré certains de ces mots avec un trait d'union notamment : compte-rendu, porte-à-porte…

CHAPITRE III
LE TRAITEMENT ET L'ANALYSE DES FORMES.

Les traitements utilisés lors de cette première étape seront baptisés "Saint-Cloud". Ces traitements obéissent à un certain nombre de principes généraux que nous présentons brièvement ci-dessous.

A l'issue de cette première phase, on obtient un fichier contenant N *mots* — N étant la *taille* du fichier — ou *formes graphiques* ou *types* différents. Un certain nombre de traitements peuvent être effectués sur ces mots et ces formes.

3.1. LE TRAITEMENT INFORMATIQUE DES FICHIERS-TEXTE

Les principes généraux auxquels obéissent ces traitements sont communs aux deux parties. Nous les exposons succinctement dans cette section.

3.11. Les principes généraux du traitement informatique

Nous sommes actuellement dans une phase expérimentale où il est indispensable de pouvoir intervenir sur telle ou telle instruction des programmes ; voire de revenir en arrière en modifiant telle ou telle convention. C'est pourquoi, d'une manière générale, nous avons décidé de découper les traitements en plusieurs phases.

A chacune de ces phases correspond une séquence de programmes dont l'enchaînement peut être interrompu à tout moment par l'opérateur. Celui-ci reçoit sur l'écran tous les messages nécessaires au suivi des opérations.

A l'issue de chacune de ces phases, un fichier transitoire est créé qui servira de fichier de travail pour la phase suivante et qui sera détruit à la fin de celle-ci. Dans des sorties papier, le programme indique l'état du texte et récapitule les opérations effectuées, les difficultés rencontrées et les solutions adoptées par l'opérateur. Ceci permet de surveiller l'exécution et de localiser les difficultés. Une faute se trouvera donc détectée rapidement et ne risquera pas de se répercuter loin en aval.

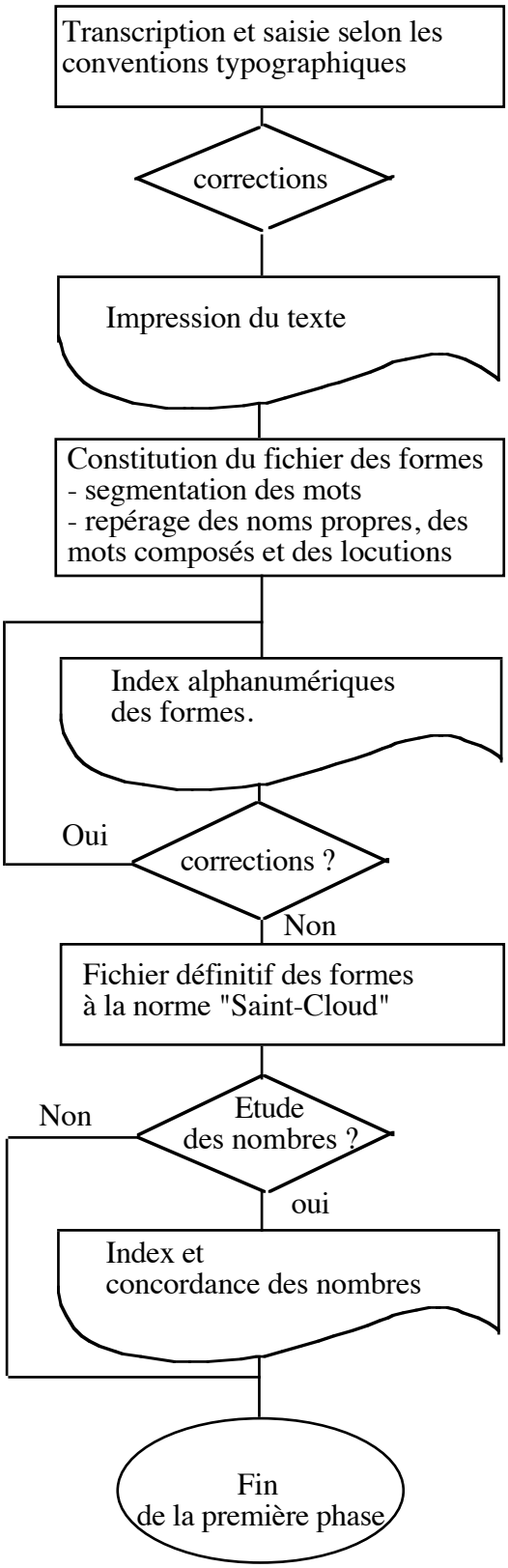
Les informations sur le déroulement des opérations sont également utiles pour juger de l'efficacité des traitements d'un double point de vue : le souci de parvenir à un "zéro-défaut" dans la fabrication des fichiers et la volonté d'augmenter au maximum la part de la machine dans ces opérations pour en accroître la sécurité et limiter les erreurs…

3.12. Les traitements issus de la norme Saint-Cloud

Nous donnons ci-dessous une description synthétique des programmes informatiques traduisant en instructions les conventions exposées dans les deux chapitres précédents (cf le schéma récapitulatif dans la page suivante).

Le premier paquet de programmes segmente les textes en mots.

Tableau 2. Les principales phases de l'analyse des formes



3.121. Le découpage des mots

A l'aide d'une table de noms propres, les majuscules sont conservées pour les seuls noms propres. En cas d'ambiguïté, le programme interroge l'opérateur. Le programme commence par les noms propres afin de ne pas réduire les noms à particules ou composés. Les autres mots à majuscules sont convertis en minuscules (une table permet de stocker les plus fréquents). En cas d'absence dans les tables, le programme interroge l'opérateur. A chaque fois, il est proposé à l'opérateur d'ajouter le mot aux tables concernées : ainsi le programme se construit lui-même…

A l'aide d'une table des locutions et des mots composés, le programme établit l'unité des "formes agglutinées" (à ce sujet voir la note sur les locutions et les mots composés). Un traitement spécial est effectué sur les verbes type "sous-tendre" "arc-bouter", etc

3.122. Le découpage des phrases

Dans cette première phase, on conserve la ponctuation et on introduit la notion de "ponctuation majeure" qui conclut une période oratoire. Pour améliorer l'exécution du programme, il a été décidé de réaliser cette interprétation à la saisie en suivant la convention typographique qui s'est maintenant affirmée :

- si le premier mot à droite du signe de ponctuation est en majuscule sans être un nom propre, une nouvelle phrase commence ;
- si le premier mot à droite du signe de ponctuation est une minuscule, la même phrase se poursuit.

Pour une transcription de l'oral, il est donc nécessaire d'opérer ce choix au moment de la saisie. Cette solution est préférable car elle donne le temps de la réflexion et d'une analyse détaillée de la phrase (cf ci-dessus § 1.1).

3.123. Le traitement des nombres

A l'issue de cette seconde étape, l'opérateur dispose du fichier définitif des formes sous un double format :

- dans le premier, l'ensemble des nombres sont transcrits en chiffres. Dans ce cas "1990" est compté comme une seule forme. Une étude des nombres peut alors être menée ;
- dans le second, les nombres sont convertis en lettres : "1990" devient "mille neuf cent quatre vingt dix" soit six mots. Un programme assure cette conversion pour éviter les erreurs.

Au cours de chacune de ces trois étapes, le programme édite un certain nombre de renseignements concernant les opérations effectuées, les difficultés rencontrées, les choix opérés par l'opérateur. Ces archives sont indispensables pour contrôler la stabilité de la norme de dépouillement tout au long des opérations.

3.2 L'INDEXATION DES TEXTES ET LEUR CORRECTION

Le fichier des formes est donc réalisé. Sur ce fichier deux ensembles d'opérations peuvent être menées à bien : la constitution d'index, les ultimes corrections et l'établissement de certaines concordances nécessaires à la lemmatisation qui sera présentée dans la seconde partie de cette note.

3.21. Les index

On établit deux index. Chacun de ces index existe en une version complète, consultable sur support magnétique ("index-électronique"), et en une version papier qui peut être synthétique par souci d'économie et de commodité ("index-papier").

3.211. L'index alphabétique

Dans l'index alphabétique, les formes sont classées par ordre alphabétique selon les conventions habituelles. On remarquera que ces conventions ne correspondent pas aux résultats obtenus en faisant appel à la fonction de tri sur chaînes de caractères offerte par les compilateurs standards. Ceux-ci suivent tous l'ordre de la table ASCII[1]. Dès lors un programme spécifique de tri doit être élaboré pour respecter les conventions lexicologiques et faciliter la consultation.

• L'index électronique

Dans un corpus de plusieurs textes, la présentation de l'index électronique doit suivre un certain nombre de règles :

- si les textes composant le corpus sont de grandes dimensions, une version de l'index doit comporter la localisation exacte de chaque forme. Par exemple, le numéro d'ordre des occurences dans le texte ; à défaut, le numéro du chapitre, de la page, du paragraphe…

- si les textes composant le corpus sont de dimensions relativement réduites (n'excédant pas 10 à 15 pages dactylographiées), on peut se contenter de relever pour chaque texte, la fréquence de la forme. En effet, tous les traitements de textes comportent aujourd'hui une fonction de recherche de chaîne de caractères qui peut permettre de retrouver rapidement un mot dans un texte.

• L'index-papier

En cas de gros corpus, un index-papier complet (comportant mention de la localisation de chaque occurrence) prend vite des proportions considérables et sa consultation devient malaisée. Deux solutions permettent de le réduire :

- si le corpus est composé de textes pas trop inégaux, on peut se contenter de la mention de la fréquence avec le numéro des textes où la forme est attestée ;

- si le corpus est composé de textes de longueurs très inégales, ou si ses dimensions excluent la solution ci-dessus, on aura recours à l'indice de répartition qui complétera l'indication de la fréquence. Pour une présentation détaillée de cet indice on se reportera à l'annexe 7 placée à la fin de cette note.

Ce premier index sert d'abord à une ultime relecture : le rapprochement alphabétique permet de détecter des fautes de frappe, des graphies erronées ou douteuses…

3.212. L'index hiérarchique

Le tri est effectué sur le nombre d'occurrences et les formes sont présentées par ordre décroissant de fréquence jusqu'à un seuil fixé a priori. Cet index est moins important que l'alphabétique mais il présente deux intérêts :

- la comparaison des fréquences avec celles observées dans d'autres corpus - ou entre les textes au sein du corpus - permet déjà de mettre en valeur certaines formes dont la présence paraît caractéristique. Une première recherche des spécificités peut également être conduite dès cet instant. Mais ses résultats ne pourront valablement être interprétés qu'après la lemmatisation.

---

1. Par exemple, dans le *Dictionnaire des fréquences* édité en 1971 par le Centre pour un trésor de la langue française de Nancy. Dans ce dictionnaire, les lettres accentuées viennent en fin de classement après le "z"…

La mention de la répartition - soit par numéro de texte et sous fréquence ou à défaut par l'indice de répartition - facilite également un premier repérage du vocabulaire spécifique.

- l'index hiérarchique permet d'anticiper sur la lemmatisation : on peut s'assurer que les cas d'homographie dépassant un certain seuil de fréquence sont bien prévus dans les programmes.

Après les ultimes corrections sur les textes, le fichier définitif est établi.

A partir des premières constatations, on choisit les formes dont l'étude mérite d'être approfondie sans attendre la lemmatisation : on éditera leurs concordances

3.22. Les concordances

Nous ne présenterons pas ici de manière détaillée la technique des concordances[1] que nous employons aussi bien sur les fichiers à la norme Saint-Cloud qu'à la norme Muller. L'étude des concordances sur les fichiers Saint-Cloud permettent de résoudre les difficultés attachées à la lemmatisation certaines formes ambiguës dont l'analyse ne peut être conduite automatiquement.

Le format actuellement retenu tient compte de la nécessité de respecter les 80 colonnes de la ligne d'imprimante et fournit les références concernant le mot recherché. On recherche un contexte significatif : au minimum 4 mots devant et derrière sauf si la forme recherchée se trouve au début ou à la fin d'un texte ou sur toute autre césure. Dans ces cas, seul le contexte significatif sera pris en compte.

Voici par exemple un extrait des concordances du mot "cohabitation" dans les discours de F. Mitterrand :

…

45, 2586	sais quelle forme de	**cohabitation**	politique mais dans une
51, 6321	du gouvernement. La	**cohabitation**	est un art difficile mais
51, 6287	depuis hier. Donc, la	**cohabitation**	, il ne faut pas la charger
51, 7018	our cela que j'ai dit : "	**cohabitation** "	à mon tour - je crois q
53, 1435	parlez de quoi ? De la	**cohabitation**	entre la France et l'Espa

…

Le premier chiffre donne le numéro d'ordre du discours. Par exemple, le numéro 45 correspond à l'émission "Ça nous intéresse, monsieur le président" du 2 mars 1986 et "cohabitation est le 2586e mot prononcé par le président depuis le début de l'émission…

On pourra ainsi éditer les concordances des formes d'homographie absolue de type "suis" (verbes être et suivre à la première personne du singulier de l'indicatif présent et de l'impératif) ou "garde" (verbe garder, substantifs masculin et féminin) dont l'analyse automatique n'est pas actuellement concevable. Nous avons ainsi un puissant outil d'aide à la lemmatisation, opération à laquelle nous consacrons notre seconde partie.

---

1. Sur ce point voir notamment Majid Sekhraoui, *op. cit.*, p 90-147.

CONCLUSION DE LA PREMIERE PARTIE

La "norme Saint-Cloud" n'est donc pas tout à fait fidèle à la "forme" entendue en son sens strict. En particulier, la prise en compte de certains phénomènes de lexicalisation introduit une faiblesse. Cependant, les défenseurs les plus intransigeants de la "forme" vont plus loin que nous dans cette voie : ainsi, dans le *Machinal*, il est proposé de faire une seule forme de "main d'oeuvre" en y ajoutant un tiret ("main-d'oeuvre"). Dans le chapitre sur les locutions et mots composés, nous avons exposé les raisons pour lesquelles il ne paraît pas possible d'aller aussi loin dans cette voie hardie mais peut-être inconséquente. D'une part, on viole les conventions graphiques du français - ce qui amène notamment des risques supplémentaires d'erreur - et, d'autre part, on fige ensemble deux termes qui ont leur vie propre. Comme on le voit dans cet exemple, aucune des règles de composition – "bric à brac", "parce que", "d'abord" – ne peuvent s'appliquer…

Cette phase est extrêmement utile.

De sa qualité dépend d'abord la fiabilité des fichiers. Le niveau "zéro-faute" est extrêmement difficile à atteindre ; il est pourtant indispensable. En effet, nous disposons maintenant de modèles statistiques puissants. Mais on a parfois le sentiment que la charrue a été placée avant les boeufs puisqu'il est difficile d'apprécier la solidité des données qui font l'objet de ces calculs…

A l'issue de ce travail, on obtient une sorte de "dictionnaire des formes". Etant donné l'aspect encore très "bricolé" des normes de lemmatisation, de tels "dictionnaires" demeurent indispensables ne serait-ce qu'à titre de contrôle…

Les fichiers à la norme Saint-Cloud permettent une recherche aisée des concordances qui sont une aide précieuse à la lemmatisation. Au-delà de cet aspect pratique, la comparaison des index de formes et de vocables devrait permettre de juger de l'impact de la norme de lemmatisation : on connaît l'état du texte avant l'opération. Une fois la lemmatisation effectuée, il deviendra possible de comparer et d'évaluer les principaux changements que l'opération a amenés dans le vocabulaire du texte.

Au-delà de ces aspects techniques, la double codification - Saint-Cloud et Muller qui va être exposée dans la seconde partie de cette note -, cette double codification est indispensable pour tester à fond les modèles élaborés pour rendre compte du vocabulaire, du lexique et de la langue. On pourra ainsi juger de leur universalité ou, à l'inverse, établir leurs limites de validité et délimiter leur champ d'application.

DEUXIEME PARTIE
LA LEMMATISATION DES TEXTES.

> "Il se peut qu'il n'y ait pas d'autre métalangue que celle qui, depuis fort longtemps et dans de nombreuses cultures, est disponible pour le simple écolier, à savoir l'ensemble des termes techniques de la grammaire, comme, en français, *singulier*, *première personne*, *préposition*, *adjectif*, *subordonnée*, etc (…)"
> "L'existence à peu près universelle, au moins dans les cultures qui possèdent une tradition grammairienne, de lexiques métalinguistiques contenant des termes comme ceux qu'on vient de citer atteste que depuis longtemps, il s'est trouvé des individus pour ressaisir en conscience le déroulement inconscient d'une démarche aussi naturelle que de parler, et pour en faire l'objet d'un discours ordonné, c'est-à-dire pour adopter à l'égard de la langue une vue scientifique."
> (Claude Hagège, *L'homme de parole*, p 289-290)

La seconde étape de nos traitements consiste à rattacher à un *vocable* chacun des mots découpés dans le texte lors de la première étape. Pour ce faire, on regroupe sous une entrée unique les formes correspondant à un même signifiant (par exemple : "mot" et "mots" ou encore "le", "la", "les", "l'", etc…). Le vocable sera symbolisé par un *lemme* ou encore *forme canonique* comportant notamment l'indication de sa catégorie grammaticale. Par exemple : "mots" est rattaché sous "mot, substantif masculin". Ainsi toutes les flexions d'un même verbe sont-elles regroupées sous l'infinitif ; les variations de genre et de nombre des adjectifs sous le masculin singulier, etc. C'est la *lemmatisation* que nous avons effectuée en suivant au plus près les principes énoncés par C. Muller[1]. Certes le travail est long - et comporte des risques d'erreur - mais il est indispensable pour trois raisons que nous présentons dans notre quatrième chapitre.

La lemmatisation est une opération délicate qui suit plusieurs phases. Le quatrième chapitre évoque succinctement ces étapes (reconnaissance des formes, détection automatique et résolution assistée des homographies, constitution d'un index lemmatisé). Les principales difficultés rencontrées proviennent de la pluralité des fonctions remplies par une même forme ("homographies") et par les frontières floues séparant les catégories de la grammaire traditionnelle sur lesquelles nous nous appuyons.

Le cinquième chapitre passe en revue les problèmes de la lemmatisation du verbe et présente les principales solutions retenues pour analyser certaines ambiguïtés attachées à cette catégorie. Celles-ci surviennent notamment à cause de la plasticité du verbe et du rôle central que jouent dans la créativité lexicale, les participes et l'infinitif.

Le sixième chapitre examine la lemmatisation du groupe nominal. Celui-ci comporte généralement, outre un substantif et un ou des adjectifs, des "déterminants" (articles, numéraux

---

1. Cf notamment Charles MULLER, "Le MOT, unité de texte et unité de lexique en statistique lexicologique", (1963), reproduit dans *Langue française et linguistique quantitative*, Paris-Genève, Slatkine-Champion, 1979, p 125-144. Voir également la note 3 page 11 de notre introduction.

et cardinaux, adjectifs indéfinis). La reconnaissance des formes du groupe nominal et la résolution des homographies propres à ce groupe sont également passées en revue.

Le septième chapitre traite de la lemmatisation des "mots invariables" (prépositions, conjonctions, adverbes).

CHAPITRE IV
PRINCIPES GENERAUX ET ORGANISATION DE LA LEMMATISATION

La lemmatisation consiste donc à rattacher chaque mot du texte à une forme canonique et à une catégorie grammaticale. Cette opération a été vivement critiquée dans le passé. On lui reprochait notamment de substituer au code qui a présidé à la création du texte, des conventions arbitraires (ce qu'on imagine être le code de la langue et qui n'est qu'un artifice de grammairien). On a opposé à la "statistique lexicale" une "lexicométrie des formes" plus respectueuse du texte[1]. En sens contraire, les défenseurs de la lemmatisation mettent l'accent sur le caractère peu maniable des index obtenus par les "formalistes"[2]. Nous ne prétendons pas trancher ce débat. La lemmatisation paraît nécessaire et comporte de nombreux intérêts à condition qu'elle n'écrase pas les "formes" c'est-à-dire le texte transcrit dans les conventions graphiques du français écrit…

4.1. NECESSITE ET INTERETS DE LA LEMMATISATION.

Nous examinerons successivement les raisons qui militent pour la lemmatisation des corpus puis les bénéfices que peuvent en retirer les lexicologues.

4.11. La nécessité d'une lemmatisation

Nous montrerons tout d'abord que la lexicométrie travaillant sur des formes hors contexte a nécessairement besoin de lever l'ambiguïté pesant sur certaines formes. Plus impérieusement, le lexicologue, quand il fabrique des index, doit se plier à certaines conventions et ne peut le faire qu'au prix d'une lemmatisation.

4.111. La confection d'index

La commodité d'utilisation de l'index n'est pas la justification la moins impérieuse : dans un dictionnaire, aurait-on l'idée de chercher des indications sur le verbe "aller" à la lettre "v" ? C'est pourtant ce qu'exigent certains auteurs d'index qui classent les formes par ordre alphabétique. On pourrait multiplier les exemples de ces difficultés de consultation des index de formes en commençant par les deux verbes les plus fréquents de la langue française (être et avoir) dont les flexions sont dispersées… Il s'agit avant tout de respecter les conventions qui se sont peu à peu ancrées dans les habitudes. Par exemple, nous faisons inconsciemment cette lemmatisation à chaque fois que nous consultons un dictionnaire. C'est pourquoi il est indispensable de se tenir au plus près des conventions admises en lexicologie malgré leurs faiblesses voire leurs incohérences… Or la lexicométrie des formes ne peut répondre à cette exigence.

---

1. Les arguments en faveur de la non-lemmatisation ont été notamment développés dans Michel Demonet, Annie Geoffroy et Al, *Des tracts en mai 68*, Presses de la Fondation nationale des sciences politiques, 1973, p 19-39. Voir aussi : Maurice Tournier, "Sur quoi pouvons-nous compter ?", *Verbum*, 1985.

2. Voir par exemple, la préface de Charles Muller au livre de Pierre Lafon, *Dépouillements et statistiques en lexicométrie*, Paris-Genève, Slatkine-Champion, 1984.

4.112. Les difficultés de la lexicométrie hors contexte

Pour éclairer ces difficultés, prenons une phrase prononcée par F. Mitterrand lors d'une de ses interventions télévisée. Elle comporte 36 mots soit environ la moyenne des phrases du premier septennat :
"Les limites coloniales, ayant été considérées comme un fait acquis, devaient devenir une base de droit dans les relations entre les Etats : tel était le cas de la bande d'Aozou qui fait partie du Tchad"

Apparemment il n'y a aucun mot ambigu dans cette phrase simple. Si on bouleverse l'ordre des mots pour adopter l'ordre alphabétique, 23 des 36 mots deviennent ambigus :

- "acquis" : adjectif ou verbe (acquérir au participe passé ou au passé simple)
- "bande" : nom féminin ou verbe (bander)
- "base" : nom féminin ou verbe (baser)
- "considérées" : adjectif ou verbe au participe passé
- "de" (3 occurrences) : préposition ou partitif
- "devenir" : nom masculin ou verbe à l'infinitif
- "droit" : adjectif ou nom masculin ou adverbe
- "entre" : préposition ou verbe (entrer)
- "été" : nom masculin ou verbe être au participe passé
- "fait" : nom masculin ou verbe faire (indicatif présent ou participe passé)
- "la" : article ou pronom
- "le" : article ou pronom
- "les" (3 occurrences) : article ou pronom
- "limites" : nom féminin ou verbe (limiter)
- "partie" : nom féminin ou verbe (partir au participe passé)
- "relations" : nom féminin ou verbe (relater, première personne du pluriel)
- "tel" : adjectif indéfini ou pronom
- "un" : pronom ou article ou numéral
- "une" : pronom ou article

Nous insisterons sur deux points. D'une part, sans le contexte, il est impossible de départager les sens possibles : la "lexicométrie des formes" travaillant généralement hors contexte accumule donc les ambiguïtés. D'autre part, cet exemple n'a rien d'exceptionnel. Dans le corpus Mitterrand, 40% des formes en moyenne - si l'on compte "de" - présentent des difficultés de ce genre… Naturellement, il reste 60% de formes utilisables. Mais là encore les index se révèlent inadaptés notamment à cause de la dispersion des flexions des verbes comme nous l'avons montré ci-dessus.

4.113 La résolution des homographies

On l'aura compris les homographies doivent être dénouées d'une manière ou d'une autre. Les partisans de la lexicométrie des formes semblent d'accord sur ce point. Ainsi, rendant compte de l'ouvrage d'Etienne Brunet sur *Le vocabulaire français depuis 1789*[1], Annie Geoffroy et Pierre Lafon citent ces propos de Brunet : "On ne trouverait rien à redire si quelqu'un séparait les formes homographes tout en refusant la lemmatisation" et ajoutent "cet idéal est aussi le nôtre". Mais Annie Geoffroy et Pierre Lafon poursuivent : "Faute de pouvoir l'atteindre, nous préférons le pis-aller des formes graphiques"[2].

---

1. Genève-Paris, Slatkine-Champion, 1981.

2. Annie Geoffroy et Pierre Lafon, "L'insécurité dans les grands ensembles", *Mots*, 5, octobre 1982, p 129.

En tant qu'usager de ces formes graphiques, nous sommes pourtant obligés de constater que leur intérêt est relativement maigre. Imagine-t-on sérieusement un lexique de la politique française qui ne distinguerait pas entre l'infinitif du verbe "pouvoir" et le substantif masculin de même graphie ? Aujourd'hui un lexique se fabrique aussi à l'aide d'index et de concordances… Or, la phrase de Mitterrand citée ci-dessus l'aura suggéré : l'homographie est au coeur de la langue française. Au début des années 70, l'équipe du Trésor de la langue française de Nancy[1] avait recensé plus de 4000 homographes (n'appartenant pas à la même catégorie grammaticale) et depuis on en a découvert beaucoup d'autres qui avait échappé à ce relevé pourtant effectué sur plusieurs dizaines de millions de mots. Les annexes de cet ouvrage donnent la liste de ces homographes pour ce qui concerne le verbe.

4.114. La distinction entre les différentes fonctions d'une même forme

La *pluralité des fonctions* que peut remplir une même forme est donc la raison essentielle qui milite en faveur de la lemmatisation. En effet, l'utilisateur d'index s'attend naturellement à voir distingués les synonymes homographes : qui s'aviserait de confondre l'adverbe "bien" avec le substantif ? De même il n'y a aucune différence syntaxique entre "un pays développé" et "un pays moderne". Si l'on trouve "moderne" sous l'adjectif pourquoi faudrait-il aller chercher l'emploi adjectival de "développé" sous le verbe "développer" alors que notre sentiment linguistique nous a fait reconnaître un adjectif dans ce participe passé ? Il en est de même pour des formes comme "immigré", "mort" ou "entreprise"… Il faut admettre qu'une même forme peut appartenir à plusieurs catégories grammaticales et que ces emplois emportent des *variations sémantiques* dont un index doit, dans la mesure du possible, rendre compte…

4.12. La qualité de la lemmatisation

Cependant les critiques des formalistes n'ont pas été inutiles. Elles obligent à constater que la lemmatisation d'un texte doit obéir à certains critères de qualité. En premier lieu, elle doit s'ajouter au texte d'origine sans le détruire. Pour désigner cette opération, nous proposons de parler d'une "lemmatisation dans le texte" : le mot reste à sa place et ne subit aucune transformation. On ajoute au document original un texte parallèle et où, à chaque mot, correspond un lemme et un code grammatical.

4.121. Le respect du texte d'origine

Il est important de comprendre que la "lemmatisation dans le texte" ne fait pas disparaître le texte d'origine. Au contraire, elle en permet la lecture : les traitements peuvent s'opérer concurremment sur les formes et les lemmes. De même, l'index comporte mention des deux les unes sous les autres : le lemme est la porte d'entrée naturelle et la forme apporte une information complémentaire qui n'est pas négligeable et ne doit donc pas être perdue. Autrement dit, une bonne lemmatisation ouvre une nouvelle voie d'accès au texte sans le mutiler.

On trouvera dans le § 4.23 de ce chapitre un exemple de texte lemmatisé qui permettra de comprendre en quoi consiste exactement cette opération.

C'est ce principe qui nous a amené à nous séparer de C. Muller sur un point : le traitement des "l'" et "t-" euphoniques (l'on dit, où va-t-il"). C. Muller préconise leur élimination puisqu'ils n'ont aucun contenu sémantique. Mais cette suppression n'étant pas réversible, il n'était plus possible de retrouver le texte original à partir de sa version lemmatisé…

---

1. Centre de recherche pour un trésor de la langue française, *Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siècle*. (IV. Table de répartition des homographes), Nancy, 1971.

4.122. Les principes de construction de la nomenclature

Il reste à définir avec le maximum de précision les catégories grammaticales pour éviter obscurités, chevauchements et contradictions. Plusieurs considérations plus ou moins complémentaires ont guidé nos choix. La norme devra avoir les qualités suivantes :

- elle comporte un minimum de cas particuliers ou d'exceptions, ceci afin d'augmenter le nombre des possibilités de résolution automatique et pour diminuer les risques d'erreurs. Selon la formule de Muller, on recherche une norme de dépouillement qui soit "très analytique en ce qui concerne la délimitation du mot et très synthétique pour celle du vocable"[1]. Par exemple, on laisse pratiquement de côté le problème de l'homonymie-homographie dont on connaît de nombreux exemples…[2] En effet, si l'on voulait donner autant de lemmes qu'il y a de sens à un mot, on risquerait de multiplier à l'infini les entrées lexicales et les difficultés. La solution retenue consiste à analyser les homographies quand celles-ci mettent en jeu des mots appartenant à des catégories grammaticales différentes. En limitant le nombre de cas particuliers, on permet un traitement de la "macro-structure" ;
- la commodité d'utilisation de l'index nous conduit à travailler au plus près des catégories utilisées dans le "français standard"[3]. Seule "innovation" aujourd'hui largement admise : la notion de *déterminant* qui regroupe les articles, les numéraux et cardinaux et les adjectifs "indéfinis" ;
- elle est exhaustive. Elle ne laisse pas dans le fichier de forme non interprétée ;
- elle est sans ambiguïté : tous les cas doivent être résolus et, au départ, il faut s'interdire la pratique consistant à admettre un "résidu" de classifications non correctes ;
- elle est automatisée autant que faire se peut : les cas non résolus par la machine, sur lesquels l'opérateur sera interrogé, doivent être aussi peu nombreux que possible et il faut prévoir des règles simples de résolution (voir règle n° 1) ;
- en cas d'appel à l'opérateur, celui-ci doit suivre des règles objectives d'analyse et ne pas se fier à son "sentiment linguistique"…

4.123. Stabilité, lisibilité, reproductibilité du dépouillement

Nous ajouterons enfin qu'une norme de dépouillement doit répondre à trois critères essentiels : stabilité, lisibilité, reproductibilité :

- stabilité : la norme ne doit pas varier en cours de traitement pour garantir l'homogénéité du matériel sur lequel on va travailler ;
- lisibilité : les procédures comme les résultats doivent être accessibles au non-spécialiste et ré-utilisables ;
- reproductibilité : les chercheurs doivent pouvoir utiliser à leur tour l'instrument si celui-ci leur paraît adapté à leurs objectifs… Telle est la raison qui nous conduit à diffuser ce document malgré ses imperfections et ses lacunes.

---

1. Charles Muller, *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse, 1967, p 29.

2. Voir à ce sujet, Charles Muller, "Polysémie et homonymie dans l'élaboration du lexique contemporain", *op cit*, p 33-38.

3. C'est-à-dire, outre les index déjà cités, essentiellement : R. L. Wagner et J. Pinchon, *Grammaire du français classique et moderne*, Paris, Hachette ; J.-C. Chevallier et C. Blanche-Benveniste, *Grammaire du français contemporain* ; Maurice Grévisse, *Le bon usage*, Gembloux, Duculot, (réed 1986). Enfin, le dictionnaire *Petit Robert* en donnant la préférence à ce dernier en cas de contradiction notamment sur le partage adverbe/préposition.

4.124. Les limites actuelles de la "norme Muller"

La "norme Muller", sous sa forme actuelle, achoppe sur quelques difficultés majeures :

- le nombre de vocables diffère du nombre de formes du fait de la décomposition des formes contractes (du, des, aux, duquel, desquels auxquels, etc). Cette décomposition en plusieurs vocables peut difficilement être évitée comme nous le montrerons plus bas ;

- les participes passés et présents adjectivés et/ou substantivés. Mêmes problèmes que ci-dessus. Suivant les époques, on retrouvera la même forme soit sous le verbe, soit avec une entrée propre…

- de manière plus générale, faut-il systématiquement départager les formes outils à haute fréquence. Faut-il distinguer entre "le" (article) et "le" (pronom) ou entre "que" (conjonction et pronom), etc ? Alors même qu'on les rencontre pratiquement dans toutes les phrases… C. Muller y renonce dans son dépouillement de Corneille. En revanche, C. Bernet le réalise presque complètement. Quant à G. Engwall, elle a adopté une catégorie "adjectif pronominal" qui semble avoir pour principale consistance de contourner en bonne partie ce problème.

Si l'on décide de traiter ces difficultés, doit-on choisir d'interroger l'opérateur pour les quelques cas douteux ou bien accepter, pour ces quelques formes à hautes fréquences, une proportion minime d'échecs à condition de contrôler ce taux ? Contre cette dernière attitude on opposera le caractère surtout syntaxique du programme d'analyse automatique : par exemple une erreur dans l'analyse d'un "le" peut se répercuter sur l'analyse de l'homographe {verbe-substantif} qui le suit…

- enfin, C. Muller recommande de se reporter au *Dictionnaire général de la Langue française* de Darmersteter, Hatzfeld et Thomas. Ce dictionnaire a l'avantage de comporter des entrées plus synthétiques que la plupart des ouvrages équivalents. Mais, réalisé au tournant du siècle, il ne peut prendre en compte les nombreuses créations lexicales du XXe siècle. Ceci permet d'ailleurs de comprendre pourquoi le renvoi à un dictionnaire est un pis-aller : il n'est vraiment valable que pour l'étude des textes antérieurs ou contemporains à sa réalisation. Mieux vaut définir des conventions et des normes, comme celles que nous avons esquissées pour le "mot" dans la première partie de cette note.

Dès lors, il faut convenir que la norme de lemmatisation est encore en devenir.

4.2 LES PRINCIPALES CARACTERISTIQUES DE LA LEMMATISATION

4.21. La reconnaissance des formes dans le texte

Jusqu'à maintenant les lemmatiseurs opéraient sur les index en s'aidant des concordances. De ce fait leur lemmatisation n'était pratiquement d'aucune utilité en cas de "retour au texte", pour l'examen des contextes de tel ou tel mot dans une optique lexicologique. Telle est la motivation essentielle de la "lemmatisation dans le texte" (cf l'exemple donné dans le paragraphe suivant).

Cette opération comporte quelques inconvénients. Notamment, on perd l'avantage méthodologique de la lemmatisation sur index : tous les problèmes de même nature sont résolus d'un coup, leur rapprochement permet d'étalonner des règles ad hoc. Cela explique sans doute le fait que les lemmatiseurs se sont généralement contentés d'un dictionnaire et de considérations assez lacunaires sur les méthodes ;

En contrepartie l'opération présente plusieurs avantages :

- elle offre une nouvelle porte d'entrée dans le texte. Par exemple, on pourra enfin rechercher toutes les occurrences - et donc tous les contextes - des substantifs "pouvoir" ou "savoir" à l'exception des verbes à l'infinitif…

- au-delà de cet intérêt immédiat, s'ouvre la possibilité de rechercher les champs lexicaux des vocables à fortes fréquences, les verbes les plus usuels, etc. Le travail sur le lemme permet de donner une nouvelle puissance aux calculs de spécificités, etc…

4.22. Les étapes de la lemmatisation.

Nous résumons dans le graphique ci-contre les étapes de la lemmatisation assistée par ordinateur qui a été mis au point pour l'analyse de notre corpus Mitterrand.

Phase 1.

Le texte à la norme Saint-Cloud est lu par la machine. Pour chaque forme, trois étapes sont possibles :

- la forme est d'abord confrontée à une table générale contenant tous les vocables exceptés les verbes. Si la forme se trouve dans la table, outre la catégorie grammaticale, le programme ajoute les indications sur le genre et le nombre. On passe à la forme suivante ;
- la forme ne se trouve pas dans la table générale ou bien une homographie avec la catégorie du verbe est mentionnée. Dans ce cas, le programme met en oeuvre une procédure de reconnaissance des verbes (décrite en détail dans le § 5.13)
- la forme demeure inconnue : l'opérateur est interrogé. Sa réponse est stockée en mémoire. A la fin des opérations, le programme propose de compléter (ou de corriger) la table concernée par le manque : ainsi les tables se complètent-elles progressivement.

A l'issue de cette première phase, chaque forme du texte se voit associer un ou plusieurs lemmes contenant toute l'information pertinente pour l'analyse syntaxique des homographies menée dans la phase suivante Par exemple le mot "bande", dans la phrase de F. Mitterrand donnée en exemple tout à l'heure :

"bande" *homographe (substantif singulier féminin"bande")-(verbe "bander", première ou troisième personne de l'indicatif présent ou du subjonctif présent ou de l'impératif)* ;

Phase 2. Résolution automatique des homographies

Le programme lit le texte phrase par phrase et tente de résoudre les homographies par analyse du contexte étroit. Par exemple, dans le cas de "bande", il détermine avec certitude qu'il s'agit d'un nom :

- "la" article supposé est de même genre et de même nombre que "bande", substantif supposé ;
- la préposition "de" qui précède "la" ne peut introduire que des verbes à l'infinitif ce qui n'est pas le cas de "bande".

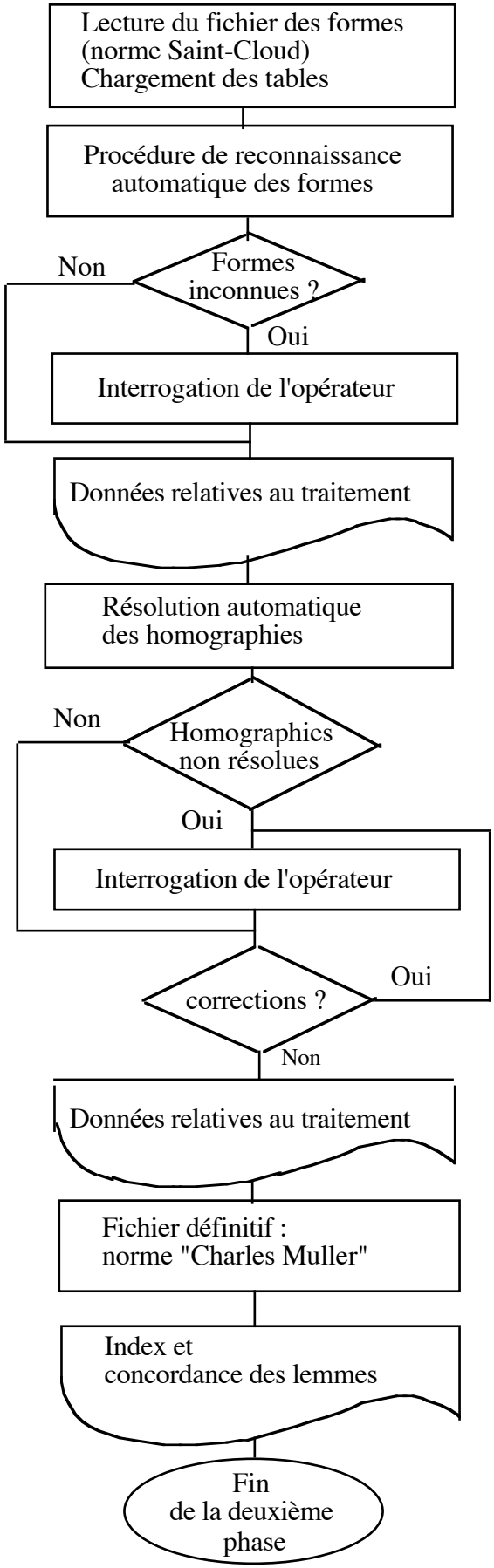
Lorsque le programme ne peut résoudre une difficulté, il interroge l'opérateur en lui demandant de choisir entre les solutions possibles. Ce choix est conservé en mémoire. A la fin du traitement, les décisions prises sont éditées sur papier (ceci permet de contrôler la qualité du programme de lemmatisation automatique, la stabilité de la norme de dépouillement et la cohérence des décisions prises par l'opérateur au long des opérations qui peuvent s'étaler sur plusieurs mois).

Phase 3. Corrections et constitution du fichier définitif

Après étude de l'index et du fichier provisoire, les corrections nécessaires sont apportées au texte lemmatisé. Ces corrections éventuelles sont également imprimées (leur étude permet d'améliorer le programme et de contrôler la qualité des textes déjà traités). Puis le fichier est débarassé des indications superflues (par exemple, pour les verbes, les indications de mode, de personne et de temps).

Une série de fichiers accessoires sont constitués (notamment un index alphabétique des lemmes et des formes).

Tableau 3. Les principales phases de la lemmatisation



4.23. La configuration des fichiers définitifs

A l'issue du dépouillement d'un corpus comportant n textes, les résultats se présentent sous la forme de (n+1) fichiers :

- un index alphabétique des vocables et des formes sous lesquels ils sont attestés. Pour chacune, l'index comporte la fréquence globale, les références des textes où des occurrences ont été constatées et les sous-fréquences correspondantes ;
- autant de fichiers qu'il y a de textes dans le corpus. En tête de chacun de ces fichiers, figurent le nombre de mots et le nombre de vocables différents. A chacun des mots du texte est attaché un lemme (comportant une forme canonique et l'indication du code grammatical).

Reprenons la phrase de F. Mitterrand sur le Tchad donnée en exemple ci-dessus. Après traitement, cette phrase se présente ainsi :

"les" (*le, déterminant*) "limites" (*limite, nom féminin*) "coloniales" (*colonial, adjectif*)"," "ayant" (*avoir, verbe*) "été" (*être, verbe*) "considérées" (*considérer, verbe*) "comme" (*comme, conjonction*) "un" (*un, déterminant*) "fait" (*fait, nom masculin*) "acquis" (*acquis, adjectif*), "devaient" (*devoir, verbe*) "devenir" (*devenir, verbe*) "une" (*un, déterminant*) "base" (*base, nom féminin*) "de" (*de, préposition*) "droit" (*droit, nom masculin*) "dans" (*dans, préposition*) "les" (*le, déterminant*) "relations" (*relation, nom féminin*) "entre" (*entre, préposition*) "les" (*le, déterminant*) "Etats" (*Etat, nom masculin*) ":" "tel" (*tel, pronom*) "était" (*être, verbe*) "le" (*le, déterminant*) "cas" (*cas, nom masculin*) "de" (*de, préposition*) "la" (*le, déterminant*) "bande" (*bande, nom masculin*) "d'" (*de, préposition*) "Aozou" (*nom propre*) "qui" (*qui, pronom*) "fait" (*faire, verbe*) "partie" (*partie, nom féminin*) "du" (*de le, préposition et déterminant*) "Tchad" (*nom propre*).

Cette présentation appelle quelques remarques :

- apparemment le texte triple de volume. En réalité, l'utilisation de codes et de différents symboles permet de limiter l'augmentation de taille des fichiers ;
- le texte original placé ici entre guillemets n'est pas modifié. Naturellement c'est lui qui est restitué à l'opérateur quand celui effectue des recherches ou lors des sorties "concordances" ;
- on remarque que "de" est toujours considéré comme une préposition ; "du" étant analysé en deux lemmes Nous nous en expliquerons dans la suite de cette note (§ 7.325).

*
* *
*

La lemmatisation des corpus est une nécessité sur laquelle les lexicographes sont maintenant pratiquement tous d'accord. Même les défenseurs de la "lexicométrie des formes" sont amenés à réintroduire les catégories grammaticales donc une norme de dépouillement sur laquelle ils ont été pour l'instant assez discrets. Ils préfèrent parler de "désambiguïsation" plutôt que de "lemmatisation" mais cela ne change rien quant au fond.

La suite de cette note présente la norme utilisée pour la lemmatisation du vocabulaire de F. Mitterrand, notamment les principales solutions apportées aux problèmes d'homographies. Suivant l'habitude des grammaires traditionnelles, cette partie sera partagée en trois chapitres :

- le groupe du verbe ;
- le groupe du nom (substantif, adjectif, déterminant, pronom) ;
- le groupe des "mots invariables" (préposition, conjonction, adverbe).

CHAPITRE 5
LA LEMMATISATION DES VERBES

Dans les dictionnaires, on a coutume de regrouper toutes les flexions d'un même verbe sous son infinitif. Nous avons voulu respecter cette tradition dans le premier but de faciliter la lecture de notre index. Pour ce faire, il fallait donc retrouver toutes ces flexions. Or la reconnaissance des verbes est la partie la plus difficile de la lemmatisation. Cette difficulté provient de trois problèmes qui se cumulent. Premièrement, c'est dans la catégorie des verbes que se rencontrent le plus grand nombre d'homographies potentielles. Deuxièmement, le nombre de flexions par lemme est beaucoup plus important que dans les autres catégories ce qui rend plus difficile la reconnaissance de chacune des formes et l'automatisation de cette tâche. En effet, en moyenne, chaque verbe peut présenter environ 100 conjugaisons et plus de 40 se traduisent par des graphies différentes. Si l'on devait stocker en mémoire ces formes pour les dix ou douze mille verbes les plus usuels, le coût serait véritablement prohibitif. D'où la nécessité, pour cette reconnaissance, d'utiliser des procédures plus sophistiquées que la simple lecture de tables. Enfin, troisièmement, les frontières de cette catégorie sont difficiles à tracer notamment entre le participe passé d'une part et l'adjectif et le verbe d'autre part.

5.1. LES PRINCIPES DE RECONNAISSANCE DU VERBE.

Tout verbe peut être analysé comme un *radical* - qui ne change pas quelles que soient les flexions du verbe - et une terminaison ou *désinence* qui, elle, varie. Au radical nous associons l'infinitif pour obtenir ce qu'on nommera une *racine* que nous analysons dans le deuxième point de cette section. A la désinence, nous associons la personne, le temps, le mode et la forme (premier point).

Remarque : pour des raisons qui seront explicitées ultérieurement, avoir, être, aller, falloir, pouvoir, vouloir, devoir sont stockés dans la table générale.

5.11. Les désinences verbales

Il s'agit donc de la terminaison du verbe, de sa partie flexionnelle. Le système de codification et de reconnaissance est inspiré du modèle utilisé par Maegard et Spang Hansen dans leur programme de segmentation automatique du français écrit[1]. L'ensemble des tableaux et des codes est présenté dans l'annexe 2 à la fin de cette note.

5.111. Les principes de classification des désinences

Nous analysons la désinence du verbe en fonction de quatre dimensions :

- les temps et modes : le futur, le conditionnel, le présent de l'indicatif, l'imparfait de l'indicatif, l'imparfait du subjonctif, le passé simple, l'impératif, le participe passé, le participe présent, l'infinitif ;
- la personne pour les formes conjuguées, le genre et le nombre pour le participe passé ;
- la classe de la conjugaison au sein du groupe en fonction des particularités de la langue française…

---

1. Bente Maegaard et Ebbe Spang-Hanssen, *La segmentation automatique du français écrit*, Paris, Editions Jean-Favard, 1978, p 92-97.

5.112. Les tables des désinences.

Ces tables sont présentées en détail dans l'annexe 2 à la fin de l'ouvrage.

5.12. Les racines verbales

Les racines verbales comportent deux parties : le radical et l'infinitif correspondant dont nous présentons successivement la codification ci-dessous.

5.121. La codification des radicaux

Dans la classification des verbes, il a été décidé de retenir la partition classique en trois groupes afin de rendre plus facilement intelligible l'analyse[1]. Au sein de ces groupes, la distinction entre sous-groupes suit au plus près la classification des frères Bescherelle mise à jour par les éditions Hatier[2].
Le tableau ci-dessous donne trois exemple de codification des radicaux. Les verbes du premier groupe (*chanter*) ont un radical unique sauf les verbes du type "placer" (pour tenir compte du *ça*), "manger" (pour tenir compte de *gea*), "lever" ou "céder" (lève, cède) et "jeter" (jette). Les verbes du deuxième groupe ont tous trois radicaux différents formés comme "finir" (sauf "haïr"). Le troisième groupe est plus divers : nous donnons ci-dessous l'exemple de "mettre". On signale en dernière colonne des homographies qui seront détaillées plus bas.

Dans le tableau ci-dessous, les premiers numéros correspondent aux groupes de désinences présentés dans le paragraphe précédent et dans l'annexe 2.

Infinitif (lemme)	radical	Groupes de classes de désinences : 1	2	3	4	5	6	7
chanter	chant-	1	1	1	1	1	1	0
accompagner	accompagn-	1	1	1	1	1	1	0
(...)								
finir	fini-	3	0	2	0	0	0	0
finir	finiss-	0	1	0	0	0	0	0
finir	fin-	0	0	0	2	2	2	0
(...)								
mettre	mett-	3	1	0	0	0	3	0
mettre	me-	0	0	26	0	0	0	1*
mettre	mi-	0	0	0	6	0	0	1*
mettre	m-	0	0	0	0	2	0	1*
(…)								

* Les homographies en cause sont : met(s), mis, mise(s).

5.122. La codification des infinitifs.

Dans les tables, nous associons à l'infinitif une série de codes qui nous servirons notamment à résoudre les homographies et à procéder à une lemmatisation automatique du texte.Cette classification des verbes repose sur les réponses apportées, a priori et pour chaque verbe, à trois séries de questions :

---

1. Cette classification présente de nombreux défauts mais elle reste la seule véritablement usitée. Pour une classification rationnelle, par exemple André Martinet, *Grammaire fonctionnelle du français*, Paris, Didier, 1984, p 91-97 qui nous a beaucoup inspiré.

2. *Le Bescherelle, l'art de conjuguer, dictionnaire de 12000 verbes*, Paris, Hatier, 1980.

5.1221. Le verbe admet-il un complément d'objet direct ?

En cas de réponse positive, la forme passive est possible et on peut rencontrer la construction "être+participe passé" même quand le verbe se conjugue essentiellement avec avoir. Et dans ce cas le participe passé doit être rattaché au verbe.

La catégorie des verbes intransitifs n'est pas homogène.

- Certains verbes sont absolument intransitifs. Il s'agit de "verbes d'état" qui ne peuvent se construire qu'avec un attribut, adjectif, substantif ou adverbe : "je suis vieux", "je suis le maître" ; "je suis très bien" ; "je suis chez moi". On trouve également quelques verbes d'action : "agir", "émigrer", "marcher", "voyager".

- D'autres verbes ont un emploi normal intransitif - et sont ainsi classés par les grammairiens - mais admettent des constructions transitives plus ou moins métaphoriques mais très courantes en français parlé ("vivre sa vie", "il pleut des cordes"…) La plupart des verbes considérés comme intransitifs peuvent donc s'employer avec un complément d'objet ("il pleure", "il pleure quelqu'un" ; 'il court", "il court cette course"…) et admettent une construction passive (c'est-à-dire l'auxiliaire "être" ("il a été pleuré par ses nombreux amis", "la course a été courue par des champions").

- Un certain nombre de verbes se conjuguant avec "être" ou "avoir" selon la nuance. Voici la liste des principaux : accroître, apparaître, atterrir, augmenter, border, camper, changer, chavirer, (dé)congeler, convenir, crever, croître, crouler, croupir, déborder, décamper, déchoir, décroître, dégeler, dégénérer, dégrossir, déménager, demeurer, démonter, dénicher, dépasser, descendre, détourner, diminuer, disconvenir, disparaître, divorcer, échapper, échouer, éclore, embellir, empirer, enlaidir, expirer, faillir, geler, grandir, grossir, nicher, maigrir, monter, paraître, passer, pourrir, rajeunir, réchapper, remonter, repasser, reparaître, ressusciter, résulter, retourner, repasser, sonner, stationner, tourner, trébucher, trépasser, vieillir…

- Enfin, il faut rappeler que les verbes transitifs peuvent être employés sans compléments : "Il boit"…

La distinction garde cependant un intérêt pour l'analyse : on établit une liste limitative de verbes réellement intransitifs se conjuguant avec l'auxiliaire avoir seulement. Ces verbes n'admettant ni forme pronominale ni construction passive, leurs participes passés précédés du verbe "être" seront considérés comme des adjectifs (cf ci-dessous § 5.233).

5.1222. Le verbe admet-il une construction pronominale même de manière exceptionnelle ?

Dans ce cas l'auxiliaire 'être' sera précédé d'un pronom réfléchi (me, te, se, nous, vous, se) :

Cette distinction est importante car elle emporte l'utilisation des auxiliaires être ou avoir et entraîne la résolution de certaines homographies (des participes passés, de s', et entre les articles et les pronoms (quand le pronom peut être inséré entre l'auxiliaire et le participe).

En fait, il faut introduire une quadruple distinction :

- les réfléchis essentiellement pronominaux sont toujours précédés d'un pronom. La construction passive est impossible ;

- les verbes intransitifs absolus n'admettent aucune construction pronominale ;

- les verbes transitifs complets admettent toutes les constructions : "ils se sont aperçus", "ils s'aperçurent", "ils aperçurent le port", "ils ont été aperçus…"

- les verbes facultativement pronominaux se construisent à la voix passive : "ils se vendirent" ; ce qui permet souvent d'obtenir une forme impersonnelle ("il s'en vend beaucoup") ou l'on peut analyser à la fois "s'" et "en".

5.1223. Avec quelle préposition ou conjonction le verbe est-il susceptible de se construire ?

Un certain nombre de mots sont susceptibles d'être attachés directement au verbe : à(au, aux), avec, contre, dans, de(du, des), en, pour, que, quoi, sur…

Nous avons utilisé à ce sujet la classification de Caput[1]. Son intérêt essentiel réside dans la possibilité de résoudre certaines homographies ('de', préposition-déterminant, 'en' pronom ou préposition, 'que', pronom ou conjonction) :

- la conjonction "que" est l'une des premières sources d'homographie. On donnera un code particulier à chaque verbe qui admet une construction complexe avec 'que' (c'est-à-dire introduisant une proposition relative) ;
- les préposition "de" et "en" peuvent introduire un verbe à l'infinitif ou un complément du verbe… Etant donné que nous avons renoncé à analyser "de", cette dernière classification perd une partie de son intérêt.

5.123. La classification des verbes.

5.1231. Classification selon l'auxiliaire

De la réponse aux deux premières questions (5.1222.a et b), on tire sept classes de verbes suivant les auxiliaires possibles :

1. "Etre" ou "avoir" selon la nuance
2. "Avoir" seul (verbe intransitif sans emploi pronominal possible)
3. "Etre" seul dans un emploi non pronominal (aller…)
4. "Etre" verbe essentiellement pronominal (pas d'emploi direct possible)
5. "Etre" ou "avoir" emploi direct et pronominal personnel possible
6. "Avoir" seul : emploi direct et pronominal impersonnel possible
7. "Etre" seul : emploi pronominal personnel ou impersonnel avec participe invariable.

5.1231. Classification des verbes selon leur construction

De la réponse à la troisième question (6.1222.c), on tire huit classes de verbes suivant les types de constructions possibles :

0. construction simple seule (verbe intransitif…)
1. verbe transitif sans construction complexe
2. construction avec "que"+relative seule admise.
3. construction complexe possible avec "à" + infinitif sans possibilité de "que + relative"
4. construction complexe possible avec "de" + infinitif sans possibilité de "que + relative"
5. construction complexe admise avec une préposition autre que "de" ou "à" sans possibilité de "que + relative"
6. construction complexe admise avec "à" ou "de" + infinitif ou "que + relative"
7. construction complexe avec des préposition autres que "de" et "à" avec possibilité de "que + relative"
8. construction complexe avec des préposition autres que "de" et "à" sans possibilité de "que + relative"

5.13. Les procédures de reconnaissance des verbes

En face d'une forme quelconque - par exemple "tînmes" - le programme consulte une première table générale dans laquelle sont codifiées toutes les formes non-verbales ainsi que les conjugaisons de quelques verbes de très haute fréquence (être, avoir, falloir, pouvoir, vouloir).

---

1. J. et J.-P. Caput, *Dictionnaire des verbes français*, Paris, Larousse, 1972.

5.131. Le cas des verbes réservés

La présence des verbes être, avoir, falloir, pouvoir, vouloir dans la table générale s'explique par deux considérations :

- leur conjugaison est si particulière qu'elle ne s'intègre que difficilement dans le schéma général présenté ci-dessus ;
- leur fréquence est si grande que, sans grand alourdissement des tables chargées en mémoire, on obtient une accélération appréciable de l'exécution du programme.

Si la forme ne se trouve pas dans la table générale ou si une homographie impliquant un verbe est signalée, la procédure de reconnaissance du verbe est mis en oeuvre

5.132. Examen préalable de la terminaison

Le mot se termine-t-il par une lettre qui ne se rencontre pas à la fin d'un verbe ? (il s'agit de : b, f, g, h, j, k, l, m, n, o, p, q, v, w, y). Dans ce cas le programme, ayant épuisé les codifications possibles, interroge l'opérateur.

5.132. La procédure de reconnaissance des verbes

La suite de la procédure de reconnaissance des verbes est résumée dans le tableau ci-dessous :

Passages :	Radical / Désinence
1	tînme / s
2	tînm / es
3	tîn / mes
4	tî / nmes
5	t / înmes

1. Le mot est coupé en deux parties et, à chaque cycle, la partie droite est augmentée d'une lettre, la partie gauche est diminuée d'autant.

- la partie droite est confrontée à la table des désinences. Ainsi, dans "tînmes", les désinences possibles sont : 's', 'es', 'mes', 'înmes'. Au quatrième passage, la procédure ne dépasse pas ce premier stade puisque 'nmes' ne se trouve pas dans la table des désinences.
- Lorsque la partie droite de la forme se trouve bien dans la table des désinences, le programme lit la table des radicaux à la recherche de la partie gauche. Si ce radical figure dans la table, les informations contenues dans celle-ci sont ajoutées à la forme. Ainsi, au cinquième passage, on obtient : "tînmes, tenir". Le lemme comporte également des indications concernant les constructions possibles du verbe (ici les codes contenus dans la table signifient que le verbe "tenir" est susceptible de se rencontrer dans des constructions transitive et intransitive ou pronominale mais non avec la conjonction "que") et sa conjugaison (ici le programme indique qu'il s'agit avec certitude de la première personne du pluriel du passé simple).

2. La procédure épuise toutes les possibilités. Même en cas de découverte d'une solution possible :

- la lecture de la table des désinences est poursuivie jusqu'à épuisement : cela permet de détecter les homographies au sein d'un même verbe entre personnes, temps, modes… Par exemple, "chante" peut être la première et la troisième personne de l'indicatif présent, du subjonctif présent et de l'impératif ;

- la table des racines est également épuisée : cela permet de détecter les homographies entre verbes (cf ci-dessous le § 5.221) ;

- la procédure s'achève lorsque la désinence excède huit lettres (taille maximale des désinences présentées dans les tableaux de l'annexe 2) ou lorsque la racine ne contient plus de lettres (dans notre exemple, la désinence "tînmes" n'est pas examinée car cela conduirait à un radical vide) ;

- lorsque plusieurs solutions sont trouvées, elles sont toutes stockées après la forme et une série de tests leur est appliquée ultérieurement pour résoudre cette homographie.

Lorsqu'aucune solution n'a été rencontrée, l'opérateur est interrogé. Soit le texte comporte une erreur de frappe, soit le vocable n'existe pas dans les tables. La décision prise par l'opérateur est imprimée sur papier. Cela permet d'assurer la stabilité de la norme, de compléter les tables et les programmes. Dans la phase actuelle, où l'ensemble de la procédure reste expérimentale, cette précaution est indispensable.

5.2. LA RESOLUTION DES HOMOGRAPHIES DU VERBE (PRINCIPES GENERAUX).

Rappelons que la résolution des homographies est confiée à un programme. L'opérateur n'intervenant qu'en cas d'échec. La machine doit permettre un gain de temps et assurer la stabilité des conventions de lemmatisation. Pour le verbe, cette résolution repose sur des classifications que nous résumons ci-dessous.

5.21. La grille générale

5.211. La codification des homographies du verbe

Dans le tableau présenté au paragraphe suivant, le code est celui utilisé par le programme :

• la première lettre désigne les modes et les temps :

A. le futur
B. le conditionnel
C. le présent de l'indicatif
E. l'imparfait de l'indicatif
F. l'imparfait du subjonctif
G. le passé simple
H. l'impératif
I. le participe passé
J. le participe présent
K. l'infinitif

• le second chiffre,

- pour les lettres A à I, ce chiffre indique la personne. Le chiffre 7 désigne une flexion commune aux première et deuxième personnes du singulier (finis) ; le chiffre 8, les première et troisième personnes du singulier (chante).

- pour les participes passés (lettre I) :

1. masculin singulier
2. masculin pluriel
3. féminin singulier
4. féminin pluriel
5. masculin singulier et pluriel identiques
6. invariable

• la dernière lettre donne,

- pour les adjectifs-substantifs homographes :

a. masculin singulier
b. masculin pluriel

c. féminin singulier
d. féminin pluriel
e. masculin singulier et pluriel (bois)
f. féminin singulier et pluriel (souris)
g. masculin et féminin singulier (garde)
h. masculin et féminin pluriel (gardes)
i. masculin singulier et pluriel, féminin pluriel (cours)

- pour les homographies avec d'autres catégories :

k. préposition
l. adverbe
m. {indicatif-impératif-subjonctif} et participe passé
y. cas particuliers

5.212. Tableau de synthèse des homographies du verbe

Les principaux cas possibles sont résumés dans le tableau ci-dessous. Un exemple est donné en dernière colonne.

N°	Code 1	Code 2	Code 3	Exemples
1	C3a	C2b		avantage(s)
2	C3c	C2e		charge(s)
3	C3g	C2h		garde(s)
4	C3Z	I1Y		dit
5	I3c	I4d		atteinte (s)
6	A3c	A2d		aura(s)
7	E4b			avions
8	C2e			as
9	G3a	G2b		but(s)
10	C3c	C2d	C4b	bouch(e,es,ons)
11	C4b			fripons
12	E4d			intentions
13	C3a			bruit
14	C2i	C3a		cour(s,t)
15	C3a	I1e		écrit(s)
16	G1b	G2b		défi(s)
17	C5m	I6f		défait(s,e,es)
18	C3k			entre
19	C2b			entretiens
20	F3a			fût
21	C6a			excellent
22	D3c	D2d		faille(s)
23	C2f	G2f	I2f	souris
24	G3a	I5e		interdit(s)
25	E7e			niais
26	C3a	D3c	D2d	sort(e,es)
27	G2b			vins
28	C4b	E4d		vis(ons,ions)
29	D3c	D2d	C2f	vi(ve,ves,s)
30	Ja			étudiant
31	Ka			pouvoir

Pour un même verbe, on ne s'occupe pas des homographies concernant la personne ni de celles concernant le groupe {impératif-indicatif-subjonctif} sauf lorsqu'elles mettent également en jeu un substantif, un adjectif…

5.22. Les homographies absolues

Nous regroupons sous ce terme une situation particulière : la forme est graphiquement la

même alors que la matière sémantique conduit à deux entrées différentes au sein d'une même catégorie. Comme nous l'avons indiqué en introduction de cette partie, en principe, cette situation ne doit pas donner lieu à analyse. Cependant, pour les verbes, ce principe admet deux types d'exceptions.

5.221. Les homographies entre deux verbes différents

Le tableau ci-dessous récapitule les homographies au sein de la catégorie des verbes. Une quatrième colonne rappelle que certaines se doublent d'une confusion possible avec une forme d'une autre catégorie (habituellement des substantifs). Certaines de ces formes ont une fréquence élevée. Ainsi en est-il de "suis" qui est certainement l'une des homographies les plus difficiles (puisque, dans les deux cas, la forme est à la première personne du singulier de l'indicatif présent). A lui seul ce problème nécessite un traitement spécial. Cependant on observera qu'il ne s'agit pas d'un cas isolé et que, souvent, l'homographie porte sur les mêmes modes, voire les mêmes temps…

Généralement, ces cas devront être résolus par l'opérateur…

Formes	verbe 1	verbe 2	Autres
admirent	admirer	admettre	
alliez allions	aller	allier	
choient	choir	choyer	
comparais	comparaître	comparer	
convient	convenir	convier	
crevasse	crever	crevasser	substantif
crevassent	crever	crevasser	
crus crue(s) crûmes, crûtes	croître " "	croire " "	substantif "
durent	devoir	durer	
dépeigne(s,ons,ez,ent) dépeignai(s,t,ent) dépeignant	dépeigner " "	dépeindre " "	
faut	falloir	faillir	
fondai(s,t,ent) fonde(s,ons,z,nt) fondi(s,ons,ez) fondant	fondre " " "	fonder " " "	 substantif
lacèrent	lacer	lacérer	
méprise(s)	méprendre	mépriser	substantif
mirent	mirer	mettre	
mise(s)	miser	mettre	substantif
moul(ons,ez,ent) moulai(s,t,ent) mouli(ons,ez) moulant	mouler " " "	moudre " " "	
murent	murer	mouvoir	
ouvre(s,ent)	ouvrir	ouvrer	
parais	paraître	parer	
pari(ons,ez)	parier	parer	

peign(e,es,ons,ez,ent)	peindre	peigner	substantif
peign(ais,t,ent)	"	"	
peignant	"	"	
plaisante(s)	plaire	plaisanter	adjectif
plu	plaire	pleuvoir	
plut	"	"	
pressent	presser	pressentir	
prise(s)	priser	prendre	substantif
recouvr(e,es,ons,ez,ent)	recouvrer	recouvrir	
revis	revoir	revivre	
reprise	repriser	reprendre	substantif
sue	suer	savoir	
suis	être	suivre	
tue(s)	tuer	taire	
vis(se,ses,ions,sent)	voir	visser	substantifs (2)
virent	voir	virer	
vi(s,t)	vivre	voir	substantif

Nous rejetons la codification "à l'aveuglette" reposant sur des probabilités observées sur des corpus plus vastes ou sur des échantillons. Acceptable quand la lemmatisation se fait sur l'index-concordance, cette solution n'est pas tenable quand la lemmatisation est opérée dans le texte car elle introduit nécessairement des erreurs. Soit ces difficultés sont codifiées à l'avance avant la mise en route du programme, soit, lorsqu'une de ces formes est rencontrée lors de l'exécution, l'opérateur est interrogé. Etant donné la fréquence de certaines d'entre elles, on a choisi d'identifier celles-ci avant le traitement (ainsi pour "suis") : règle 5.235 ci-dessous.

5.222. Les homographies dans les flexions d'un même verbe

Pour certains verbes, le participe passé s'écrit comme certaines formes conjuguées. Il en est systématiquement ainsi avec les verbes du deuxième groupe pour le présent, le passé simple, l'impératif. Comme nous l'indiquons plus bas (§ 5.231), les homographies entre les différents temps et modes ne sont pas dénouées. En revanche, celles qui concernent les participes doivent l'être car ils sont (comme les infinitifs) "des classes à part distinctes de celles du verbe"[1] proprement dit. Un examen de ce problème permettra de comprendre comment procède la résolution automatique des homographies en s'appuyant sur l'examen du contexte étroit du mot à codifier.

On sait que le participe passé peut se rencontrer dans trois grands types de constructions :

- à la voie active ou passive, précédé d'un auxiliaire, qui peut être, selon les cas, "avoir" ou "être" on y reconnaîtra un verbe, en tenant compte des constructions plus ou moins complexes qui peuvent éloigner le participe de son auxiliaire ;
- précédé d'un auxiliaire au participe présent ("La cigale ayant chanté…") ;
- précédé d'un nom, d'un déterminant ou d'un autre adjectif et accordé à eux en nombre et en genre, il appartient au groupe du nom sauf s'il s'agit d'un homographe absolu car le verbe peut ici être conjugué.

Dans cette dernière hypothèse, on tiendra compte des indications suivantes :

- pour les première et seconde personnes du singulier et du pluriel, le verbe fléchi est généralement précédé d'un pronom. Le test à gauche résout la plupart des cas : l'absence du pronom permet généralement de trancher en faveur du nom.

1. André Martinet, *op. cit.*, p 113.

- pour les troisièmes personnes, la présence à gauche d'un pronom personnel accordé (il, elle, on, ils, elles) ou d'un relatif (que, qui…) permet généralement de trancher en faveur du verbe à un mode autre que le participe ;
- pour les troisièmes personnes, la présence d'un nom, d'un déterminant ou d'un adjectif à gauche, lorsque l'accord en genre est réalisé, permet rarement de résoudre l'homographie absolue en l'absence d'autre indices comme la construction négative (ne… pas, plus, etc)… Après une recherche à gauche et à droite à la recherche de ces indices, l'appel à l'opérateur devient alors indispensable.

5.23. Les principes généraux de résolution des homographies du verbe

Les quatre principes de résolution des homographies du verbe sont discutés dans les paragraphes suivants de cette section :

5.231. La règle "finis"

Les flexions homographes d'un même verbe ne sont pas distinguées sauf quand l'homographie met en jeu le participe passé et d'autres modes. Ainsi 'finis' peut-être un participe passé (masculin pluriel) ou un indicatif (présent ou passé). Nous discutons ce problème dans le § 5.34 ci-dessous.

5.232. La règle "étudiant, étudiante"

Seules les formes se terminant en "ant" peuvent être des participes présents. Les formes se terminant en "ants", "ante", "antes" sont rattachées à d'autres catégories (substantif, adjectif…). On établit une liste limitative de participes présents susceptibles d'être utilisés comme substantifs ou adjectifs (masculin singulier) : annexe 4 ;

5.233. La règle "immigré"

A priori tout participe passé peut être utilisé comme adjectif ou substantif : seule la place dans la phrase détermine si l'on a affaire à un verbe, à un adjectif ou à un substantif. Par exemple :
- "il a immigré" est rattaché au verbe "immigrer" ;
- "il est immigré" est considéré comme un adjectif puisqu' "immigrer" est un verbe intransitif ;
- "un immigré algérien" est considéré comme un substantif.

Cependant, on peut établir une liste à peu près exhaustive des participes passés substantivés (annexe 3). En dehors de cette liste, on considérera que l'homographie se réduit au couple verbe-adjectif. Le participe passé précédé du verbe "être" sera toujours rattaché au verbe sauf pour une liste limitative de verbes absolument intransitifs.

5.234. La règle "faire affaire"

Dans de nombreuses expressions, des verbes usuels peuvent être combinés avec des substantifs homographes sans en être séparés par un déterminant : "faire affaire", "rendre compte", "faire partie", "donner prise", etc. Une liste complète de ces exceptions est donnée dans le tableau placé en annexe 6. En dehors de ces cas, le substantif est séparé du verbe par un déterminant ou une préposition

5.235. La règle "suis"

L'opérateur doit effectuer un traitement préalable de certaines formes impossibles à analyser dans l'état actuel de la réflexion ou dont la fréquence très élevée risque de ralentir les opérations.

Ainsi "suis" (verbe être et verbe suivre), "est" (point cardinal et verbe être…[1]) ne peuvent être reconnus comme verbes qu'en position d'auxiliaires lorsqu'ils sont suivis d'un participe passé (Pour une liste complète, cf ci-dessus le § 5.22).

De même, toutes les formes qui demeurent ambiguës à la fin de la procédure automatique sont soumises à l'opérateur. A priori, aucune codification par défaut n'est admise car elle pourrait conduire à des erreurs.

5.3. L'HOMOGRAPHIE DU VERBE AVEC D'AUTRES CATEGORIES (ETUDES DE CAS)

5.31. Les homographies des formes conjuguées

Nous traitons plus loin du cas des formes conjuguées homographes du participe passé. La présente sous-section concerne uniquement les flexions des verbes formellement identifiées comme appartenant à l'indicatif, au subjonctif, à l'impératif ou au conditionnel.

5.311. La première et la troisième personne

Ce sont les homographies les plus nombreuses. Par exemple, des centaines de substantifs et d'adjectifs se terminant par "e" sont homographes avec des verbes à la première ou la troisième personnes du singulier ("je formule", "une formule" ; "tu formules", "des formules"…) Nous donnons une liste non limitative de ces homographies dans l'annexe 5 à la fin de cette note. A titre d'illustration des programmes de la lemmatisation assistée par ordinateur, nous détaillons ci-dessous les règles suivies pour permettre une résolution automatique de ces homographies.

Dans le cas présent, une difficulté supplémentaire provient de l'homographie entre pronom et déterminant qui se combine très souvent avec celle du verbe. Exemples :

(a) "je formule une demande";
(b) "je la formule" ;
(c) "je les formule" ;
(d) "je ne la formule pas"
(e) "je la leur formule" ;
(f) "je ne la leur formule pas" ;
(g) "je donne la formule"
(h) "la vieille formule"
(i) "le porte-parole la formule"
(j) "les formule-t-il ?"

Si on se contentait de la présence d'un déterminant (ou supposé tel) à gauche de l'homographe du verbe, dans (b), (d), (e), (f) et (i) on codifierait "formule" comme un substantif. Les combinaisons possibles sont si nombreuses qu'il est illusoire de penser éviter absolument l'appel à l'opérateur mais quelques tests peuvent résoudre la grande majorité des cas. Dans l'ordre, le programme applique les règles suivantes:

• La règle de l'accord à gauche

1. le substantif supposé doit être accordé en genre et en nombre avec le déterminant supposé situé à sa gauche :

- en cas de réponse négative, le cas est tranché en faveur du verbe. L'accord de celui-ci avec un sujet est requis à titre de contrôle (cf 2 ci-dessous). Ainsi dans (c) la présence de "je" en (n-2) permet de décider avec certitude. Le sujet peut également se trouver à droite en cas de construction interrogative (j). A défaut d'un pronom, un groupe nominal comportant un substantif au singulier doit se trouver à gauche en position sujet ;

---

1. Sur ce point, la difficulté est résolue en suivant la convention typographique dominante qui place les points cardinaux en majuscules.

- en cas de réponse positive, une présomption existe qui doit être confortée par les tests suivants ;

2. le verbe supposé doit être accordé en nombre avec le sujet supposé (pronom ou groupe nominal) :

- à gauche - en (n-1) ou sous des conditions restrictives en (n-2) -, la présence d'un pronom (sans homographe) de la première ou de la troisième personne du singulier permet de trancher sans équivoque en faveur du verbe ("je formule") ;

- la présence d'un pronom en (n-2) ne permet pas de trancher en faveur du verbe sans avoir auparavant vérifier l'absence d'inversion du sujet (type construction interrogative). Par exemple : "Voit-il l'annonce ?"

- sous la réserve précédente, la règle est étendue aux pronoms réfléchis. Par exemple : "il me l'annonce" est interprété sans équivoque comme un groupe verbal ;

- à droite, la présence d'un pronom (sans homographe) de la première ou de la troisième personne du singulier permet de trancher en faveur du verbe s'il n'est pas suivi d'un groupe verbal, lui-même accordé ;

- à gauche, la présence d'un substantif singulier permet de passer à la règle n° 4 ;

• la règle de l'adverbe et de la construction négative

3. La présence d'une négation ("ne", "ni", "pas"…) permet de trancher avec certitude en faveur du verbe si la règle de l'accord est satisfaite - cas (d) et (f) - ou de remonter en (n-3) - ou (n-4) sous certaines conditions - à la recherche du sujet : "il ne me l'annonce pas" (du même coup l'homographie "pas" se trouve résolue en faveur de l'adverbe) ;

4. la présence d'un adverbe permet de décaler le programme d'un pas à gauche, ou à droite, et de recommencer les tests après avoir contrôlé que les règles de combinaison des adverbes ne conduisent pas à arbitrer en faveur du substantif ou du groupe verbal (cf § 7.122) ;

• la règle du substantif unique dans le groupe nominal

5. on ne peut rencontrer deux substantifs dans un même groupe nominal. Ils doivent être séparés par un signe quelconque : conjonction, préposition, virgule ou autre ponctuation mineure :

- le premier substantif doit être identifié avec certitude et satisfaire à la condition 2 ci-dessus. Par exemple, dans (i), "porte-parole" satisfait à ces trois critères. On peut donc en déduire avec certitude que "l'annonce" est composé d'un pronom et d'un verbe ;

6. une virgule ou une conjonction ("et", "ou", "soit"…) en (n-1) ou (n-2) oblige à aller plus à gauche : si l'on rencontre, en (n-2) - ou (n-3), (n-4) sous conditions - :

- un verbe certain de même conjugaison (personne, temps, mode), on en tire que la forme est un verbe ;

- un substantif ou un adjectif de mêmes genre et nombre, on est en face d'un adjectif ou d'un substantif ;

Remarque : cette règle pose le problème des locutions de type adverbial composées à partir d'un substantif. A plusieurs reprises, chez F. Mitterrand, nous avons rencontré plusieurs groupes nominaux juxtaposés sans conjonction. Seule une virgule peut éviter de placer le programme de reconnaissance dans une situation indécidable. L'un des premiers cas était le suivant : "J'ai vu se défaire en *même temps un peuple* et un pays" où "un peuple" était inanalysable par le programme. D'une part "un" n'étant pas précédé du déterminant "le", il ne pouvait s'agir d'un pronom et, d'autre part, l'absence de virgules encadrant "en même temps" faisait coexister deux substantifs (temps et peuple) dans un même groupe nominal… Cet exemple montre combien la ponctuation d'une transcription doit être effectuée soigneusement. Evidemment à l'écrit de telles situations se traduisent par un blocage complet et un appel à l'opérateur ;

• la règle de composition des déterminants

7. Le test le plus puissant à gauche utilise la loi selon laquelle deux déterminants de catégorie 1 ne peuvent pas coexister dans le même groupe nominal (§ 6.423). Cette loi permet de reconnaître dans (e) et (f) la présence de deux pronoms et donc d'un verbe ;

• la règle de l'adjectif antéposé

8. A gauche la présence d'un adjectif susceptible d'antéposition (§ 6.23) peut amener à reconnaître un substantif. Attention : il ne doit pas y avoir de doute sur le déterminant ou d'ambiguïté sur la nature de l'adjectif antéposé : si l'on admet que "vieille" peut aussi être employé comme substantif (tournure familière pour "une vieille personne"), le test n'est plus concluant et (h) doit être considérée comme une forme ambiguë nécessitant l'appel à l'opérateur ;

• La règle du verbe fléchi unique dans le groupe verbal

9. Deux verbes conjugués ne peuvent se suivre dans le groupe verbal. Le second est nécessairement au participe passé derrière "être" et "avoir", à l'infinitif ou au participe présent dans les autres cas. Cette règle permet de résoudre les nombreuses tournures où le substantif est placé immédiatement derrière un verbe (règle "faire affaire") ;

• Si les neuf premiers tests ne donnent pas de résultats, la présence d'une forme homographe, {déterminant/pronom} devant la forme, oblige à remonter plus loin à gauche. Suivant qu'on rencontre en (n-2) - ou (n-3) sous certaines conditions :

- une ponctuation majeure, on arbitre en faveur du couple déterminant et substantif ;
- la présence d'un groupe nominal complet (déterminant et substantif) oblige à aller à droite rechercher soit un groupe nominal (complément d'objet) soit une préposition compatible avec le verbe pour trancher en faveur de ce dernier…

Une fois parvenu à ce point, on débouche sur un appel à l'opérateur. Mais l'application des instructions précédentes aura permis de résoudre plus de neuf sur dix homographies du verbe à la première et à la troisième personne. La résolution du dernier dixième paraît à la fois très coûteuse et comporter trop de risques d'échecs ou d'erreur…

5.312. La première personne du pluriel (avions…)

Un nombre assez important de formes sont concernées par cette homographie. La prononciation les différencie habituellement mais non la graphie. Voici les cas recensés pour la seule lettre "a" : acceptions, adoptions, affections, agressions, aiguillons, attentions, avions…

La présence du pronom "nous" devant la forme ou derrière celle-ci (forme interrogative) permet de détecter pratiquement avec certitude le verbe. Il est parfois nécessaire de remonter plus en amont sur la gauche pour éviter des erreurs : "nous les acceptions" voire "nous ne les acceptions guère…"

"Avions" peut également servir d'auxiliaire : la présence d'un participe passé à droite permet de trancher en faveur de "avoir".

La présence à gauche d'un déterminant ou d'un adjectif antéposé accordés en genre et en nombre permet de trancher avec certitude en faveur du substantif.

5.313. La troisième personne du pluriel (parent…)

Un peu moins fréquente mais concerne des formes de fréquences souvent élevées : couvent, dément, évident, équivalent, parent, président… Cette dernière est employée 300 fois par F. Mitterrand lors de son septennat ! L'analyse repose essentiellement sur :

- présence d'un pronom de la troisième personne du pluriel ;
- un groupe nominal pluriel pluriel devant la forme ;

- un déterminant, un adjectif avant ou après au masculin singulier.

5.32. Les homographies de l'infinitif

D'accord avec C. Muller[1], nous séparons le substantif de l'infinitif dont il est dérivé. Les critères adoptés sont les suivants (dans l'ordre de l'analyse) :

- la présence, devant la forme analysée, d'un déterminant (ou d'un adjectif susceptible d'être placé devant le substantif) : "la venue *au* pouvoir de la gauche" ; "un *impérieux* devoir"…
- la présence devant la forme analysée d'un verbe à la forme active ou d'un autre infinitif permet d'arbitrer en faveur du verbe : "il voudrait pouvoir faire…" ; "il pense *devoir dire*" ;
- l'absence de déterminant amène généralement à trancher en faveur du verbe même dans quelques cas litigieux : nous voyons dans "*vouloir* c'est *pouvoir*" deux verbes…
- la présence d'un "le" devant la forme peut dans certains cas provoquer des situations inanalysables sans le secours de l'opérateur. Par exemple : "il voudrait le pouvoir" peut signifier que la personne souhaite pouvoir faire quelque chose (pronom+ verbe) ou qu'elle veut accéder au pouvoir (déterminant et nom)… Seul le contexte large - voire la situation d'énonciation - permet de résoudre ce problème.

5.33. Les homographies du participe présent

5.331. Discussion

Il semble que, en général, la dérivation du participe au substantif s'accompagne soit d'un changement orthographique dans la racine (fabriquant/fabricant, intriguant/ intrigant) soit de la formation d'un couple "a/e" (par le latin) : précédant/précédent, adhérant/adhérent… Cependant à l'époque moderne cette fragile barrière est rompue et l'on observe une multiplication des homographies. L'annexe 4 récapitule les formes ambiguës que nous avons rencontrées au cours de nos lectures.

La solution restrictive est dominante chez les lexicographes. Voici quelques exemples relevés dans l'index inverse des formes établis par G. Engwall : "immigrants" est placé sous le verbe "immigrer" ; en revanche "émigrants" est considéré comme un nom masculin ; "lancinantes" est un adjectif mais "passionnantes" un verbe ; "militantes" est un verbe mais "étudiant" et "étudiante" sont des noms et l'on ne trouve pas ces formes sous le verbe "étudier"… Comme pour le participe passé que nous examinons plus loin, il semble difficile de trouver une rationalité à ces classements…

C'est A. Lyne qui propose la solution la plus raisonnable et la plus élégante. A. Lyne utilise le critère de l'accord[2] :

- aucune forme en "ants", "ante(s)" n'est rattachée à un verbe. Par exemple : "étudiant" est ambigu mais "étudiante" ne peut être qu'adjectif ou substantif (féminin singulier) ;
- les formes en "ant" susceptibles de s'accorder ne sont pas rattachées à un verbe. Ainsi "un militant" donne "des militant*s*" par changement dans la détermination : il s'agit donc d'un substantif. "Un électeur hésitant" est un adjectif car on peut dire : "une électrice hésitante". En revanche : "un électeur hésitant entre la droite et la gauche" est bien un verbe car on aura : "des électeurs hésit*ant* entre…"

---

1. Charles Muller, "Le Mot, unité de texte et unité de lexique", *Langue française et linguistique quantitative, Paris-Genève*, Slatkine-Champion, 1979, p 146.

2. Anthony A. Lyne, "L'élaboration des listes de fréquence. A la recherche d'une solution aux problèmes d'affectation des mots-occurrences dans les classes de mots", *Cahiers de lexicologie*, 1973, II, p 83-108. Notamment, p 91-94.

- les substantifs en "ant" sont en nombre limité. Il sont enregistrés dans une table a priori qu'il est naturellement possible de compléter à la rencontre de créations nouvelles (cette table est reproduite en annexe 4).

- dans le cas de l'adjectif, il est impossible d'établir une limite stricte car le français contemporain permet de former un adjectif à partir de pratiquement n'importe quel participe présent. On ne peut donc en dresser une liste exhaustive.

Ces quelques exemples permettent de comprendre les critères utilisés par la machine pour départager les verbes des autres emplois.

5.332. Les formes en "ants", "ante(s)"

Ces formes accordées sont comptées en substantifs ou adjectifs. On se trouve alors dans le cas normal d'une homographie entre substantif et adjectif traitée dans le § 6.32 Les règles générales de lemmatisation sont les suivantes :

- les formes en "ants" sont rattachées au masculin singulier "ant". Si aucune homographie n'est mentionnée dans les tables, il est automatiquement considéré comme un adjectif ;

- les formes en "ante(s)" ont deux lemmes possibles. Elles sont rattachées soit à un substantif féminin, soit à un adjectif (dont la forme canonique sera au masculin). Si le cas est prévu dans les tables d'homographie, on se retrouve dans le cas standard de l'homographie du substantif et de l'adjectif. Si le cas n'est pas prévu, le programme tranche automatiquement en faveur de l'adjectif.

5.333. Les trois homographies des formes en "ant"

Trois cas sont à envisager :

- le cas habituel entre le participe présent et l'adjectif. Ici pas de limites strictes. On admettra que la position dans la phrase est le seul indice d'un emploi en tant que verbe ou en tant qu'adjectif.

- un cas complexe : l'homographie entre le participe présent et le groupe {substantif-adjectif}. La dérivation est arrivée à son terme et, à côté du verbe, nous avons un substantif qui a trouvé son autonomie. A chaque fois que le programme détectera un participe présent, il devra consulter la table des homographes pour déterminer s'il se trouve dans ce cas… La forme sera soumise à une batterie de tests et en cas d'échec, l'opérateur sera interrogé…

- quelques cas particuliers. On rencontre l'homographie entre le participe présent et le groupe {préposition-adverbe} parfois {adverbe-substantif-adjectif} : approchant, concernant, considérant, durant, maintenant, moyennant, partant, pendant, suivant… Les principaux tests utilisés pour dénouer ces ambiguïtés sont décrits dans les § 7.135 et 7.322.

5.334. Les tests de reconnaissance des participes présents en "ant"

De manière générale, les tests portent sur la position dans la phrase et sur l'entourage immédiat. Dans l'ordre :

• des tests à gauche :

1. "Etant" ou "ayant" sont toujours rattachés aux verbes "être" et "avoir". En position d'auxiliaire ils sont suivis d'un participe passé) ;

2. derrière "en" : participe présent (du même coup 'en' sera une préposition et non pas un pronom) ;

3. derrière un substantif ou un adjectif féminin ou pluriel : verbe (les électeurs *hésitant* entre…) ;

4. derrière une négation "ne *souhaitant* pas…" ;

5. derrière un pronom réfléchi "se *voulant* au-dessus des partis" ;

6. derrière un déterminant(s) et (éventuellement) adjectif antéposé au masculin singulier, le programme passe aux tests à droite :

• des tests à droite permettent de détecter le verbe dans les constructions suivantes :

7. devant un verbe à l'infinitif : "un grave événement *pouvant* survenir…"

8. devant la conjonction "que" si le verbe admet la construction en "que" : "*supposant* que" ; "*admettant* que"…

9. devant un attribut du sujet : "les circonstances paraissant favorables" ;

10. devant une préposition acceptable avec le verbe en question : "*marchant* vers"

• cas ambigus :

- le dernier critère n'est pas d'une grande solidité. Nous avons rencontré plusieurs cas d'adjectifs en "ant" complétés par des prépositions ("les conditions existantes à ce moment") sans qu'on puisse affirmer qu'il s'agit d'une faute ou d'un lapsus… Il n'est donc pas programmé et on le signale à l'opérateur comme étant une présomption en faveur du verbe ;

- "le cas échéant". Il s'agit d'une locution adverbiale qui ne répond pas à la règle "bric à brac" exposée dans la première partie de cette note car on rencontre "échéante(s)", "échéants". Puisqu'on rencontre "échu" en position d'adjectif ("le terme échu"), nous avons décidé de voir également dans "échéant" un adjectif mais, à la réflexion, le critère de l'accord ne semble pas rempli et nous ne maintiendrions peut-être pas cette position aujourd'hui.

- "ayant cause", "ayant droit". Ces formes sont considérées comme des mots composés. Le critère de l'accord les rattache au substantif ("ayants droit", "ayants cause") ;

5.34. Les homographies des participes passés

5.341. Discussion

Le problème se pose pour de nombreux verbes (voir la liste des homographies du participe passé présentée dans l'annexe 3). Suivant la tendance dominante, deux solutions sont envisageables. La première, très radicale, consiste à regrouper toutes ces formes sous les verbes dont on considère qu'elles sont dérivées[1]. Cette règle a l'avantage de la simplicité mais elle entraîne des incohérences. Par exemple, dans l'index de G. Engwall, on trouve 157 "mort, nom féminin" mais aucun "mort, substantif masculin" ; en revanche sous le verbe "mourir" on trouve 158 mort(s) et morte(s) dont plusieurs sont plutôt des substantifs masculins ou des adjectifs (masculins et féminins). D'autre part, la convention amène à "recréer" des verbes dont on peut douter qu'ils puissent jamais être conjugués. Ainsi le verbe "désentimentaliser" inventé par G. Engwall pour lemmatiser : "Il voit des être durs, *désentimentalisés* par leur propre discipline"[2]. Alors même que, par ailleurs, on trouve plusieurs cas où l'auteur a renoncé à cette opération trop artificielle. Par exemple : "accidenté" (adjectif) n'est pas rattaché à un verbe "accidenter" pourtant plus plausible que "désentimentaliser"… La seconde solution, proposée par C. Muller, consiste à utiliser des critères sémantiques : "Le participe exprime-t-il encore le résultat d'une action ? Ses limites sémantiques concordent-elles avec celles du verbe ?"[3]. Ces critères sont clairs mais, reposant sur le "sentiment linguistique" de l'opérateur, ils ne peuvent garantir la stabilité de la norme de dépouillement.

Il nous semble donc nécessaire d'analyser complètement ces homographies. Ainsi trouverons nous "un *immigré*" sous le nom, "un travailleur *immigré*" sous l'adjectif et "il a *immigré*" sous le verbe. Comme pour les autres catégories, l'analyse des homographies repose sur des critères syntaxiques et fonctionnels ; lorsque nous codons "désentimentalisé" comme adjectif, nous

---

1. C'est la solution préconisée par C. Muller pour les participes passés adjectivés : art. cit, p 137. Il apportera par la suite quelques amendements sur lesquels nous revenons ci-dessous : *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse, 1967, rééd : Paris-Genève, Slatkine-Champion, 1979, p 33-34.

2. Gunnel Engwall, *op. cit.*, p XXXI.

3. Charles Muller, *op. cit.*, p 33.

sousentendons qu'il s'agit d'un participe passé faisant fonction d'adjectif… Leur nombre est infini, comme le soulignait Littré, car tout verbe contient potentiellement un adjectif même si celui-ci ne s'actualise pas. Ceci étant, il n'en est probablement pas de même pour le substantif. Le mouvement qui va du participe passé au nom en passant par l'adjectif est forcément plus long et demande une certaine ratification sociale. On pourrait songer à en établir une liste limitative qui éviterait de prendre en compte des créations arbitraires et sans lendemain. Cependant, plus on s'éloignerait de la date de sa création, plus la liste deviendrait obsolète. Par exemple, il est probable qu'il y a trente ans, on n'aurait pas admis "immigré" dans la liste de ces substantifs : dans le vocabulaire du roman des années 1960 analysé par Engwall, "immigrer" n'apparaît qu'une fois sous la forme traditionnelle ("immigrant" à propos des Etats-Unis !).

Dès lors, la solution raisonnable est d'étendre le critère fonctionnel aux substantifs. Cependant, la complexité des constructions augmente le risque d'erreur. Nous avons donc adopté la solution suivante :

- si l'homographie est connue (type "mort" ou "immigré"), nous retombons dans le cas général de l'homographie verbe-substantif dont nous analysons ci-dessous les caractéristiques particulières induites par le participe passé (dans l'annexe 3, ces vocables sont placées en italiques) ;

- si l'homographie n'est pas tabulée, en face d'un participe passé en position de substantif présumé, le programme interroge l'opérateur pour confirmation ;

5.342. Les homographies entre participe passé, substantif et adjectif

On postule ici que l'homographie verbe-verbe a été tranchée sans ambiguïté en faveur du participe passé (notamment pour les verbes du deuxième groupe). La position dans la phrase détermine la catégorie.

1. précédé d'un auxiliaire "avoir" il s'agit d'un verbe au participe passé à condition que l'accord soit réalisé.

2. précédé du verbe être, la même solution est retenue si le verbe est transitif et/ou s'il admet une construction pronominale. Dans le cas contraire, nous avons un substantif ou un adjectif.

On voit que la catégorie du verbe est extensive et englobera quelques formes ambiguës. Par exemple, dans "il est tué", on pourrait considérer "tué" comme adjectif épithète. En revanche, on verra des verbes dans : "il est tué par ses ennemis, "il s'est tué" ou "il les a tués". La nuance est relativement claire pour un habitué de l'analyse mais elle est extrêmement difficile à programmer et nous avons dû y renoncer. D'une part, certains verbes peuvent contenir une dose importante d'ambiguïté. Ainsi "mourir" cité ci-dessus : "Ils sont morts" peut-être considéré comme un verbe ou un adjectif ou - pourquoi pas ? - comme un substantif ("ils sont morts de peur…") D'autre part, le locuteur dispose d'une marge importante. Il peut employer intransitivement certains verbes transitifs (par exemple : "gagner"). Pratiquement tous les verbes transitifs peuvent être utilisés pronominalement sur le modèle "se faire", "se dire" : se convaincre, se louper, se chercher…

Dès lors, nous avons adopté une règle plus simple : *sauf pour un petit nombre de verbe absolument intransitifs et non susceptibles d'une construction pronominale, la forme est rattachée à un verbe quand elle est précédée d'un auxiliaire "être"*.

3. Le participe passé peut être séparé de l'auxiliaire par :

- un ou plusieurs adverbes : "il les a *sauvagement* tués", "il les a *très probablement* assassinés", etc. Dans ces cas, il ne doit pas être confondu avec l'adjectif précédé d'un adverbe : "un homme sauvagement tué". Pour cela le test remonte à gauche à la recherche d'un substantif ou d'un déterminant (cf. ci-dessous le problème des locutions) ;

- une négation "il ne les a *pas* tués" (également : ni, plus, moins, jamais…) ;

- une locution : "il les a *un peu* tués". Une table des locutions insécables est établie. En remontant la phrase à gauche, à la recherche d'un auxiliaire, le programme vérifie qu'il n'est pas en face d'une de ces locutions ;

- si la forme est séparée de l'auxiliaire par plusieurs mots, dans lesquels ne figure pas de substantif, la présence d'une préposition à droite est un indice supplémentaire en faveur du verbe.

4. La forme se trouve à droite d'une conjonction ("et", "ou"), on examine les mots placés devant la conjonction :

- la présence d'un participe passé permet de trancher en faveur du verbe ("il les a *frappés et tués*" ;

- la présence d'un substantif ou d'un adjectif accordé en genre et en nombre permet de trancher en faveur de l'adjectif ou du substantif ;

5. Le programme ne doit pas trancher définitivement en faveur du verbe avant d'avoir examiné la possibilité d'une locution figée construite avec un substantif sans déterminant :

- avoir *prise* (sur…)

- être *partie* ("prenante", "à…", "dans…)

- citons pour mémoire : faire et prendre *parti(e)*, faire *nuit*, faire et donner *prise*, porter *atteinte*… qui peuvent poser problème au passé composé, plus que parfait, passé antérieur… Un tableau complet de ces difficultés est présenté en annexe 6.

En cas d'échec de ces cinq tests du verbe, on examine la possibilité d'un adjectif ou d'un substantif (s'il est présent à ce titre dans la table des homographes).

6. La forme se trouve-t-elle à droite d'un substantif accordé en genre et en nombre. Nous sommes alors devant un adjectif ;

7. La forme se trouve à gauche d'un substantif et à droite d'un déterminant. Si ces trois mots sont accordées et que la forme figure dans la liste des adjectifs susceptibles d'antéposition, nous sommes en présence d'un adjectif ;

8. La virgule peut servir d'auxiliaire dans la décision suivant le modèle proposé pour les conjonctions de coordination. L'examen des mots placés devant la virgule permet de clarifier certaines situations. Par exemple, on verra un adjectif dans : "Un homme gisait, sauvagement *tué*… et un verbe dans : "un homme a été frappé et *tué*…"

9. En cas d'échec ou d'ambiguïté, le programme interroge l'opérateur.

5.343. Les homographies entre participe passé et préposition

Cette homographie concerne : attendu, entendu, excepté, passé, pourvu, supposé, vu. Elle est discutée dans le chapitre 7 (§ 7.322).

*

* *

*

A l'issue de ce chapitre, nous avons opéré une distinction essentielle entre la catégorie du verbe et toutes les autres. Deux remarques de conclusion :

D'une part, la frontière suit au plus près les conventions et la pratique du français contemporain telle que nous pouvons la connaître à travers la presse, les romans, les dictionnaires ou les grammaires… Il a fallu toutefois préciser et arbitrer lorsqu'il y avait discussion. Dans ce cas, un souci de commodité et de simplification nous a souvent amené à trancher en faveur des séparations les plus nettes aux dépens d'une parfaite cohérence qui, au surplus, n'existe pas chez les grammairiens.

D'autre part, la procédure peut paraître lourde mais l'essentiel est confié à la machine. Il n'en reste pas moins que la présence de l'opérateur demeure indispensable. En ce qui concerne le traitement des homographies du verbe dans le corpus Mitterrand - notamment lorsqu'elles concernent le participe passé - notre objectif n'a pas été atteint (nous voulions que au moins neuf difficultés sur dix soient résolues par le programme). Il est vrai que F. Mitterrand a un style déroutant et que, appliquées à d'autres (C. de Gaulle, V. Giscard d'Estaing, J. Chirac), ces procédures atteignent de meilleurs rendements…

Chapitre 6.
LA LEMMATISATION DU NOM

Plutôt que de "nom" il faudrait sans doute parler de "groupe nominal". En effet l'étude du substantif ne peut être détachée de celle de l'adjectif et du déterminant qui l'accompagnent dans ce "groupe nominal". Nous y ajoutons le pronom qui, comme son nom l'indique, a un comportement très proche de celui du groupe nominal.

6.1. LE SUBSTANTIF

Ici encore le principe consiste à prendre comme base la définition courante et à rechercher une lemmatisation qui l'épouse aussi étroitement que possible tout en levant les ambiguïtés dont celle-ci peut être porteuse.

6.11. Définition

Le substantif se caractérise d'un double point de vue :

- syntaxiquement, le substantif est le noyau d'un groupe nominal. Du fait de cette fonction de noyau, c'est le substantif qui impose son genre et son nombre au reste du groupe notamment les adjectifs ; il est obligatoirement accompagné d'un déterminant (sauf quelques exceptions recensées dans le paragraphe "déterminants" de ce chapitre) ;
- sémantiquement, le genre du substantif est fixe. Si le substantif est commun son nombre est variable. En revanche, la plupart des noms propres ont un nombre unique. Les noms de peuples font exception : "un Américain", "des Américains" (sur cette distinction voir le § 1.12).

En fonction de cette définition, la lemmatisation des substantifs suit les principes généraux exposés ci-dessous.

6.12. Les règles de lemmatisation des substantifs

Le traitement assisté par ordinateur des homographies entre substantifs et adjectifs et entre ce groupe et les autres catégories obéit au règles suivantes.

6.121. Le lemme est au singulier

Puisque le nombre est variable, le lemme sera le singulier (le pluriel est considéré comme une flexion). Notons quelques exceptions :

- certains substantifs sont invariables ; sont notamment toujours au pluriel : ténèbres, obsèques, fiancailles, moeurs… Nous considérons ces noms comme invariables.
- d'autres substantifs ne portent pas de marque du pluriel (noms se terminant par **s, x** et **z** (riz, nez, prix, vis, français, fils…).
- du fait de leurs pluriels particuliers les noms composés font l'objet d'un traitement spécifique dès lors qu'ils ont été retenus comme des formes uniques dans la norme de Saint-Cloud (§ I.22).

6.122. Le genre est attaché au lemme

En vertu de ce principe général, "(un) mort" (substantif masculin) sera distingué de "(une) mort" (substantif féminin). Cet exemple de forme "bisexuée" est assez rare. Le plus souvent la

distinction suivant le genre donne naissance à deux formes différentes (mari-femme ; cheval-jument ; rat-rate, lion-lionne…) Ce mécanisme général prouve le bien-fondé de notre définition.

6.123. La règle "air"

Les homographies ne sont pas résolues au sein de la catégorie des substantifs. Contrairement à C. Muller, nous ne distinguons pas "air" (que l'on respire) et "air" (apparence) ; "vol" (d'un voleur) et "vol" (d'un oiseau), "cours" (d'eau) et "cours" (de bourse)… car il s'agit de substantifs masculins dont il serait d'ailleurs possible de montrer qu'ils ne sont pas sémantiquement totalement indépendants[1].

On pourra tenir compte de particularités attachées à un corpus donné. Ici le discours politique. Nous considérons que, dans le champ du vocabulaire politique, il existe une résistance à l'homographie qui se manifeste habituellement par l'emploi de la majuscule : "Etat" vs "état" ; "Constitution vs constitution", etc. Cette brèche dans le principe paraît dangereuse ; en tout cas, la liste de ces exceptions doit être aussi brève que possible et reposer sur des conventions typographiques bien assises.

6.124. Les substantifs "bisexués" (règle "garde")

On rencontre un certain nombre de substantifs bisexués homographes assez important (garde, livre, manche, mémoire, mort, moule, office, page, poste, tour, voile…) Lorsque le sens féminin est nettement différencié du masculin, on aura deux lemmes. Au singulier, le substantif masculin se distingue aisément de l'homographe féminin : la présence d'un déterminant ou d'un adjectif accordé permet d'opérer quelques tests simples d'assez bon rendement.

6.125. Le pluriel des substantifs bisexués

Les pluriels des substantifs homographes bisexués sont difficiles à analyser automatiquement puisque la plupart des déterminants pluriels (les, des, aux) ne portent pas de marque du genre (il en d'ailleurs de même des nombreux adjectifs). Ainsi, "les pages", "les postes", "les gardes", "les livres", "les morts" ont deux sens possibles qui doivent être, bien souvent, résolus par l'opérateur à l'aide du contexte large. Remarque : certains de ces verbes cumulent également une homographie avec le verbe (garde, poste, livre, mort…)

6.126. Les homographies propres au pluriel

Le cas des homographies propres au pluriel se rattache à la règle exposée au paragraphe précédent. Ainsi "cours" peut être le pluriel du substantif féminin "la cour", ou du substantif masculin "le(s) cours" (ou le verbe "courir", sur ce point : § 5.31). L'homographie du verbe écartée, deux solutions sont possibles :

-un autre élément du groupe nominal permet de départager les genres. Par exemple un adjectif : "des grandes cours", "des cours élevés" (voir également : fonds, fils…)

- en l'absence de cet élément, le recours à l'opérateur est nécessaire.

6.127. La règle "enfant"

Un certain nombre de substantifs sont employés indifféremment au masculin ou au féminin : "enfant", "élève", "disciple"… C'est également le cas de certains noms de professions ou de

---

1. Jacqueline Picoche montre que par suite d'un "regroupement" sémantique, les trois sens possibles de "air" ne peuvent plus être détachés (*Structures sémantiques du lexique français*, Paris, Nathan, 1986, p 82-83).

fonction (ministre, concierge, juge, procureur…). Nous avons choisis de les placer systématiquement au masculin pour éviter la constitution d'une catégorie supplémentaire et surtout de situations impossibles à dénouer dans le cas des pluriels signalés ci-dessus…

6.2. L'ADJECTIF

L'adjectif assume principalement la fonction d'attribut-épithète. Il est flexible en genre et en nombre et s'emploie généralement au comparatif et au superlatif. Suivant cette définition, les formes féminines et plurielles seront regroupées sous le même lemme (masculin singulier). Certains adjectifs ne s'emploient plus que dans une forme ou deux : "bouche bée", "pierre philosophale", "droit régalien", etc… En revanche, il faut rejeter la notion d'"adjectif invariable" inventée par les grammairiens à l'intention expresse de "bien" ("il est bien", "ils sont bien") mais ce trait "invariable" marque le caractère définitivement adverbial de "bien".

Pour le traitement par ordinateur des formes ambiguës, on tiendra compte de ce que trois grandes constructions sont possibles :

6.21. L'adjectif dans le groupe nominal

En règle générale, l'adjectif épithète est placé après le substantif. Cependant certaines formes peuvent venir s'intercaler entre le déterminant et le substantif :

- des ordinaux et des cardinaux : "les *deux* hommes", "le *deuxième* homme" (cf ci-dessous 6.4) ;
- les "adjectifs indéfinis" : aucun, certain, chaque, différents, nul, plusieurs, quelque, tout, etc. Suivant la tendance dominante chez les grammairiens, nous avons considéré que ces formes sont qualifiés à tort d'adjectifs et qu'il s'agit de déterminants. La détection de ceux-ci se fera donc par leur place dans le groupe nominal (devant le substantif). Par exemple : "un tort certain" (adj), "un certain tort" (déterminant). A ce sujet, cf § 6.413 ;
- certains adjectifs qualificatifs. Pour la plupart, il s'agit de formes usuelles qui peuvent se trouver également devant ou derrière le substantif qu'ils qualifient suivant que le locuteur met l'accent sur le substantif ou sur l'adjectif (cf ci-dessous les adjectifs antéposés) ;
- pour des raisons de style, ils peuvent être détachés du nom : "Il écoutait, attentif,…". Cette construction sera détectée par des virgules ou des tirets.

6.22. L'adjectif attribut et l'emploi adverbial

L'adjectif attribut se trouve placé derrière un verbe. Ici se pose le problème de l'homographie adjectif-adverbe. A priori, tout adjectif est susceptible d'un emploi adverbial. Cependant le souci de ne pas multiplier les difficultés nous a amené à restreindre l'examen des homographies à un petit nombre d'adjectifs : bas, bon, bref, cher, clair, dur, fort, haut, juste, mauvais, vrai…

Deux règles permettent de résoudre la majorité de ces ambiguïtés :

- derrière "être", la forme est toujours adjectif : "il est bon", "c'est cher"… Il faut ajouter à cela un certain nombre de verbes d'état : sembler, devenir, paraître… dont la liste exhaustive est difficile à établir ;
- pour les quelques adjectifs cités ci-dessus, leur emploi derrière un verbe autre que "être" - et les quelques verbes d'état mentionnés ci-dessus - leur donne un caractère d'adverbe ("aller bien", "dire vrai", "parler fort"…). Nous rattachons à ce cas les constructions pronominales comme "se montrer dur", "se faire mal", "se montrer bon", etc, dont certaines sont ambiguës.

Toutes les autres formes appartenant à cette catégorie et placées derrière un verbe sont considérées comme des adjectifs (sur cette question, voir également le § 7.131)

6.23. L'adjectif antéposé, le déterminant et l'adverbe

De manière générale nous postulons que l'adjectif vient après le substantif qu'il détermine. Un certain nombre d'adjectifs peuvent toutefois être placés avant ce substantif. Ces adjectifs qualificatifs antéposés sont parfois difficiles à séparer des adjectifs indéfinis et des adverbes.

6.231. Discussion

Il est difficile de dresser une liste limitative des adjectifs susceptibles d'être placés devant le substantifs qu'ils qualifient. A vrai-dire, l'usage est assez souple et beaucoup d'adjectifs les plus usuels peuvent être antéposés dans une construction emphatique. La liste ci-dessous doit être nourrie et complétée afin de parvenir à couvrir plus de 90% des occurrences.

Seuls les adjectifs contenus dans la liste ci-dessous peuvent être antéposés. Dans tous les autres cas, le programme considère que le nom vient d'abord. Cette règle permet de résoudre les cas du genre "une belle capitale" et "la peine capitale" (capitale est reconnue comme substantif dans le premier cas et comme adjectif dans la seconde).

6.232. Liste des adjectifs susceptibles d'une antéposition

La liste ci-dessous pourra être complétée grâce aux observations ultérieures. Nous plaçons en italiques les homographies avec des substantifs :

ancien, artificieux, *bas*, beau, bel, *blanc*, *bon*, *brillant*, *chaud*, cher, clair, *commun*, court, *curieux*, *demi*, dernier, *droit*, *drôle*, *dur*, éminent, énorme, épais, épouvantable, étrange, excellent, *faible*, *faux*, fol, formidable, *fort*, *froid*, gai, gentil, *grand*, gras, gris, gros, *haut*, heureux, honnête, honorable, honteux, horrible, immense, imperceptible, important, impossible, impressionnant, inadmissible, incomparable, indispensable, indomptable, *innocent*, inquiétant, insupportable, intelligent, intolérable, *jeune*, joli, joyeux, judicieux, *juste*, large, léger, légitime, lent, libre, lointain, long, lourd, *mâle*, maigre, magnifique, *malheureux*, *maudit*, mauvais, meilleur, *merveilleux*, minuscule, modeste, moindre, mortel, mince, nécessaire, *noble*, *noir*, nombreux, *nouveau*, nul, obscur, orgueilleux, pacifique, pâle, pareil, *parfait*, *particulier*, *pauvre*, *permanent*, *petit*, pire, *plein*, principal, *prochain*, *proche*, profond, propre, prudent, *puissant*, *pur*, rapide, rare, récent, *riche*, rigoureux, ruineux, rude, sacré, sain, *saint*, *sage*, sale, *sauvage*, *secret*, sensible, sereine, sérieux, seul, sévère, *simple*, solennel, solide, sombre, *sourd*, strict, stupide, superbe, tendre, tiède, *tragique*, tranquille, triste, ultime, *vague*, véritable, vert, *vieux*, vieil(le), vif, violent, vrai.

6.233. L'adjectif antéposé et l'adverbe

Le problème se pose pour quelque adjectifs pouvant être employés comme adverbes (cf liste plus haut : § 6.221). Si, l'un de ces adjectifs précède un autre adjectif sans en être séparé par une virgule ou une conjonction, il semble faire office d'adverbe : "il arrive *bon* dernier", "un *fort* bel homme", "le *bon* vieux temps"… Cependant rien ne distingue vraiment ces emplois des autres adjectifs. Ainsi, dans "un *parfait* honnête homme", il est difficile de voir un adverbe dans "parfait". Au total, nous avons décidé de voir dans ces constructions des adjectifs en acceptant le risque de laisser passer dans cette catégorie quelques formes ambiguës.

6.3 LES HOMOGRAPHIES DU GROUPE DES SUBSTANTIFS ET ADJECTIFS

Nous entendons par là les homographies ne comportant que deux solutions : la forme est rattachée soit aux substantifs, soit aux adjectifs.

6.31. L'homographie entre deux substantifs

En fonction de la règle "air" l'analyse des substantifs homographes n'est effectuée que lorsque ces substantifs sont de genre différent

Il faut distinguer trois cas :

1. les substantifs distingués par le genre (garde, manche, poste…) :

Attention beaucoup de déterminants masculin et féminin sont homographes au pluriel. S'il n'y a pas dans le groupe nominal d'autres indices (autres déterminants, adjectifs) permettant de trancher, le programme fait appel à l'opérateur. Pour éviter une cascade d'interrogations, il serait possible de prolonger l'analyse au groupe verbal (notamment l'accord du participe passé)…

2. les substantifs distingués par le nombre :

Par exemple, "un cours d'eau" & "des cours de ferme". Normalement le déterminant et/ou l'adjectif doivent permettre de trancher dans les mêmes conditions qu'en 1.

3. les substantifs de même nombre :

"des fils à papa" & "des fils de fer" ; "des cours d'appel" & "des cours d'eau" & "des cours de bourse"). Sauf présence d'un déterminant ou d'un adjectif permettant de trancher, il est fait appel à l'opérateur. Naturellement, des programmes plus volumineux peuvent aisément traiter ces difficultés au cas par cas en tenant compte des locutions dans lesquelles ces formes entrent habituellement…

6.32. Les homographies entre adjectifs et substantifs

On rencontre ici une des difficultés majeures de la lemmatisation : un grand nombre de substantifs peuvent aussi être employés comme adjectifs et vice-versa. Par conséquent, on pourrait songer à faire trois groupes : substantifs seuls, adjectifs seuls et (adjectif ou substantif). Mais l'importance de cette troisième catégorie serait telle qu'elle interdirait les comparaisons. Il faut donc conserver les deux catégories exclusives et lever les ambiguïtés.

La qualité d'adjectif ou de substantif n'est généralement pas intrinsèque au nom. Il lui est conféré par sa place dans la phrase. L'identification doit d'abord reposer sur des critères syntaxiques.

Par exemple : "une amie est venue" ; "amie" est reconnu comme substantif du fait de la présence d'un déterminant devant lui et d'un verbe à la troisième personne derrière. Dans "une amie fidèle", nous avons deux homographes substantif-adjectif mais le programme reconnaît sans difficulté dans "amie" le substantif et dans "fidèle" l'adjectif car "ami" ne figure pas dans la liste des adjectifs susceptibles d'être antéposés. En revanche, dans "Une main amie est venue me secourir" - "amie" est reconnu comme un adjectif puisque précédé d'un substantif auquel il est accordé en genre et nombre.

La copule placée devant la forme homographe permet une procédure du même genre : si elle vient derrière un substantif, la forme est un substantif, si elle vient derrière un adjectif, elle est un adjectif. En revanche les signes de ponctuation mineure entre la forme analysée et la forme qui précède introduit un doute : dans le cas d'énumération (plusieurs adjectifs ou substantifs séparés par des virgules) la première forme déterminera les autres : "des visages haineux, *crispés*, *révulsés*…" (adjectifs), mais : "Adieu veaux, vaches, cochons, *couvées*…" (substantifs) !

La table générale contient une liste non limitative des formes partagées en trois catégories : substantif seul ; adjectif seul et {adjectif ou substantif} distinguables par leurs emplois et leurs places dans la phrase.

6.33. Les autres homographies du groupe {substantifs-adjectifs}

Les homographies entre verbe et substantif et entre verbe et adjectif ont été examinés dans le chapitre précédent (§ 5.3).

6.331. Les homographies entre {substantif-adjectif} et adverbe

Nous avons déjà examiné cette question dans les § 6.22 et 6.233

• pour l'adjectif, nous ajoutons aux cas déjà envisagés les expressions toutes faites. Dans les expressions ci-dessous, et celles construites de la même manière, la forme est un adverbe :

- "allons bon !", "ah bon !", "bref !"…
- locutions prépositives : "en bas", "de haut", "en haut", "en vrai"…
- locutions adverbiales : "tout bas", "tout haut", "bien vrai", "…

• pour le substantif (bien, pas, bas, près…), la présence d'un déterminant ou d'un adjectif signale habituellement le groupe nominal dont le substantif homographe est le pivot. Mais ces cas comportent souvent des homographies en cascade. Par exemple : "Faites le bien" peut s'analyser comme verbe+pronom relatif+adverbe ou comme : verbe+article+substantif. Sauf présence d'autres indices dans le contexte immédiat, l'appel à l'opérateur est inévitable…

6.332. Les autres homographies du groupe {substantif-adjectif}

Cinq cas peuvent se rencontrer :

- substantif-conjonction : ce groupe concerne : or, car. Il sera traité dans le chapitre suivant sous le § 7.223 ;
- substantif (et adjectif)-préposition : avant, contre, derrière, outre, plein, proche, sauf, sous, vers… Voir le § 7.323 ;
- substantif-pronom : rien, personne… Cf. ci-dessous, § 6.574 ;
- substantif-déterminant : son, ton…. Cf. ci-dessous, § 6.432 ;
- substantif-adverbe-pronom-déterminant : tout(s,e,es), cf. § 7.137.

6.4. LES DETERMINANTS

Par rapport à la grammaire classique, la notion de "déterminant" est le seul écart que nous nous sommes permis. La notion de "déterminant" n'est d'ailleurs pas neuve (elle date de Port Royal semble-t-il) mais elle ne s'est imposée que récemment. Elle est utilisée dans pratiquement toutes les grammaires modernes mais avec une portée différente.

6.41. Caractéristiques des déterminants

6.411. Définition

Les déterminants forment un ensemble de mots qui, à l'intérieur du groupe nominal, ont un même comportement de satellite du noyau. De ce fait l'une des manières de reconnaître le déterminant homographe est la possibilité de le commuter avec un dont la catégorie ne présente pas d'ambiguïté.

Les principaux déterminants sont les articles, les cardinaux et ordinaux et les adjectifs possessifs, démonstratifs, interrogatifs, exclamatifs et indéfinis.

6.412. Les articles

L'article est toujours placé avant le nom mais peut en être séparé par un ou plusieurs mots (particulièrement, adjectifs antéposés, prépositions). On distingue :

- les articles indéfinis : un, une
- les articles définis : le, la, les

On rattache à cette seconde catégorie le partitif "des" analysé en "de"+"les" (cf § 7.325) ainsi que les articles contractés "au(x)" et "du, des" analysés en deux lemmes "à le" ou "à les", "de le" ou "de les". Cette solution a l'avantage de rétablir la symétrie entre le masculin et le féminin singulier "à la" et "de la"

6.413. Les adjectifs non-qualificatifs.

On distingue quatre catégories ayant des lois de combinaison assez diverses

6.4131. Les adjectifs possessifs :

Ils comportent trois informations (le genre, le nombre, la personne)

Possédé	singulier		pluriel
	masc	féminin	masc/féminin
Possesseur singulier	mon	ma	mes
	ton	ta	tes
	son	sa	ses
Possesseur pluriel	notre		nos
	votre		vos
	leur		leurs

NB : "leur" pronom personnel est invariable donc toutes les formes "leurs" sont automatiquement des possessifs (pronoms ou déterminants). De même l'accent circonflexe distingue les pronoms possessifs "vôtre" et "nôtre" des déterminants de la même personne.

6.4132. Les adjectifs démonstratifs (ce, cet, cette, ces)

Au pluriel le féminin ne se distingue pas du masculin. "Cet" se rattache à "ce" (problème euphonique).

NB : dans des expressions comme "cet homme-ci" "ce politicien-là", ci et là sont des adverbes. En revanche, "celui-ci" et "celui-là" - et les autres pronoms démonstratifs de même construction - sont comptés comme un seul vocable.

6.4133. Les adjectifs indéfinis

On range dans cette classe les déterminants qui expriment les nuances les plus floues ou les plus complexes (notamment la quantité non mesurée). Les adjectifs indéfinis se différencient des autres non-qualificatifs par leur nombre élevé et par la difficulté que l'on éprouve à en dresser une liste complète. Citons notamment : aucun, autre, certain, chaque, différent, même, plusieurs, quel, quelque, seul, tel, tout…

Les critères d'analyse sont les suivants :

- l'accord en nombre avec le nom : certain, quel, tout, tel, même, quelque, autre ;
- certains n'ont qu'une forme : aucun, nul, chaque…

- nous rattachons à cette catégorie les exclamatifs et les interrogatifs ("quel") qui s'accordent en genre et en nombre avec le nom-noyau de la phrase.

6.4134. Les numéraux

Les ordinaux (premier, deuxième…) et les cardinaux (un, deux…) présentent plusieurs caractéristiques :

- "un" est indissolublement article et numéral et la séparation paraît impossible en dehors de la combinaison avec d'autres numéraux (par exemple : 1981). L'homographie avec le pronom est analysée ci-dessous (§ 6.572).

- les numéraux peuvent se comporter comme des noms ou se trouver placés à gauche comme à droite de celui-ci. Ce comportement, assez semblable à l'adjectif, justifie qu'on en fasse une catégorie à part .

6.42. Les règles d'utilisation des déterminants

6.421. La portée du caractère obligatoire des déterminants

Les déterminants ont un caractère quasi-obligatoire : dans le groupe nominal, le nom-noyau est généralement accompagné d'un déterminant. La suppression du déterminant rend la phrase incorrecte. Toutefois :

• cette règle admet une exception générale : le déterminant peut être remplacé par une préposition. Ainsi dans les expressions toutes faites qu'on nomme également "locutions prépositionnelles" : en train, par avion, avec plaisir, sans envie, de grâce, à merçi, par bonheur… (ces expressions sont innombrables…) ;

• d'autre part, on rencontre de nombreux cas particuliers :

- les noms propres n'ont pas besoin de déterminant,

- le déterminant est supprimé dans les invocations (oh ! rage ! oh ! désespoir !…) ;

- le déterminant n'existe pas dans les locutions verbales : avoir envie, faire part, prendre garde… NB : on remarque que ces trois substantifs sont homographes avec des verbes (envier, partir, garder…). Un test spécial est mis en place pour les détecter (§ 5.234) ;

• le déterminant peut être également omis dans :

- les énumérations : "hommes et femmes, jeunes et vieux…" ;

- lorsque le substantif est placé en attribut du sujet : "il est président" ;

- lorsque le substantif est placé en apposition : "F.Mitterrand, président de la République" ;

- dans certaines expressions toutes faites : "Noblesse oblige" ; "par pertes et profits"…

6.422. La place du déterminant dans la phrase

Les déterminants sont toujours placés devant le nom et s'accordent avec lui (sauf les ordinaux et les cardinaux qui ont le comportement d'un adjectif). Cette règle est importante pour les tests d'homographie. Par exemple, pour analyser les homographies :

- entre substantif et verbe : l'absence d'un déterminant à gauche (ou d'une préposition) fera penser que l'on est en présence d'un verbe sauf cas particuliers énoncés ci-dessus ;

- entre déterminant, pronom et conjonction : si la forme à gauche ne s'accorde pas en genre ou en nombre avec le nom, on est probablement en présence d'un pronom et d'un verbe homographe ("le facteur le porte" : déterminant+nom+pronom+verbe).

- entre substantifs de genre opposé : l'analyse du genre ou du nombre du déterminant permet souvent de résoudre ces homographies gênantes. Par exemple, "le voile de la mariée ; la voile du navire" ; "le manche du couteau ; la manche de la veste" ; "le garde du château ; la jeune

garde" (cet exemple illustre la nécessité de tenir compte des adjectifs antéposés, cf § 6.322 ci-dessus).

Remarque : ces règles ne sont pas d'application générale car, outre les réserves exprimées au § 6.421, on notera que la graphie de certains déterminants ne varie pas en fonction du genre et du nombre.

6.423. Les combinaisons de déterminants

Les combinaisons de déterminants sont possibles et peuvent même être multiples. Toutefois des régularités sont à noter :

Les déterminants peuvent être groupés en deux catégories :

1. articles, les adjectifs possessifs, démonstratifs et interrogatifs…

2. les adjectifs indéfinis, les numéraux et les cardinaux…

Dans un groupe nominal, on peut rencontrer ensemble plusieurs déterminants de la catégorie 2 mais jamais plus d'un de la catégorie 1… ("tous *les* deux jours"). Cette règle est extrêmement intéressante pour résoudre les principales homographies mettant en cause les déterminants

6.43. Les homographies des déterminants

Les déterminants comportent de nombreuses homographies. L'analyse de ces homographies repose sur les règles énoncées ci-dessus et notamment sur le comportement assez différent des deux groupes. On utilise surtout la règle selon laquelle le déterminant du groupe 1 ne peut être séparé du nom que par un déterminant du groupe 2 ou par un adjectif antéposé ou par certaines prépositions.

6.431. L'homographie entre le déterminant et le pronom

- "le", "la", "les", "un", "une", "leur" peuvent être pronoms ou articles. C'est incontestablement l'une des principales sources d'ambiguïté (cf. § 6.573). Ce cas est également discuté sous "le" dans la dernière section de cette note (§ 6.572).

- "aucun", "autre" (pronoms) ; "tout" (également adverbe) appartiennent au deuxième groupe des déterminants : la présence d'un substantif et/ou d'un adjectif (accordés en genre et en nombre) dans le même groupe nominal peut conduire à identifier le déterminant. Par exemple : "je pense qu'*aucun* des amis ne viendra (pronom)" vs "je pense qu'*aucun* ami ne viendra (déterminant)", "Tout homme est mortel"… (Voir également § 6.573 et 7.137).

- "certain" et "nul" peuvent également être des adjectifs ("un certain tort" : adjectif indéfini ; "un tort certain" : adjectif qualificatif ; "certaines personnes" : déterminant 1). La nuance est faible et nous avons décidé de rattacher ces deux vocables aux adjectifs susceptibles d'être antéposés[1] (beaucoup d'entre eux jouent cette fonction d'indéfinis lorsqu'ils sont placés avant le substantif. Voir également § 6.573).

6.432. L'homographie entre le déterminant et le substantif

- "son", "ton" se rattachent au premier groupe des déterminants. La présence d'un autre déterminant du groupe 1 dans le contexte immédiat conduit à reconnaître automatiquement le substantif.

---

1. A la réflexion cette solution engendre plus de difficultés qu'elle ne procure d'avantages. Nous serions finalement enclins à rétablir pour "certain" et "nul" une triple codification.

-"première", "second", "seconde", appartiennent au groupe 2. Ils peuvent donc se combiner avec des déterminants du premier groupe. Seule la présence d'un autre substantif (accordé en genre et en nombre) dans le même groupe nominal peut conduire à identifier le déterminant .

6.433. L'homographie entre le déterminant et l'adjectif

• "différent", "divers" ne sont pas reconnus comme déterminants mais comme des adjectifs susceptibles d'antéposition : cette solution escamote la fonction de déterminant joué par le pluriel ("Différentes personnes…"). A la réflexion "différent" et "divers" devraient être traités comme "autre" ou "certain".

• "certain" et "nul" : cf § 6.573.

• le cas de "neuf"

Comme numéral, "neuf" :

- est habituellement associé à d'autre numéraux pour former des chiffres (cent neuf, neuf et demi…) ;
- est suivi d'un groupe nominal au pluriel ("neuf hommes") ;
- est précédé d'un déterminant au pluriel ("les neuf plus grands") ;
- est précédé du verbe "être" au pluriel ("nous sommes neuf") alors que l'adjectif serait accordé ("ces avions sont neufs") ;

Comme adjectif, "neuf" :

- ne peut être placé devant le substantif ;
- est précédé d'un substantif masculin singulier ("un homme neuf") ;
- est précédé d'un verbe conjugué à une personne du singulier ("je suis neuf dans le métier") ;

Enfin rappelons que neufs et neuve(s) sont toujours des adjectifs…

Un soin particulier doit être apporté à la résolution de ces cas. D'une part, leur très haute fréquence oblige à limiter le recours à l'opérateur. D'autre part, une erreur d'analyse risque de se répercuter sur d'autres formes : par exemple, une confusion entre l'article et le pronom risque de se répercuter sur l'analyse de l'homographie entre le substantif et le verbe… Telle est la raison pour laquelle des traitements spécifiques ont été élaborés pour certaines constructions syntaxiques particulières dont les principales ont été évoqués ci-dessus.

6.5. LES PRONOMS

Classiquement, les grammairiens définissent le pronom comme un mot qui remplace le plus souvent un nom ou un groupe nominal ; il peut aussi se substituer à un adjectif ou à une proposition toute entière. De ce fait, il se comporte comme l'élément auquel il se substitue.

6.51. La classification et la lemmatisation des pronoms

La définition purement sémantique des pronoms est de peu d'utilité car elle ne donne guère de critères permettant de désigner avec certitude les pronoms par rapport aux formes homographes. C'est pourquoi nous lui ajoutons un élément syntaxique : généralement, dans la phrase le pronom se trouve dans la même position syntaxique qu'un nom ou un groupe nominal, sujet ou objet, attribut ou épithète.

Suivant les cas, sa nature sera différente comme permet de le comprendre la classification des pronoms.

6.511. La classification des pronoms

Rappelons que, en fonction du rôle rempli dans la phrase, on distingue les pronoms :

- personnels (je, moi, tu, toi, il, elle, lui, le, soi, nous, vous, ils, elles, leur…)
- personnels réfléchis (me, te, se, nous, vous, se)
- démonstratifs (ce, ceci, cela, ça, celui, celle…)
- possessifs (mien(s), mienne(s), tien(s), tienne(s), sien(s)…)
- interrogatifs (qui, que, lequel, lesquels…)
- relatifs (qui, que, quoi, dont…)
- indéfinis (aucun, autre, autrui, certain, chacun, nul, on, personne, quelqu'un, quiconque, rien, tout, un…)

6.512. Les problèmes de lemmatisation des pronoms

La liste ci-dessus permet d'apercevoir deux difficultés :
1. Les homographies au sein des pronoms
• sémantiquement, la même forme peut correspondre à plusieurs genres ou nombres :
- la distinction masculin ou féminin est souvent impossible (tu, toi, lui, le, qui, que…)
- le nombre ne peut être défini : se, qui, que…
• certains pronoms assurent plusieurs fonctions. D'où les homographies entre :
- les personnels et les réfléchis : nous, vous
- le personnel et le possessif : leur
- les interrogatifs et les relatifs : qui, que
2. Les homographies entre les pronoms et les autres catégories. Par exemple :
- pronom-déterminant : ce, le, autre, un, tout…
- pronom-substantif : rien, personne
- pronom-conjonction : que
- pronom-verbe : tiens, tiennes…
- pronom-adverbe : tout…

6.513 Les principes généraux de lemmatisation des pronoms

Pour les homographies entre des pronoms et d'autres catégories, l'analyse est complète même pour les formes à haute fréquence (notamment "le", "ce", "un"…)

Suivant le principe général selon lequel il n'est pas fait de distinction parmi les homographes au sein d'une même classe, les homographies entre pronoms ne seront pas analysées. Cette décision implique que :

• le lemme définitif indique simplement l'appartenance de la forme à la catégorie des pronoms sans apporter plus de précisions. Sans doute y-a-t-il là une limitation qui pourrait se révéler gênante et qui devrait être dépassée si les études lexicologiques se développaient dans cette direction ;

• au cours des traitements automatiques, la codification des pronoms doit être complète. On a donc un code pour :

- les sept catégories homogènes,
- trois catégories supplémentaires (personnel et réfléchi ; personnel et possessif ; relatif et interrogatif).
- autant de code qu'il y a d'homographies entre les pronoms et d'autres catégories.

6.52 Les pronoms personnels

6.521. La classification des pronoms personnels

La forme du pronom dépend de la ou les personnes qu'ils évoquent, la ou les choses auxquelles ils font référence et la fonction remplie dans la phrase. Nous présentons dans le tableau ci-dessous la classification opérée habituellement par les grammairiens (les formes élidées - j', m', t', l' - ne sont pas mentionnée).

Tableau de classification des pronoms personnels

Personnes	Sujet	C. O. D.	C. O. I.	Autres compl.	
Singulier	1	je, moi	me	me,	moi
	2	tu, toi	te	te,	toi
	3	il, elle, on	le, la, en	lui, en, y	elle, lui, en, y
Pluriel	1	nous	nous	nous	nous
	2	vous	vous	vous	
	3	ils, elles,	les	leur, en, y	eux, elles, en, y

Nota : en position de complément du nom ou du verbe, après une préposition : moi, toi, lui, elle, nous, vous, eux, elles.

6.522. La lemmatisation des pronoms personnels

Le tableau ci-dessus suggère qu'il est difficile de se régler sur les distinctions de genre ou de nombre. La difficulté explique le grand nombre des solutions adoptées. Ainsi les critères adoptés par Juilland, Muller, Bernet et Engwall diffèrent. Bernet comme Muller mettent l'accent sur la fonction (sujet, objet…) Engwall donne la priorité au genre et au nombre (un seul lemme pour "elle", "la", "lui", "y"… mais "elles" a une autre entrée) Cette dernière position paraît logique mais entraîne un inconvénient majeur : elle oblige à interpréter manuellement des milliers de "lui","y"…

Nous proposons un compromis qui amène à une position très proche de celle adoptée par C. Muller et présente l'avantage de respecter la philosophie générale de la classification des pronoms. Dans la liste ci-dessous, la forme canonique est soulignée :

je : moi, me
tu : toi, te
il : elle, lui
nous
vous
ils : elles, leur
le : la, les
en
y

6.523. Les homographies des pronoms personnels

Plusieurs problèmes d'homographie se posent entre certains pronoms et déterminants : le, la, les, leur ou avec la préposition "en"… Ce problème est compliqué par la pluralité des fonctions que peuvent assurer ces pronoms personnels et qui s'accompagne d'une grande mobilité dans la phrase. D'une manière générale, on peut utiliser le fait que la position la plus courante du pronom se trouve soit juste devant le verbe soit entre celui-ci et un auxiliaire (pour le complément d'objet direct), avec au besoin un adverbe (ce qui ne peut être le cas d'un déterminant).

"Le" est analysé dans la section consacrée aux déterminants (§ 6.572).

Le cas de "en" est examiné dans le chapitre suivant (§ 7.322).

"Tu" (2e personne du singulier et participe passé de "taire") n'est pas attesté dans le discours politique contemporain.

6.53. Les pronoms démonstratifs

6.531. Classification et lemmatisation des pronoms démonstratifs

La forme des pronoms démonstratifs varie en fonction du genre et du nombre des êtres et des

choses qu'ils représentent. La fonction qu'ils occupent dans la phrase n'entraîne aucune variation dans leur forme…

La classification habituelle aboutit à la grille présentée dans le tableau ci-dessous. En ce qui concerne la lemmatisation de ces formes, il n'y a pas de divergeances importantes dans les conventions proposées notamment par Juilland, Muller, Lyne, Bernet ou Engwall. Dans le tableau, la forme canonique retenue est soulignée. Les formes à droite se regroupe sur la première forme de la ligne.

Formes	Singulier		Pluriel		Neutre
	Masculin	Féminin	Masculin	Féminin	
Simples	celui	celle	ceux	celles	ce (c')
Composées	celui-ci	celle-ci	ceux-ci	celles-ci	ceci
	celui-là	celle-là	ceux-là	celles-là	cela (ça)

6.532. Le cas de "ce"

La catégorie des pronoms démonstratifs ne comporte qu'une seule homographie : "ce".

Le pronom démonstratif "ce" peut remplacer un mot, un groupe de mot, voire une proposition entière. Il est généralement suivi d'un verbe et le plus souvent il s'agit de "être". Quand il signifie "une chose", "un événement", il est nécessairement suivi d'une relative qui en précise la signification. Très souvent l'utilisation emphatique entraîne une relative et l'emploi d'un pronom ("c'est… qui, c'est… que"), "que" est alors pronom relatif… En présence d'un "que", l'un des tests possibles sera donc de remonter à gauche à la recherche d'un "c'".

En revanche, l'adjectif démonstratif est toujours suivi d'un groupe nominal masculin singulier pouvant comprendre : substantif, adjectif, déterminant du groupe 2 dans des combinaisons plus ou moins complexes. Par exemple : "ce même beau jour". De ce fait, le programme commence par ce test. La recherche du pronom n'est déclenchée qu'en cas d'échec du test du déterminant.

6.54. Les pronoms relatifs

6.541. Définition

La construction relative est un moyen permettant à un nom - l'antécédent - qui a déjà une fonction dans une proposition d'en assurer une autre à l'intérieur d'une proposition différente. Pour atteindre ce résultat, on fait suivre le nom d'un pronom dit "relatif".

Cette construction pose deux problèmes :

- elle occupe une place fluctuante dans la phrase. Ainsi certains pronoms relatifs peuvent être placés entre le sujet et le verbe…
- elle comporte des formes résultant de la contraction de plusieurs vocables.

6.542. Les pronoms relatifs simples

• Les relatifs simples sont invariables en genre et en nombre. Ce ne sont en réalité que des formes diverses du "qui" et ils suivent généralement le nom auxquels ils sont relatifs.

On distingue suivant la fonction :

- qui : en fonction sujet de la relative
- quoi : forme neutre de "qui" employée habituellement en complément d'objet.
- que : en complément d'objet direct
- où : en complément circonstanciel de lieu, parfois de temps (homographie avec l'adverbe de lieu)
- dont : en complément de nom

suivent généralement le nom auquel ils sont relatifs

• lemmatisation des relatifs simples.

Etant donné leur profonde parenté, on pourrait choisir de les regrouper sous "qui" (c'est la solution partiellement adoptée par Engwall qui en sort toutefois "où" et "dont").

Finalement, nous avons choisi de les maintenir chacune sous un lemme différent. Cette solution se justifie par leur fréquence et par des constructions bien différentes qui manifestent la séparation nette de leurs fonctions.

6.543. Les pronoms relatifs composés

• Les relatifs composés sont variables en genre et en nombre. Ils proviennent de la contraction de plusieurs formes. On trouve :

lequel (le+quel) : lequel, laquelle, lesquels lesquelles

auquel (à+le+quel) : à laquelle, auxquels, auxquelles…

duquel (de+le+quel) : de laquelle, desquels, desquelles…

• la lemmatisation des relatifs composés

En bonne logique, il faudrait décomposer ces formes contractes en autant de lemmes. Cependant, on remarquera que "lequel" et ses flexions sont toujours collés en une seule forme. En revanche la fusion avec les prépositions "à" et "de" n'est pas complète puisque on retrouve deux formes au féminin. Dès lors deux solutions sont possibles :

- trois lemmes "incomplets" (lequel, auquel, duquel) ; "à laquelle" et "de laquelle" étant analysés en deux formes, la première rattachée à la préposition, la seconde au pronom "lequel". Cette solution sera obligatoirement retenue si l'on a décidé d'avoir le même nombre de vocables que de formes (G. Engwall) ;

- un seul lemme : "lequel". Les formes contractes étant systématiquement décomposées. Ainsi "auxquelles" est lue comme "à"+"lesquelles" ; "desquels" comme "de"+"lesquels", etc. Cette solution a l'avantage d'être cohérente avec celle adoptée pour traiter "du" et "des". Elle est préconisée par Muller et Bernet. Elle présente l'inconvénient de conduire à un nombre différent de vocables et de formes (sauf si l'on opère cette décontraction dès le premier stade des traitements). Nous avons retenu cette dernière solution.

6.55. Les pronoms possessifs

Le pronom possessif varie en fonction de la personne, du genre et du nombre. La personne dépend du possesseur ; le genre et le nombre dépendent de l'objet possédé.

Personne qui possède		l'objet possédé			
		Singulier		pluriel	
		Masculin	féminin	Masculin	féminin
Singulier	1 pers	le mien	la mienne	les miens	les miennes
	2 pers	le tien	la tienne	les tiens	les tiennes
	3 pers	le sien	la sienne	les siens	les siennes
Pluriel	1 pers	le nôtre	la nôtre	les nôtres	les nôtres
	2 pers	le vôtre	la vôtre	les vôtres	les vôtres
	3 pers	le leur	la leur	les leurs	les leurs

La forme soulignée est celle sous laquelle sont lemmatisées les formes à gauche

Deux problèmes d'homographie sont posés :

- "leur(s)" peut aussi être pronom. D'une manière générale le pronom se distingue de l'adjectif possessif par la présence devant lui des déterminants 'le, la, les' et par le fait qu'il ne

peut être suivi d'un substantif. Il est généralement suivi d'un verbe accordé en genre et en nombre avec l'objet possédé…

NB : les pronoms possessifs de la 1e et 2e personnes du pluriel se distinguent des adjectifs possessifs non seulement par la présence de 'les' devant mais aussi par l'accent circonflexe sur le o (à vérifier lors des relectures…)

- "tiens", "tiennes" (également verbe tenir à la deuxième personne du singulier). Le verbe n'est attesté que par la présence d'un pronom personnel "tu" à gauche ou à droite (interrogatif) ou employé seul et suivi d'un point d'exclamation.

6.56. Les pronoms interrogatifs

6.561. Les caractéristiques particulières des pronoms interrogatifs

La forme varie en fonction de l'objet de la question. Si elle porte sur une personne, on emploie "qui" ; si elle porte sur un objet : "que" ou "quoi" :

- "que" ne peut être employé dans la fonction sujet ;
- "qui" peut se combiner avec toutes les prépositions ;
- "que" ne s'emploie pas en complément d'objet indirect : "quoi" ;
- "qu'" remplace toujours "que" jamais "qui" : "qu'avez-vous fait ?" (vous avez fait quoi) mais "qui avez-vous vu ?"

Des formules de renforcement sont souvent utilisées (surtout à l'oral) : "Qui est-ce qui…", "Qu'est-ce que", "à (ou de) qui est-ce que"… Dans ces formules "ce" et "que" sont des pronoms…

6.562. La lemmatisation des pronoms interrogatifs

Selon Engwall, l'ensemble de ces pronoms doit être regroupé sous qui quelle que soit la fonction. On peut aussi envisager de regrouper "quoi" sous "que". Cette dernière solution a l'avantage de la simplicité et de la cohérence (différence dans le genre et la personne) mais ne respecte pas le principe posé plus haut pour les relatifs. La solution la plus raisonnable consiste donc à conserver trois lemmes : qui, que (qu'), quoi… Là encore cette dernière solution est dictée par la haute fréquence et par le principe de départ consistant à refuser d'analyser les formes pronominales homographes.

6.57. Les pronoms indéfinis

6.571. Caractéristiques des pronoms indéfinis

La plupart des mots que la tradition range sous ce titre de "pronoms indéfinis" ne représentent aucun mot, aucune proposition, aucune idée précisément exprimés dans le discours. Ce ne sont pas des pronoms au sens propre du terme mais des "nominaux" comme les appelle Grevisse. De là vient leur grande diversité et les nombreux problèmes d'homographie.

Dans cette catégorie, on trouve essentiellement : aucun, autre, autrui, certain, chacun, nul, on, personne, quelqu'un, quiconque, rien, tout, un…

6.572. L'homographie entre le pronom et le déterminant : "le"

Par sa fréquence, il s'agit de la première source d'homographes et sa résolution détermine en bonne partie la qualité de la lemmatisation. La règle "pas plus d'un déterminant du premier groupe dans le groupe nominal" permet de résoudre la majorité des cas (à condition naturellement que la ponctuation ait été correctement réalisée).

• "Le" est un pronom :
- quand il est compris dans un groupe verbal sans ambiguïté ("je le dis") ;
- quand il est encadré par d'autres pronoms ("je le leur dis" : ici se trouve également résolu l'homographie de "leur" en faveur du pronom) ;
- quand, après échec les tests du déterminant, on constate qu'il précède un verbe transitif fléchi. C'est pourquoi la résolution de cette homographie est souvent liée à celle des formes conjuguées du verbe (cf. notamment § 5.31). Dans tous ces cas, il est nécessaire de passer aux tests du déterminant présentés ci-dessous.

• "Le" est un déterminant quand :
- il précède un substantif, un adjectif antéposé ou un numéral (de même genre et même nombre) ;
- devant un nom propre ("la Renault") ;
- devant un adjectif féminin dans une locution ("à la française", "à l'anglaise"…) ;
- devant un adverbe de comparaison ("le plus", "le moins", "le mieux", "le pire", "le pis"…) ;
- devant un pronom personnel possessif ("le mien", "le tien", "le sien"…) ;
- devant un pronom "l'un", "l'autre", "le même", "le tout"…
- "la plupart" est un nom féminin ;
Ces 9 critères résolvent la majorité des cas (plus de 7 fois sur 10, "le" est un article…)

• L'homographie "leur" présente certains traits communs mais la solution est beaucoup plus simple. On commence également par la recherche du pronom. Le pronom personnel est invariable. Il toujours placé dans un groupe verbal en position de complément avant le verbe. Le pronom possessif est nécessairement précédé d'un déterminant de catégorie 1 : le(s) ("des" étant analysé en "de les"). En dehors de ces constructions, "leur(s)" est toujours déterminant.

6.573. Les autres homographies entre le pronom et le déterminant

Le principe de l'analyse a été présenté ci-dessus (§ 6.431). Ajoutons que deux catégories peuvent être distinguées :
• "nul", "aucun", "tel. Comme déterminants, ils sont généralement employés seuls devant un substantif ("aucun homme"). Le test de l'accord joue donc un rôle clef.
• "autre" et "certain"[1]. Comme déterminants ils sont accompagnés d'au moins un autre déterminant (articles, possessifs, adjectifs indéfinis…) : "un autre homme". Le test de la compatibilité entre les déterminants peut être également employé.

En outre, deux cas particuliers doivent être mentionnés :
• "même" : déterminant (adjectif indéfini) devant le nom ; adverbe dans une série de locutions et pronom indéfini (les mêmes, aux mêmes…) ;
• "un" n'est pronom que dans quatre constructions :
- "un… de…" : "un de mes amis", "un parmi la foule") ;
- "l'un …de…" : "l'un de mes amis", "l'un d'eux") ;
- "un…qui…" (ou que) : "En voilà un qui n'est pas fier" ;
- "l'un…+ verbe" : "l'un dit blanc, l'autre noir"
Pour le reste, "un" est toujours déterminant (nous avons indiqué plus haut que l'homographie entre l'article et le numéral n'est pas analysée (§ 6.4134). Enfin, rappelons que "uns" et "une(s)" sont toujours des pronoms.

---

1. Dans l'index placé à la fin du *Vocabulaire de F. Mitterrand*, il y a un pronom "certain". Vérification faite, cette codification est erronée. Elle entraîne des distinguos trop complexes et superflus. Il est préférable de n'admettre comme pronoms que "certains" et "certaines"…

• "Tel" n'est pronom que quand il vient en tête de phrase (ou dans la locution "un tel") et qu'il est suivi d'un groupe verbal ou d'un pronom relatif. Dans tous les autres cas il est un déterminant. Tels, telle(s) sont toujours adjectifs indéfinis.

6.574. L'homographie entre le pronom et le substantif

• "personne" s'emploie en pronom à la place de "quelqu'un" ou comme équivalent de "quiconque", "aucun", "nul" :

- il est sujet du verbe ou complément d'objet. Dans la fonction sujet, il est toujours accompagné d'une construction négative ;

- nous avons considéré que "personne" employé sans déterminant était toujours un pronom.

Le substantif est précédé d'un déterminant et parfois d'un adjectif antéposé au féminin ("une grande personne"). Au pluriel, il s'agit toujours d'un substantif.

• "rien"

- substantif lorsqu'il est précédé d'un déterminant ("un rien"…) ou d'un adjectif antéposé : "un petit rien". Naturellement "riens" est toujours substantif.

- pronom indéfini, il est le plus souvent construit avec "ne" ("Qui ne risque rien n'a rien") ou précédé d'une préposition (presque _, de _, moins que_…), d'un adverbe de manière (absolument _…). On le rencontre également en tête d'une locution prépositive (_ de, _ que…) "rien moins que…", "rien de moins que…" et dans des constructions familières (Ne + verbe + pas + rien) : "ce n'est pas rien", etc ;

- l'adverbe n'est pas distingué du pronom.

Le cas de "tout" est examiné dans le chapitre suivant à propos des adverbes (§ 7.137)

*
* *
*

Le cadre qui vient d'être présenté ne prétend pas épuiser les problèmes posés par l'étude du groupe nominal. Nous voudrions faire deux remarques pour conclure.

Avec les verbes usuels, le substantif est le lieu par excellence de la polysémie. Les conventions adoptées interdisent que l'on puisse rendre compte de ce problème dans toute sa complexité. Par exemple, on a bien conscience qu'il y a plus qu'une nuance entre "le pouvoir" et "les pouvoirs". Une lemmatisation brutale fait perdre ces nuances. C'est ici que la technique de la "lemmatisation dans le texte" trouve un de ses justifications. Elle permet en effet au lexicologue de revenir au texte et de comparer les contextes de ces deux vocables. Si l'on en était resté aux formes ce retour aurait été rendu très difficile par le parasitage de l'infinitif "pouvoir". Ainsi, nous avons relevé dans le corpus Mitterrand : 80 "pouvoir" (verbe à l'infinitif) ; 104 "pouvoir" (substantif masculin singulier) et 50 "pouvoirs" (substantif masculin pluriel). Le programme de concordance les éditera séparément.

Pour les pronoms, et notamment les personnels, la situation est encore plus difficile puisque leur contexte d'emploi est susceptible de leur donner une infinité de nuances. Cependant, nous avons acquis une certitude : la lemmatisation est indispensable à cette étude. Par exemple, dans une phrase où est employé un pronom personnel sujet, on a beaucoup de chances de rencontrer, à proximité, un verbe à la forme active et cette probabilité est beaucoup plus grande que dans la moyenne des phrases du corpus : des tests statistiques réalisés sur des formes - c'est-à-dire sans que puisse être prise en compte leurs catégories grammaticales - ne pourra donc donner une idée exacte du contexte du pronom en question. Grâce à la lemmatisation dans le texte des "formes-outils" à très haute fréquence comme les pronoms et les déterminants, on peut enfin envisager d'analyser leur univers lexical pour mieux connaître la manière dont ils sont réellement utilisés.

CHAPITRE 7
LES MOTS INVARIABLES
ADVERBES - CONJONCTIONS - PREPOSITIONS

Leur caractère invariable amène souvent à considérer ensemble ces trois catégories. Elles ont en commun plusieurs autres caractéristiques qui établissent une véritable parenté. C'est pourquoi, en principe, nous avons renoncé à résoudre les homographies entre ces trois groupes.

7.1. LES ADVERBES

L'adverbe est un mot invariable que l'on joint à un mot, ou à un groupe de mots, pour en modifier le sens. On peut étudier les adverbes du point de vue de leur formation ou de leur fonction dans la phrase.

7.11. La formation des adverbes.

A part les adverbes usuels hérités du latin ou du vieux français (§7.113 ci-dessous), trois grands procédés de formation des adverbes méritent d'être évoqués rapidement car ils ont des conséquences sur la délimitation des frontières séparant les adverbes des autres catégories et sur les procédures de reconnaissance.

7.111. La dérivation

La forme normale de création de l'adverbe est la dérivation par ajout d'un suffixe : *-ement, -on, -ons* (dans les seules locutions : "à califourchon", "à reculons"). Les adverbes de manière dérivés d'un substantif ou d'un adjectif ou d'un verbe+*ement* sont proprement innombrables et il s'en fait de nouveaux tous les jours. Les dictionnaires inverses permettent d'en établir une liste indicative à partir de la terminaison "-ment"[1]. Ces adverbes sont chargés dans la table générale afin que le programme puisse les reconnaître. Aucune homographie majeure n'est à signaler dans ce groupe.

7.112. La composition

La composition est pour les adverbes l'autre lieu de la créativité lexicale. Tantôt l'usage aboutit au regroupement en un seul mot (bientôt, aussitôt, davantage, dedans,…), tantôt les éléments composants sont restés séparés (en dehors, au-delà…). Ainsi se sont formées des locutions adverbiales extraordinairement nombreuses en français. Par exemple, avec les prépositions "au" ou "par" : au-deçà, au-dedans, au-dehors, au-delà, au-dessous, au-dessus, au-devant, par-delà, par-dessus, par-dessous… Suivant la règle posée dans notre première partie, nous ne retenons que les formes unies par un tiret (comme ci-dessus les locutions composées avec au-) ou celles dont l'un des éléments n'a plus d'autre emploi que dans l'expression en question (règle "parce que"). Cette dernière convention pose le problème des locutions adverbiales.

---

1. Alphonse Juilland, *Dictionnaire inverse de la langue française*, La Haye, Mouton, 1965. Egalement, dans Gunnel Engwall, un index inverse des formes du *Vocabulaire du roman français (op. cit.)*. On peut enfin se reporter à John Chandioux, Conrad Sabourin, *l'adverbe français : essai de catégorisation*, Paris, Jean Favard, 1977.

7.113. Les locutions adverbiales

Ne sont pas codifiées comme des locutions figées des formules comme : "à côté de", "à part", "de suite", "en effet"… parce que les substantifs qui les composent conservent leur emploi original et donc leur existence propre… Ceci élimine quasiment toutes les locutions adverbiales excepté les locutions latines. Cette décision entraîne des conséquences qui, du point de vue sémantique, ne sont pas toujours cohérentes : par exemple, on ne verra qu'un mot dans "grosso-modo" et deux dans "en gros"… Mais, comme nous l'avons indiqué plus haut, les critères sémantiques doivent tenir une place aussi faible que possible car nous voulons construire une norme synthétique qui limitera au maximum la marge laissée à la libre interprétation de l'opérateur, source inévitable d'erreurs et d'incohérences

Cependant, il nous a paru intéressant d'isoler certaines locutions adverbiales comme "d'abord" ou "d'accord"…

- "d'abord". Nous avons rencontré quelques attestations du substantif précédé de la préposition "de" : dans ce cas le nom est toujours au pluriel ("d'abords difficiles"). Il n'y a donc aucun recouvrement entre la locution et le substantif. La règle "parce que" s'applique donc.

- "d'accord" est une locution fréquente dans le français parlé. Ses emplois peuvent être très significatifs. Nous avons considéré qu'il était nécessaire de l'isoler des emplois du substantif "accord". Considérant que la locution adverbiale est bien présente dans "être (ou ne pas être) d'accord", le seul recouvrement attesté vient des formulations négatives construites sur le modèle "ne pas trouver d'accord"… Leur traitement est aisément programmable.

- "d'ailleurs". Dans le même ordre d'idée, on pourrait trouver intéressant d'isoler la locution adverbiale "d'ailleurs" fréquente en langue parlée. Elle est aisément distinguable de "ailleurs" également adverbe (dans "aller ailleurs", "par ailleurs"…). Une routine du programme devrait alors traiter le cas de "venir *d'ailleurs*" seul attestation - à notre connaissance - de l'adverbe précédé de la préposition "de" et homographe de la locution adverbiale qui vient d'être isolée. Ayant posé en principe que nous n'analyserions pas les homographies au sein d'une même catégorie, nous y avons renoncé dans notre dépouillement du vocabulaire de F. Mitterrand. Mais, comme on le voit, cela crée une certaine discordance avec le sort fait aux locutions "d'abord" et "d'accord".

7.114. Les adjectifs en emplois adverbiaux

Dans la langue parlée, pratiquement n'importe quel adjectif courant peut être employé comme adverbe : "l'avoir mauvaise" ; "boire sec" ; "manger gras (ou "maigre" ou "léger")" ; "frapper fort", "parler vrai", etc. Mais ces formes restent essentiellement des adjectifs et n'acquièrent pas un sens particulier dans cet emploi adverbial. En revanche, un petit nombre d'adjectifs ont donné naissance à des adverbes employés comme tels : ils apparaissent non plus dans une ou quelques constructions figées (comme celles citées plus haut) mais dans une infinité de combinaisons possibles. Dans ce cas, il devient nécessaire de dénouer ces homographies. Cependant distinguer l'emploi adverbial de l'adjectif est très difficile puisque l'attribut se place normalement après le verbe tout comme l'adverbe. Dès lors il convient de limiter strictement cette source d'homographie où l'opérateur risque d'être assez souvent mobilisé au secours de la machine et où les frontières sont parfois floues entre l'adjectif et l'adverbe. Pour éviter les flottement, on a décidé de limiter l'analyse à un nombre de cas énumérés restrictivement ci-dessous. En dehors de cette liste, les emplois "adverbiaux" de l'adjectif ne seront pas pris en compte.

7.115. Liste des principaux adverbes usuels

Nous plaçons en italiques les homographes :
ailleurs, ainsi, assez, autour, avec, *bas*, *beau*, *bien*, *bon*, *bref*, certes, *cher*, *clair*, comment, *court*, davantage, debout, *dehors*, demain, *demi*, *dessous*, *dessus*, *droit*,*ensemble*, *environ*, exprès, *faux*, *fort*, franco, *grand*, gratis, *gros*, *haut*, hier, impromptu, incognito, *juste*, là, *loin*, *maintenant*, *mal*, *même*, mieux, moins, *net*, où, *pas*, *petit*, peu, pis, *plus*, plutôt, *point*, prou, quand, quasi, recta, *soudain*, *si*, tant, tard, tôt, *tout*, très, vite, voire, volontiers, *vrai*, *y*.

7.12. La classification des adverbes

La classification habituelle des adverbes est fondée sur le sens qu'ils impriment au mot ou au groupe de mots auxquels ils sont adjoints (manière, temps, lieu, etc). Cette classification sémantique nous est de peu d'utilité dans l'élaboration de procédures automatisées de lemmatisation et de résolution des homographies. Celles-ci vont reposer sur les positions possibles de l'adverbe de la phrase et dans les règles de combinaison entre eux et avec les autres catégories.

7.121. La position de l'adverbe dans la phrase

Les règles générales sont les suivantes :
- avec un verbe à un temps simple, l'adverbe est normalement placé après le verbe, sauf construction négative ;
- avec un verbe à un temps composé, l'adverbe se place souvent entre l'auxiliaire et le participe mais les exceptions sont nombreuses. La place normale de l'adverbe de lieu est après le participe (il a été partout, il est ailleurs…) ;
- "ne" précède toujours le verbe ainsi que les "adverbes pronominaux" (que nous classons comme pronoms : "en, y") sauf à l'impératif ("vas-y", "prends-en") ;
- l'adverbe se place, en général, avant le substantif, l'adjectif ou l'adverbe qu'il détermine ;
- l'adverbe de manière se place parfois en tête de la phrase ou du membre de phrase qu'il détermine.

7.122. Les règles de combinaison des adverbes

Pour compléter les principes généraux énoncés ci-dessus, nous employons la classification proposée par Martinet[1] qui repose sur le comportement particulier de chacun des adverbes et qui permet de formaliser quelques règles d'analyse dont certaines se révèlent assez productives.

Groupe 1 (Seulement). Les adverbes de cette classe remplissent toutes les fonctions possibles de l'adverbe. Ils déterminent à la fois le nom, l'adjectif, les adverbes, les pronoms et les verbes : aussi, non…*plus*, ni…ni, *juste*, *même*, plutôt, seulement, *surtout*.

Groupe 2 (Ailleurs). Les adverbes de ce groupe peuvent se construire avec le verbe, les adjectifs et la plupart des autres adverbes. Ils peuvent également se construire avec "de" ou "en". Dans cette construction, ils accompagnent des verbes ou déterminent des substantifs : ailleurs, alentour, alors, après, aujourd'hui, autrefois, *bas*, *dedans*, *dehors*, demain, *derrière*, *dessous*, *dessus*, *devant*, hier, *haut*, ici, jadis, là, là-bas, *maintenant*, naguère, partout, toujours.

---

1. André Martinet, *Grammaire fonctionnelle du français*, Paris, Didier, 2e édition, 1984, p 135-136.

Groupe 3 (Plus). Outre les verbes et les adjectifs, les adverbes de cette catégorie, seuls ou combinés avec d'autres adverbes, peuvent déterminer des cardinaux : *plus*, moins, trop. Dans ce cas, ils sont déterminés par les adverbes : beaucoup, peu, tellement…

Groupe 4 (Assez). Par rapport au groupe 3, ces adverbes ne sont pas susceptibles de déterminer des cardinaux et ne sont pas déterminables par "beaucoup", "tellement", "peu"… : autant, aussi, beaucoup, très, tant, si, tellement… Certains peuvent déterminer un groupe verbal quand celui-ci contient un pronom "en".

Groupe 5 (Souvent). Ces adverbes sont déterminables par ceux des deux groupes précédents (groupes 3 et 4). On peut distinguer parmi eux :

- les adverbes de manière composés grâce au suffixe "-ment" ;
- *bien*, *fort*, loin, longtemps, *mal*, *près*, souvent, tard, tôt, vite, volontiers…

Cette dernière règle peut aider à régler un grand nombre d'homographies de l'adverbe.

Groupe 6 (Ensemble). Ces adverbes n'admettent pas de combinaisons avec les adverbes des groupes 3 et 4. Ils ne peuvent pas déterminer des noms par l'intermédiaire de "de" : ainsi, autour, bientôt, certes, ci-contre, comment, debout, déjà, désormais, encore, enfin, *ensemble*, ensuite, exprès, jamais, là, parfois, peut-être, pourquoi, quelquefois, *soudain*…

Dans les tables, les adverbes devront donc être accompagnés du numéro de leur group, ce classement pouvant aider l'analyse de la phrase et le dénouement des homographies.

7.13. L'homographie des adverbes

L'homographie de l'adverbe avec la préposition est discutée ci-dessous au § 7.3.

Plusieurs autres cas sont possibles appelant chacun un traitement spécifique :

7.131. L'homographie entre l'adverbe et l'adjectif

Les cas retenus sont les suivants : bas, beau (bel), bref, cher, clair, faux, fort, grand, gros, haut, soudain, vrai… La plupart appartiennent au groupe 2 ci-dessus, ce qui permet de résoudre sans problème l'homographie dans la plupart des cas :

1. lorsque l'adverbe se trouve situé entre l'auxiliaire et le verbe ("il a beau faire", "c'est cher payé"…) ;
2. lorsque l'adjectif se trouve en épithète dans un groupe nominal (avant ou après un substantif) ;
3. lorsque l'adverbe est employé dans certaines constructions particulières comme :

- les constructions impersonnelles ("il fait bon", "il pleut fort", "c'est bon", "c'est fort"…) ;
- après le verbe dans des expressions toutes faites dont l'inventaire ne peut être complet mais qui mérite pourtant d'être effectué aussi soigneusement que possible étant donné leur haute fréquence. Par exemple pour "bas" : "mettre bas", "parler bas", "tomber bas", "viser bas", "voler bas"… ;
- les constructions avec des prépositions comme "de" et "en" (caractéristiques du groupe 2). Par exemple, pour "bas" : "en bas", "de bas (en haut)", "à bas"…

4. lorsque l'adjectif attribut supposé n'est pas accordé à la personne du verbe ("ils marchent droit", "elles parlent fort"). Il s'agit alors avec certitude d'un adverbe ;
5. l'adjectif attribut suit un verbe d'état (être, paraître, sembler, devenir…) et ne précède pas un adjectif ou un substantif. Par exemple : "cet homme est *fort*" (adjectif) mais "cet homme est *fort* aimable" (adverbe)… Comme nous l'avons indiqué ci-dessus, il est impossible de dresser une liste complète des verbes d'état. C'est ici que se situe l'une des deux difficultés ;

6. l'adjectif attribut supposé accompagne un nom ou un adjectif. Si les règles de composition (§7.122) ne permettent pas de départager avec certitude l'adjectif de l'adverbe :

- le substantif ou l'adjectif ou le déterminant ne sont pas au masculin singulier et l'on peut trancher avec certitude en faveur de l'adverbe ;
- la forme suit immédiatement un substantif masculin singulier : adjectif
- la forme se trouve placée devant un substantif au masculin singulier : adjectif
- la forme précède un adjectif au masculin singulier épithète ou attribut : adverbe ("un fort bel homme", "un objet cher payé"…).

La plupart des adjectifs cités ci-dessus ne peuvent être employés adverbialement dans ces constructions ce qui limite beaucoup les difficultés à soumettre à l'opérateur ;

7. L'homographie avec l'interjection n'est pas analysée (bon!, bref!, bien!…)

7.132. L'homographie entre l'adverbe et le substantif

La liste des cas retenus est la suivante : ailleurs, arrière, bas, bien, bon, dehors, demi, dessous, dessus, droit, ensemble, faux, fort, haut, impromptu, juste, mal, net, pas, petit, point.

Les règles exposées dans le paragraphe ci-dessus s'appliquent mais doivent être complétées par l'analyse de l'homographie au sein de l'ensemble {adjectif-substantif} : sur ce point cf la note sur les substantifs (§ 6.32).

Il faut tenir compte de ce que la plupart de ces adjectifs :

- entrent dans de nombreuses expressions toutes faites ;
- se placent souvent avant le substantif ("bas étage", "bonnes affaires"…").

7.133. Le cas de "bien"

"Bien" est l'un des cas d'homographie les plus fréquents. On le rencontre en :

• adverbe de manière :

- il est placé devant ou derrière de nombreux verbes dans des emplois très courants : "faire bien", "vouloir bien", "devoir bien", "penser bien", "(se) sentir bien", "savoir bien", "aller bien", etc…
- il peut être employé en renforcement d'un autre adverbe : "tout à fait bien", "très bien", "trop bien", "bien mieux", "bien pis", "bien bas", "bien longtemps", "bien plus", etc…
- il peut s'associer à une préposition : "ou bien", "ni bien ni mal", "bien devant", "bien derrière", "bien après"…
- dans des interjections : "eh bien ! ah bien ! oh bien ! etc…
- associé à une conjonction pour former une sorte de locution conjonctive : "bien que", "ou bien",

Dans tous ces cas, la liaison entre les éléments n'est pas suffisante pour que la locution puisse être traitée comme un seul mot.

• "adjectif invariable" : habituellement placé en position d'attribut après un verbe d'état : "c'est bien", "il se sent bien"… Cet emploi se distingue difficilement de l'adverbe et le caractère invariable marque bien le glissement progressif de l'adjectif dans la fonction adverbiale… C'est pourquoi nous rattachons ces cas à l'adverbe.

• substantif masculin :

le substantif se distingue aisément de l'adverbe par la présence d'un déterminant et, éventuellement d'un adjectif à ses côtés ("faire le bien", "le souverain bien"). Cependant, une analyse spécifique paraît indispensable du fait de la plasticité de l'adverbe. Par exemple, "faites le bien" peut s'analyser comme (verbe+pronom+ adverbe) aussi bien que comme (verbe+déterminant+substantif). Si l'on veut éviter des erreurs, un recours assez fréquent à l'opérateur sera inévitable.

7.134. Le cas de "pas"

"Pas" est un cas extrêmement fréquent :

• l'adverbe ne pose pas de problèmes mais se présente dans un grand nombre de constructions possibles :

- il suit normalement un verbe fléchi à condition qu'une négation précède ce verbe ("ne… pas") ;

- un adverbe de manière peut d'intercaler entre le verbe et "pas" ("il ne veut *absolument* pas") ;

- il est placé devant un pronom personnel pour former une interjection ("pas moi !" ; "pas lui !") ;

- il peut être employé avec une préposition ou un autre adverbe pour former une locution ou une interjection ("surtout pas!", "même pas", "certainement pas" …) ;

• le substantif est généralement précédé d'un déterminant ou d'un adjectif antéposé au masculin. Une difficulté vient de ce que le substantif ne porte aucune marque du pluriel. De plus, il peut être employé sans déterminant dans certaines locutions généralement construites avec "à" : "pas à pas", "mettre au pas", "à pas de velours", "à pas de géant", "à pas comptés", "à petits pas"…

7.135. L'homographie entre l'adverbe et le verbe

• "maintenant" (participe présent du verbe "maintenir") et "plus" (formes conjuguées du verbe "plaire"). Ces verbes ne sont pas attestés dans nos corpus et notamment dans le corpus Mitterrand. C'est pourquoi nous avons renoncé pour l'instant à mettre en oeuvre une procédure spécifique qui se serait essentiellement inspirée de celle mise en oeuvre pour les adjectifs et substantifs (§ 5.33) ;

• "quitte" (verbe "quitter"). Les emplois adverbiaux (quitte à, être quitte, tenir quitte…) ne sont pas attestés dans nos corpus mais semblent assez fréquent dans la langue littéraire…

7.136. Le cas de "y" et de "où"

L'homographie avec le pronom ("y", "où") apparaît insoluble à l'aide de critères de morphologie de la phrase puisque, d'après les grammairiens :

- "y" est pronom quand il renvoie à des personnes, des animaux, des objets des idées ("nous y pensons") ;

- "y" est adverbe quand il renvoie à des lieux : "il s'y rend", "il y va".

On voit que la position dans la phrase est souvent très proche. Etant donné la haute fréquence des "y", il a été décidé de ne pas séparer le pronom de l'adverbe et de regrouper l'ensemble de ces formes sous le pronom.

Quand à "où", si le pronom interrogatif ne pose pas de problème, les écarts entre les grammaires et les dictionnaires sont tels que, là encore, nous avons renoncé à séparer le pronom de l'adverbe.

7.137. Une quadruple homographie : "tout(e,es,s)"

Troisième cas d'homographie par sa fréquence, "tout" présente un problème d'une rare complexité car il peut être rattaché à quatre catégories différentes : adverbe, pronom, déterminant, substantif. Pour "tous" et "toute(s)", le choix se réduit à trois en l'absence du substantif. De plus, "tout" présente le cas unique d'un "mot invariable" qui peut s'accorder…

• adverbe :

- dans des locutions toutes faites mais interprétée en plusieurs formes : "tout à fait", "tout de même", "tout en bas"…

- devant des adjectifs ("autre", "jeune", "entier"…). Dans ce cas il sera parfois précédé d'un déterminant du groupe 1 ("un tout jeune homme"…) mais il peut aussi arriver seul : "tout seul" ;

- dans un groupe verbal : "il veut tout faire" ;

- devant un pronom personnel : "tout lui"…

• déterminant (adjectif indéfini) :

- quand il est suivi d'un substantif ou d'un adjectif antéposé de mêmes genre et nombre en particulier dans de nombreuses expressions toutes faites : "en tout cas", "en toute chose", "tout compte fait", "de toute façon", "en tout genre", "en tout lieu", "à tout moment" ;

- comme tous les adjectifs indéfinis, il peut être employé avec un déterminant du groupe 1 ("tout le monde" "toute cette affaire"…), dans un groupe nominal comportant un substantif accordé en genre et en nombre ;

• pronom : tout (invariable), "tous", "toutes" :

- quand il est précédé d'une préposition (à, pour, de, dans…) et n'est pas accompagné d'un déterminant, substantif, adjectif… : "de tout", "en tout", "pour tout", etc). On retiendra que, pour certains grammairiens, on a ici un substantif comme dans les expressions citées ci-dessus ;

- quand il est suivi : d'un verbe fléchi (attention au réfléchi), d'une ponctuation, d'un pronom ("tout ce que…", "toutes celles qui…") accordé en genre et en nombre, d'une préposition (de, chez, dans) ;

- précédé d'un pronom personnel d'une des personnes du pluriel ("nous tous", "eux tous"…") ;

- quand il suit un verbe et n'est pas compris dans un groupe nominal : "il veut tout", "il les prend tous"…

- dans des expressions comme : "c'est tout", "capable de tout", "bonne à tout faire"….

• le substantif masculin est toujours au singulier et accompagné d'un déterminant du groupe 1 et non suivi d'un substantif : (un, le, son…). Se rencontre souvent dans des expressions comme : "jouer le tout pour le tout", "du tout au tout", "pas du tout", etc.

7.2. LA CONJONCTION

La conjonction comme la préposition unit des groupes de mots entre eux. Mais, alors que la préposition relie généralement des mots de catégories différentes entre eux, la conjonction unit des propositions ou des unités de même nature. On s'attend donc à trouver derrière cette dernière, soit un mot de même nature que celui placé avant la conjonction, soit un groupe de mots contenant un verbe.

7.21. Nature de la conjonction

7.211. La classification des conjonctions

On distingue deux types de conjonctions

• conjonction de coordination peuvent avoir plusieurs fonctions :

- liaison entre plusieurs élément (et) ;

- l'opposition entre ces éléments (mais, pourtant,…)

- l'alternative ou la négation (*soit*, ni, ou…)

- la conséquence (donc)

- la conclusion, la conséquence ultime (ainsi, enfin…)

• la subordination : (comme, quand, *que*…) ouvre généralement une proposition plus ou moins complète.

7.212. L'analyse des locutions conjonctives

L'association d'un mot quelconque avec certaines conjonctions peut donner une locution conjonctive. La conjonction "que" est la plus productive :

- rentrer *avant* la nuit est une préposition ;
- rentrer *avant qu*'il fasse nuit donne une locution conjonctive ;

Plusieurs locutions conjonctives sont considérées comme une forme unique (cf ci-dessus la discussion sur les formes § 2.23) : parce que, c'est-à-dire, tandis que…

En revanche, les règles énoncées dans la première partie de cette note conduisent à voir deux vocables dans la plupart des locutions conjonctives : ainsi que, afin que, afin de, alors que, après que, autant que…

7.213. Les principales conjonctions

Dans la liste des principales conjonctions, nous plaçons en italiques les homographies et les locutions que nous analysons en plusieurs mots :

attendu que, *afin que*, *après que*, *avant que*, avec, *car*, cependant, c'est-à-dire, comme, *contre*, *depuis que*, donc,*entre*, et, *excepté*, lorsque, mais, ni, *or*, ou, *outre*, parce que, *pendant que*, *pourvu que*, puisque, quant, *que*, quoique, *sauf*, si, sinon, *tandis que*, *vers*, *vu que*,…

7.22. Les conjonctions homographes

Des listes d'Engwall on retient que certaines homographies paraissent hypothétiques : contre (verbe), durant (verbe), sauf (adjectif) ne sont pas attestés dans l'index d'Engwall portant sur *Le vocabulaire du roman français des années 60* ni dans nos dépouillements. En revanche, les cas suivant sont à noter :

- attesté une fois : outre (substantif),
- entre : verbe 31 fois dans le roman français, 9 fois dans le corpus Mitterrand,
- or : substantif 49 fois dans le roman français, 6 fois dans le corpus Mitterrand,
- car : substantif 18 fois dans le roman français, non employé dans le corpus Mitterrand,
- soit : chez Engwall : 136 flexions du verbe "être", pas de conjonction ; dans le corpus Mitterrand : 47 conjonctions et 311 verbes "être". Chez Juilland, on a 66 conjonctions et 155 verbes "être". Il est donc probable qu'il a une erreur dans le dépouillement d'Engwall…
- que : pronom 1662 fois dans le roman français (contre 7249 conjonctions) et, chez F. Mitterrand 5547 conjonctions contre 2326 pronoms.

7.221. L'homographie entre le verbe et la conjonction

• L' homographie avec un verbe actif (contre, entre) se résout de la manière suivante :

- le verbe est attesté derrière un pronom, dans une négation, suivi d'un adverbe ou d'une préposition ("il entre dans…") ;
- la conjonction est attestée derrière un verbe ("choisir entre"…) ;
- la conjonction est attestée derrière un groupe nominal au pluriel.

• Le cas de "soit" (adverbe ou conjonction ou verbe "être" à la troisième personne du singulier du subjonctif présent) :

l'adverbe est toujours suivi d'un point d'exclamation. Suivant le principe énoncé au début de ce chapitre, on n'analyse pas l'homographie entre l'adverbe et la conjonction. L'homographie entre la conjonction et le verbe est dénouée en suivant la procédure suivante :

- en tête de phrase, la conjonction est toujours attestée ;
- "soit" est un verbe quand il est précédé ou suivi d'un pronom de la troisième personne du singulier ou inclus dans une construction négative ;

- le verbe est attesté quand il est précédé en amont de la phrase par une conjonction "que" et qu'aucun autre "soit" n'est présent dans la phrase ;

- si les tests précédents ont échoué et que l'on trouve un autre "soit" dans la phrase, on considère qu'il s'agit de deux conjonctions ;

- en cas d'échec, l'opérateur est interrogé.

7.222. Le cas de "que"

Le mot "que" est, avec "le", la source d'homographie la plus fréquente puisqu'il peut être pronom et conjonction. C'est de loin le cas le plus difficile à résoudre. Il est traité par un véritable sous-programme qui repose sur les caractéristiques suivantes :

• La conjonction "que" :

- placée après un verbe, elle introduit une complétive. Nous avons indiqué plus haut que la codification des verbes comporte l'indication d'une possible construction avec "que" (annoncer, dire, déclarer, penser, croire, falloir, vouloir, savoir, faire, prouver, affirmer, démontrer, douter, nier, refuser, prétendre…). Attention : dans certaines constructions, la conjonction peut être assez éloignée du verbe auquel elle se rattache, tout particulièrement dans les phrases négatives. Par exemple : "ne vouloir à aucun prix que…" ;

- devant un verbe fléchi ou, plus rarement à l'infinitif ("Il ne sait que faire", "il ne sait que critiquer") ;

- dans des locutions conjonctives et dans les formules d'insistance (voilà que…, c'est que… Ce sont des locutions conjonctives que la norme de dépouillement nous interdit de considérer comme des locutions figées. Ces locutions sont normalement insécables. Cette caractéristique n'est pas absolue. Par exemple : "à condition que" : "à condition expresse que…!. Voici quelques locutions de ce genre : "à mesure que", "attendu que", "de façon que", "de sorte que", "avant que", "dès que", "pendant que", "après que", "pis que", "pourvu que", "tandis que". Le programme peut reconnaître ces locutions mais il ne remonte pas au-delà de (n-1), le caractère conjonctif est considéré comme acquis dès qu'il n'y a pas, en (n-1) d'homographie impliquant une forme fléchie d'un verbe (par ex : "la mesure que…" et "à mesure que…") ;

- dans des locutions sécables. Par exemple, les formules de comparaison : "autant que", "plus que", "moins que", "plutôt que", "mieux que", "autre que", "même que"… Ici l'analyse est moins productive puisque plusieurs formes peuvent venir s'intercaler dans la formule. On ne pourra conclure que lorsque ne s'intercalent pas entre les deux termes de la formule des groupes nominaux ;

- en corrélation avec *ne* pour marquer la restriction (également "*ni*"). Comme précédemment, la négation peut se trouver très loin de "que" et les mots présents dans l'intervalle peuvent interdire toute conclusion…

- après un adjectif ou un nom dans une construction impersonnelle : sujet + verbe + {adjectif ou nom en emploi adverbial} + {que ou de} + {infinitif ou subordonnée}. Par exemple : "il est clair que", "il paraît évident que". La langue semble imposer que "de" soit suivi de l'infinitif et "que" d'une proposition subordonnée. Substantifs ou adjectifs susceptibles d'être construits ainsi avec "que" : clair, vrai, certain, acquis, évident, indéniable, possible, probable, étonnant, exceptionnel, rare, nécessaire, bon, bien, dommage…

• L'adverbe "que" est confondu avec la conjonction. On le rencontre :

- dans un emploi exclamatif : "Que si !"; "Que non !" ;

- en substitut d'un mot-outil ("pour…que", "afin que", "quelque… et que"). Dans ce dernier cas, la préposition ou la conjonction sont des indices mais des indices seulement car un pronom peut aussi se trouver répété grâce à la conjonction ("La fleur que vous m'avez offerte et que vous m'avez reprise…") ;

- les formules interrogatives sont également douteuses. Dans "Que dites-vous ?" la forme "que" est parfois analysée comme conjonction alors que dans "Qu'en dites-vous ?" elle serait

pronom. En fait il est toujours possible de remplacer ces formes par le pronom interrogatif "quoi" ce qui prouve bien leur caractère pronominal. C'est pourquoi nous regroupons ces formes interrogatives sous le pronom.

• Le pronom relatif "que" :

- il se place normalement après le nom ou le groupe nominal auquel il est "relatif". En règle générale, le pronom personnel ne peut que rarement être identifié en tant que tel : comme la conjonction, il peut se trouver entre noms et verbes. On examinera donc cette hypothèse seulement lorsqu'on se trouvera en dehors des cas où l'on peut trancher en faveur de la conjonction ;

- du fait de la règle posée ci-dessus, certains emplois du pronom relatif peuvent être identifiées à la ponctuation : "Que vouliez-vous qu'il fit ?" ; "Que faire ?" ; "Qu'en dîtes-vous ?" ;

- "que" est également pronom dans : "ce que", "c'est que", "c'est ce que" et, dans : "Qu'est-ce que… ?", les deux "que" sont analysés comme des pronoms.

7.223. L'homographie entre la conjonction et le substantif

Cette homographie concerne deux cas assez semblables (car et or) :

- le substantif est attesté lorsqu'il est précédé d'un déterminant, d'un adjectif susceptible d'être antéposé (au masculin singulier) ou des préposition "de" ou "en" ;

- la conjonction est attestée lorsqu'elle est précédée d'un verbe, d'un adverbe ou d'un signe de ponctuation.

7.224. L'homographie entre l'adverbe, la conjonction et la préposition

• Règle générale : on n'analyse pas les homographies internes à ce groupe.

Tout au moins si l'on admet le postulat ci-dessous selon lequel il est impossible d'analyser les homographies au sein de l'ensemble {adverbe-conjonction- préposition} qui concerne un grand nombre de formes au premier rang desquelles "si" (préposition et adverbe : cf ci-dessous)

• "Si" mérite cependant un traitement à part car - par la forme élidée "s'" - existe aussi une homographie avec le pronom réfléchi "se". Enfin, il y a également une homographie avec le substantif (note de musique). Dès lors une analyse complète s'impose :

- "s'" est une conjonction quand elle est immédiatement suivie de "il" ou "ils" ;

- "s'" est un pronom réfléchi quand elle est précédée d'un pronom de la troisième personne ;

- "s'" est un pronom réfléchi quand elle est suivie d'un verbe à la troisième personne ou d'un pronom "en" ou "y" ;

- "si" est une conjonction quand elle introduit une proposition de type sujet+verbe ;

- "si" est un adverbe quand il est placé devant un adjectif ou un autre adverbe ;

- "si" est un adverbe quand il est suivi d'un point ou d'un point d'exclamation.

En cas d'interrogation de l'opérateur, l'adverbe est reconnu par le fait qu'on peut le remplacer par "ainsi" ou par "oui". La conjonction par "puisque" ou "supposons que"…

7.3. LES PREPOSITIONS

La préposition est un mot invariable qui introduit un complément en marquant le rapport qui unit le complément au mot complété. Ce rapport de lieu (dans) ; rapport de temps (depuis) ; d'appartenance (de) ; moyen (à)…

7.31. Nature de la préposition

Du fait des rapports extrêmement divers qu'elles sont censées exprimer, les prépositions offrent une large gamme de possibilités et de cas particuliers.

7.311. Les principales prépositions

Dans la liste ci-dessous les homographies sont signalées en italiques :

- principales prépositions : à, après, au-delà, *avant*, avec, chez, *contre*, dans, de, depuis, *derrière*, dès, *en*, en-deça, *envers*, hors, jusque(s), malgré, *outre*, par, parmi, *plein*, pour, près, *proche*, sans, *sauf*, selon, *sous*, sur, *vers*, via, voici, voilà…
- les anciens participes devenus prépositions (les homographies sont placées en italiques) : *attendu*, *devant*,*durant*, *entendu*, *excepté*, hormis, moyennant, nonobstant, *passé*, *pendant*, *pourvu*,*suivant*, *supposé*, *touchant*, *vu*… En revanche, à l'instar de Lyne[1], nous ne conservons pas "concernant" ;
- les locutions prépositives. Pour ces locutions, les règles énoncées dans notre première partie nous amènent à analyser en plusieurs mots celles que nous plaçons en italiques : *à cause de*, *à côté de*, *à défaut de*, *afin de*, *auprès de*, *d'après*, *jusqu'à*, *en dehors de*, *loin de*…
- emploi spéciaux ou vieillis : delà, devers, environ, ès, ex, fors, in…

7.312. Les règles de composition des prépositions

- avec des groupes verbaux : les prépositions "de" et "à" seront suivies d'un infinitif et jamais d'une forme fléchie ("demander à voir", "la volonté de faire"). Cette règle est d'une grande utilité pour la résolution des homographies du verbe. Par exemple, dans "il n'a de cesse" nous pouvons voir en "cesse" un substantif puisqu'il suit "de" et ne peut donc pas être une forme fléchie du verbe "cesser"…
- les autres prépositions introduisent des groupes nominaux : *les formes homographes du verbe précédées de prépositions autres que "de" ou "à" ne sont pas des verbes*. Cette règle simple est d'une grande utilité pour l'analyse des formes ambiguës, notamment les couples "homographe déterminant-pronom"+ "homographe substantif-verbe". Par exemple : "après *l'annonce*", "dans *la lutte*" sont nécessairement des déterminants et des substantifs. Cette règle simple permet de résoudre près d'un tiers des couples homographes (pronom+verbe) ou (déterminant+substantif).

7.32. L'analyse des prépositions

7.321. La frontière entre la préposition et l'adverbe

"Il existe, entre la préposition et l'adverbe, des rapports fort étroits : plus d'un adverbe (avec, après, avant) a pu dès l'origine jouer un rôle de préposition. Dans le français moderne, la distinction s'est nettement établie, mais non sans que le passage reste parfois possible entre les deux catégories : dans la langue familière, les prépositions "après, avant, contre, depuis,

---

1. Antony Lyne, *The Vocabulary of French Business Correspondence*, Paris-Genève, 1985, p 91.

derrière, devant, entre, hors, outre, parmi, proche, sans, selon" s'emploient couramment comme adverbes (surtout avec)" (Grévisse). On en tire qu'il n'est pas souhaitable de les départager comme nous l'indiquions au début de ce chapitre.

7.322. Les homographies entre la préposition et le verbe

• Les participes passés des verbes attendre et voir ("vu", "attendu"). Ces deux cas doivent être traités à part car, les verbes comme les préposition peuvent être accompagnées de "que". On recherche, à gauche de la forme, un auxiliaire "être" ou "avoir" dont elle peut être séparée par un pronom, une négation, un adverbe - ou par une locution adverbiale - pouvant s'intercaler entre l'auxiliaire et le participe passé.

• durant, excepté, pourvu. Ces verbes ne pouvant se construire avec "que", la présence d'une conjonction "que" derrière la forme indique donc avec certitude une préposition. En cas d'échec, le test de l'auxiliaire est appliqué aux homographes du participe passé comme décrit dans le paragraphe ci-dessus, et le test du gérondif aux homographes du participe présent.

7.323. Les homographies entre la préposition, l'adjectif et le substantif

Ce cas concernent les prépositions : avant, contre, derrière, envers, outre, plein, proche, sauf, sous, vers. Leur éventuelle homographie avec les adverbes correspondants n'est pas analysée. Les règles de reconnaissance sont les suivantes :

- on est en présence d'un substantif quand la forme est précédée d'un déterminant, d'un adjectif susceptible d'être antéposé ou suivi d'un adjectif et si ces formes sont accordées en genre et en nombre ;

- on est en présence d'un adjectif (plein, proche, sauf) quand la forme est précédée d'un substantif ou d'un déterminant de même genre et de même nombre. Cette règle amène à classer dans les adjectifs quelques formes ambiguës : "un ami *proche* de lui" ;

- pour les adjectifs, on est en présence d'un attribut derrière le verbe être - "le moment est proche" - et d'une préposition derrière les autres verbes : se sentir proche de…

- on est en présence d'une préposition quand celle-ci est placée en tête d'un groupe nominal ou quand elle suit immédiatement un verbe (autre que les verbes d'état)…

- certaines prépositions autres que "en" peuvent être associées à elle pour former une locution (adverbiale donc non départagée de la préposition) : "en avant", "en arrière", "en outre", de proche en proche…

7.324. Le cas de "en"

"En" peut être un pronom ou une préposition.

• "En" est préposition dans les positions suivantes :

- devant un nom sans déterminant ou avec un déterminant autre que l'article défini ;

- devant un indéfini, un adjectif neutre, un adverbe, pour former des locutions adverbiales : "en tout", "en rien", "en général", "en particulier", "en gros", "en vain", "en dehors", "en bas"…

- devant un verbe au participe présent forme le gérondif ("en marchant").

• "En" est un pronom parfois nommé "adverbial" dans les constructions suivantes :

- complément de nom servant d'appui à des quantitatifs et des indéfinis : "donnez m'en", "mettez m'en", "en voilà un", "tu en aimes un autre"…

- complément de verbe : complément de lieu (j'en viens)…, d'objet (j'en ai) ;

- dans un groupe verbal entre le réfléchi et le verbe : "on s'en va", "je m'en tiens là"…

- dans le groupe verbal devant l'auxiliaire : "on en a…"

7.325. Le cas de "de"

"De" peut être un article partitif (déterminant) ou une préposition.

C'est, avec "que", "le" et "tout", l'homographie la plus embarrassante. D'une part, il s'agit de la forme dont la fréquence est la plus élevée. Dès lors, le traitement ne doit pas trop reposer sur l'opérateur… Or, d'autre part, la lecture des grammaires et des principaux manuels laisse perplexe tant cette question donne lieu à des interprétations divergentes et à des distinguos peu opératoires. Nous avons cependant élaboré un cadre qui permet de résoudre au moins partiellement la question :

• Les formes "des", "du".

- on peut les envisager comme des partitifs quelle que soit leur construction. Ainsi retrouverait-on le caractère invariable des prépositions. Mais de nombreuses prépositions passeraient dans le partitif ("ne pas manger du tout")…

- on peut les décomposer en "de les" et "de le" que l'on analyse comme "préposition et article". Cette solution est brutale mais elle permet de considérer que "de" est préposition derrière tous les verbes admettant cette construction. Par exemple, dans "il part de loin" et "il part des hypothèses…", "de" sera toujours considéré comme une préposition et l'article sera rétabli. C'est la solution que nous avons retenue car elle est conforme à la norme Muller. Elle présente toutefois un inconvénient important : la taille du texte lemmatisé sera légèrement allongée par rapport au texte à la norme Saint-Cloud.

Cette opération de "décontraction" effectuée, l'homographie se résout à "de" et "d'".

• La préposition est examinée en premier :

- elle suit le plus souvent un verbe admettant la construction en "de" (ces verbes portent un code particulier). C'est le seul moyen de départager "il pose de nombreux problèmes" (partitif) et "il parle de problèmes nombreux " (préposition) ;

- elle précède un verbe à l'infinitif (attention aux multiples homographies entre le verbe à l'infinitif et les substantifs : "il a trop de pouvoir") ;

- elle est couramment placée entre deux noms ("taille des arbres") ou entre un adjectif et un nom ("blanc de peur")… Ici on peut appliquer la règle pas plus d'un déterminant de même catégorie dans un groupe nominal…

- elle suit un adverbe ou une autre préposition ("plus de peur que de mal", "moins de bruit", "trop de pouvoir"….

- elle précède un infinitif ou les pronoms "en", "y" "ce" ("d'y aller", "d'en dire", "de ce que") ou un nom propre ("le discours de Mitterrand") ;

- elle est employée dans des constructions impersonnelles : "il est {substantif, adjectif, participe présent ou passé} de faire" ("il est courant de faire" ; "il est mal de dire"…

• Le partitif se trouve :

- devant un substantif au singulier ou un adjectif antéposé également au singulier (ce test doit donc être effectué après avoir examiné l'hypothèse de la préposition) ;

- il suit les règles générales des déterminants. Par exemple, il ne peut pas y avoir un autre déterminant de la classe 1 dans le groupe nominal. L'ensemble du groupe nominal doit être au singulier ;

Cependant, la règle de l'accord ne peut être généralisée car l'usage permet d'utiliser le partitif "de" au singulier devant certains adjectifs antéposés au pluriel ("Ce sont de grosses bêtises", "de nombreux cadeaux", "de multiples personnes"…) ;

- en début d'une phrase ou après un signe de ponctuation quelconque, on considérera que nous sommes en présence du partitif.

Malgré ces conventions assez complètes, nous avons dû nous résoudre à ne pas disjoindre le partitif "de" et la préposition homographe. Des tests nous ont montré que le taux de réussite n'était pas fameux et que plusieurs erreurs se glissaient dans le dépouillement (en particulier du

fait des verbes pouvant se construire avec la préposition "de"). En effet, comment permettre à la machine de repérer "les choses qui ne se comptent pas et les noms abstraits" que le partitif est censé introduire ?

Personne, jusqu'à maintenant n'a résolu ce problème. Généralement les index font passer "de, d', du, des" dans les prépositions sans autre forme de procès. Dans ces conditions, la décomposition des formes contractes préconisée par C. Muller apporte une amélioration incontestable sans résoudre le cas du singulier.

7.325. Le cas de "au"

Le problème des formes contractes se retrouve avec "au" et "aux" dont la fréquence est très élevée. Nous avons adopté la solution préconisée par C. Muller. Elle s'inspire de celle mise en oeuvre pour "du" et "des" et présentée au paragraphe précédent. Soit, le lemme étant placé entre parenthèses et en italiques :

- "au" est analysé comme : "à" (*à, préposition*) "le" (*le, déterminant masculin singulier*) ;
- "aux" : "à" (*à préposition*) "les" (*le déterminant masculin ou féminin pluriel*)

Cette opération de "décontraction" a le mérite de faire apparaître l'ambiguïté de la préposition "à" qui joue, dans la langue française, un rôle proche de celui assuré par "de" (voir à ce sujet, l'article que lui consacre N. Ruwet dans sa *Grammaire des injures*)

Plus généralement cette discussion illustre la difficulté que l'on éprouve à tracer des frontières stables et définitives entre les mots invariables ou entre ceux-ci et certains déterminants. Finalement l'important ne réside pas dans la rigueur de ces partitions que dans la clarté et la simplicité des conventions qui les fondent. La stabilité de la norme de dépouillement est à ce prix.

CONCLUSION GENERALE

Pour l'essentiel, nos études sur le vocabulaire politique reposent sur des données obtenues grâce à un traitement des mots qui les maintient à leur place dans les textes, sans faire subir à ces derniers d'altération notable. En 1984, au début de l'expérience, nous avions fixé trois objectifs aux programmes informatiques que nous commencions à réaliser. D'une part, nous souhaitions qu'ils codifient automatiquement la quasi-totalité des mots et, en particulier, qu'ils résolvent la grande majorité des difficultés homographiques. Deuxièmement, nous voulions parvenir à un "zéro faute" : l'ordinateur choisit à coup sûr ou interroge l'opérateur. Troisièmement, nous souhaitions que les deux phases (Saint-Cloud et Muller) représentent pour l'opérateur une charge de travail aussi faible que possible.

A la fin des traitements sur le vocabulaire de F. Mitterrand, plus de neuf homographies sur dix étaient dénouées par l'ordinateur. Etant donné que, dans un texte quelconque, il semble y avoir, en moyenne, à peu près 32% des mots qui présentent une difficulté de codification - sans compter le problème du partitif "de" -, cela signifie que des programmes capables de reconnaître 98% des mots d'un corpus sont parfaitement concevables.

En revanche, les deux autres objectifs n'ont pas été atteints.

Jusqu'à la fin de cette expérience - malgré la complexification des programmes -, la lemmatisation contenait quelques scories dues à des constructions de phrases ou à des tournures imprévues, à des mots inconnus, à des homographies que nous n'avions pas aperçues avant qu'elles ne se traduisent par des aberrations. Pour venir à bout de ces quelques erreurs, il faudrait augmenter considérablement le volume des tables et le nombre des règles que doivent gérer les programmes. Une telle augmentation, pose un double problème. D'une part, ces règles devraient être classées en niveaux hiérarchiques et il faudrait examiner attentivement leurs incompatibilités éventuelles. D'autre part, ce saut nécessite des moyens intellectuels et matériels importants, notamment pour la programmation et l'expérimentation sur des corpus plus étendus. Le problème du rapport entre les coûts et les avantages se pose inévitablement et la réponse à peu de chance d'être favorable à une poursuite de ces recherches !

A l'heure actuelle, la "chasse à l'erreur" de lemmatisation oblige à relire les fichiers après leur passage en machine, d'où une augmentation importante du temps de travail. Sur ce point, nous ajouterons le problème de l'orthographe française : jusqu'à la fin de la recherche sur F. Mitterrand, des fautes nouvelles ont été découvertes dans le texte original. Certaines étaient détectées grâce aux programmes de lemmatisation qui se heurtaient à des formes inconnues ou à des incohérences. D'autres conduisaient à des codifications erronées. Toutes ont coûté de nombreuses heures de relecture fastidieuse.

Au total, nous estimons que, pour l'instant, le temps de travail requis par l'application des normes "Saint Cloud" et "Muller" correspond au moins à l'équivalent d'une nouvelle saisie du texte.

Enfin, la norme de dépouillement pose encore certains problèmes. Les frontières entre les mots ne sont pas toujours claires. Le comportement de certains d'entre eux est pour le moins équivoque : ils nécessitent des analyses fouillées et un traitement "sur mesure". De plus, les catégories de la grammaire traditionnelle ne sont pas totalement adaptées à une analyse syntaxique rigoureuse. Il faudrait descendre plus dans le détail que nous ne l'avons fait et faire éclater certaines catégories (les adjectifs indéfinis, les pronoms, les adverbes…) Or ces catégories résistent mal aux contraintes de l'informatique qui ne tolère pas le flou ni les

frontières malléables. Toute complexification de la grille risque de multiplier les homographies et de paralyser les programmes, d'où une inflation prévisible des appels à l'opérateur.

Pour surmonter ces faiblesses, une plongée profonde dans la langue française paraît indispensable. Elle devrait s'accompagner d'une discussion entre les lexicographes pour que soient définies, par consensus, les grandes lignes des normes de dépouillement et de lemmatisation des textes français contemporains.

Après avoir dépouillé, dans le texte, près d'un demi-million de mots, nous sommes donc obligé de confesser un demi-échec. Cependant, nous avons acquis la certitude que cette voie est fructueuse et que l'expérience mérite d'être conduite sur des corpus plus étendus que celui auquel nous nous sommes limités du fait de la faiblesse de nos moyens. Nous espérons que ce modeste document pourra y aider.

ANNEXE 1
TABLE DES LOCUTIONS ET MOTS COMPOSES

La liste ci-dessous reproduit la table des locutions utilisée pour le traitement des discours de F. Mitterrand et de C. de Gaulle. Etablie a priori, elle ne prétend pas être exhaustive.

A à bras-le-corps, a cappella, a contrario, à croupetons, à Dieu vat, a fortiori, a giorno, à hue et à dia, à jeun, à l'avenant, à l'encontre, à l'envi, à l'instar, a pari, a posteriori, a priori, a quia, à rebrousse-poil, à reculons, à tâtons, à tire-d'aile, à tire-larigot, à tue-tête, à vau-l'eau, à-côté(s), à-coup(s), à-propos, à-valoir, ab intestat, ab ovo, abat-jour, abat-son, abat-vent, abat-voix, accroche-coeur(s), accroche-plat, ad hoc, ad hominem, adjudant(s)-chef(s), ad libitum, ad litem, ad patres, ad valorem, ad vitam aeternam, agro-alimentaire(s), aide(s)-comptable(s), aide(s)-soignant(e,es,s), aigre(s)-doux, aigue(s)-marine, aller-retour, allocation(s)-logement, …-chômage, …- maladie, alma mater, alter ego, américano-soviétique, amour-propre, amuse-gueule, anarcho-syndicalisme (te,s), anglo-arabe, anglo-normand(e,ex,s), anglo-saxon(ne,nes,s), année(s)-lumière, **anti** : anti-atomique(s), anti-économique(s), anti-inflammatoire(s), anti-inflation, anti-monte-lait, anti-sous-marin, anti-tout, apprenti(s)-médecin(s), apprenti(s)-sorcier(s), appui-bras, appui-main, appui-tête, après-demain, après-dîner, après-guerre, après-midi, après-ski, arabo-persique(s), arc(s)-boutant, arc(s)-en-ciel, arrache-clou, arrache-pied, **arrière** : arrière-ban, arrière-bec, arrière-bouche, arrière-boutique(s), arrière-cerveau, arrière-choeur, arrière-corps, arrière-cour(s), arrière-cuisine(s), arrière-fleur, arrière-fond(s), arrière-garde(s), arrière-gorge, arrière-goût, arrière-grand(s)-mère(s), arrière-grand(s)-oncle(s), arrière-grand(s)-père(s), arrière-grand(s)-tante(s), arrière-main, arrière-neveu(x), arrière-pays, arrière-pensée(s), arrière-petit(s)-fils, arrière-petite(s)-fille(s), arrière-plan(s), arrière-port(s), arrière-saison, arrière-salle(s), arrière-train(s), assurance(s)-automobile, assurance(s)-chômage, assurance(s)-invalidité, assurance(s)-maladie, assurance(s)-vie, assurance(s)-vieillesse, attrape-mouche(s), attrape-nigaud(s), au-dedans, au-dehors, au-delà, au demeurant, au-dessous, au-dessus, au-devant, au fur et à mesure, aujourd'hui, auto-accusation, auto-allumage, auto-école(s), auto-imposition, auto-induction, auto-intoxication, auto-intoxiquer, auto-stop, auto-stoppeu(r,se,s), auto-suffisant(e,s,ce), **avant** : avant-bassin, avant-bras, avant-centre, avant-corps, avant-coureur(s), avant-courrier, avant-dernier(e,es,s), avant-garde(s), avant-goût, avant-guerre, avant-hier, avant-main, avant-mont, avant-port(s), avant-poste(s), avant-première, avant-projet(s), avant-propos, avant-scène(s), avant-toit, avant-train(s), avant-veille, avocat(s)-conseil(s),

B b-a ba, bain(s)-marie, balai(s)-brosse, ballon(s)-sonde, bar-bibliothèque, bar(s)-tabac(s), bas-allemand(e,es,s), bas-côté(s), bas-fond(s), bas-relief(s), bas-ventre(s), basse(s)-cour(s), basse-fosse, bat-flanc, **Beau** (bel) : beau(x)-fils, beau(x)-frère(s), beau(x)-papa, beau(x)-père(s), beaux-arts, bec-d'âne, bec-de-cane, bec-de-lièvre, bec-fin, belle(s)-dame(s), belle(s)-de-jour, belle(s)-famille(s), belle(s)-fille(s), belle(s)-mère(s), belle(s)-soeur(s), béni-oui-oui, bernard-l'hermitte, best-seller, **Bien** : bien-aimé(e,es,s), bien-dire, bien-être, bien-fondé, bien-fonds, bien-pensant(e,es,s), bien-portant(e,es,s), bio-chimie, bio-chimique(s), bio-technologie(s), black-jack, black-out, blanc-bec, blanc-seing, blancs-manteaux, bleu-gris, bleu-noir, bloc-moteur, bloc-notes, bloc-système, blue-jean(s), bon(s)-papa(s), bonne(s)-maman(s), borne-fontaine, bouc(s) émissaire(s), bouche-à-bouche, bouche-trou(s), boui-boui, bout(s)-dehors, boute-en-train, bouton-d'or, bouts-rimés, bow-window, boy-scout(s), bracelet(s)-montre, brain-trust, branle-bas, bric à brac, brise-bise, brise-fer, brise-glaces, brise-lames, brise-mottes, brise-tout, brise-vent, brûle-gueule, brûle-pourpoint, buisson-ardent, bull-terrier,

C c'est-à-dire, cache-cache, cache-col, cache-flammes, cache-misère, cache-nez, cache-pot, cache-poussière, cache-radiateur, cache-sexe, cache-tampon, café-concert, café-crème, café(s)-

restaurant(s), cahin-caha, camion(s)-citerne(s), cap-hornier(s), carte(s)-lettre(s), carton-pâte, casse-cou, casse-croûte, casse-gueule, casse-noisettes, casse-noix, casse-pattes, casse-pieds, casse-pierres, casse-pipes, casse-tête, casus belli, celle(s)-ci, celle(s)-là, celui-ci, celui-là, cérébro-spinal(e,es,s), cerf(s)-volant(s), cessez-le-feu, ceux-ci, ceux-là, chaise(s)-longue(s), chasse-clou(s), chassé-croisé, chasse-goupille(s), chasse-marée, chasse-mouche(s), chasse-neige(s), chasse-pierre(s), chasse-roue(s), chassés-croisés, château(x)-fort(s), chaud(s)-froid(s), chauffe-eau(x), chauffe-pied(s), chauffe-plat(s), chausse-pied(s), chausse-trappe(s), chauve(s)-souris, chef(s)-d'oeuvre, chef(s)-lieu(x), chef(s)-machiniste, chêne(s)-liège(s), chèque(s)-vacances, cheval(eaux)-vapeur, chevau-légers, chien(s)-loup(s), chou(x)-fleur(s), chou(x)-rave(s), ci-après, ci-dessous, ci-dessus, ci-devant, ci-contre, ci-gît, ci-inclus, ci-joint(e,s), ciné-club(s), ciné-roman(s), cirro-cumulus, cirro-stratus, clair(s)-obscur(s), claire-voie, clin d'oeil, clins d'yeux, cloche-pied, clopin-clopant, coca-cola, cocotte-minute, coffre-fort(s), coin-coin, col-de-cygne, colin-maillard, colin-tampon, commedia dell'arte, commis-voyageur(s), commissaire(s)-priseur(s), compte-gouttes, compte-tours, congé(s)-conversion(s), congé(s)-formation(s), contrat(s)-type(s), **contre** : contre-alizé, contre-allée(s), contre-amiral(aux), contre-analyse(s), contre-appel(s), contre-assurance(s), contre-attaque(s), contre-attaquant(s), contre-avis, contre-chant(s), contre-courant, contre-courbe, contre-dénonciation, contre-digue, contre-écrou(s), contre-empreinte(s), contre-enquête(s), contre-épreuve(s), contre-espionnage, contre-essai(s), contre-expertise(s), contre-extension, contre-feu, contre-fugue, contre-haut, contre-indication(s), contre-jour, contre-lettre(s), contre-manifestant(s), contre-manifestation(s), contre-marche(s), contre-mesure(s), contre-mine(s), contre-offensive(s), contre-ordre(s), contre-performance(s), contre-pied(s), contre-placage(s), contre-plaqué, contre-plongée(s), contre-poil(s), contre-pointe, contre-porte(s), contre-projet(s), contre-proposition(s), contre-réforme(s), contre-révolution(s), contre-révolutionnaire(s), contre-sens, contre-terrorisme, contre-timbre(s), contre-torpilleur(s), contre-vagues, contre-valeur(s), contre-vérité(s), contre-visite(s), contre-voie(s), coq-à-l'âne, cordon-bleu, corps-à-corps, corps-mort(s), couci-couça, coup-de-poing, coupe-choux, coupe-cigare(s), coupe-circuit(s), coupe-coupe, coupe-feu(x), coupe-file(s), coupe-gorge(s), coupe-jarret(s), coupe-légume(s), coupe-ongle(s), coupe-papier(s), coupe-racine(s), coupe-vent(s), court-bouillon, court(s)-circuit(s), couvre-chef(s), couvre-feu(x), couvre-lit(s), couvre-livre(s), couvre-pied(s), cow-boy(s), crayon(s)-feutre(s), crédit(s)-bail, crédit(s)-formation, crève-coeur(s), crève-la-faim, croc(s)-en-jambe, croche-pied(s), croque-mitaine(s), croque-monsieur, croque-mort(s), croque-note(s), cross-country, cul-de-basse-fosse, cul(s)-de-jatte, cul(s)-de-lampe, cul(s)-de-porc, cul(s)-de-poule, cul(s)-de-sac, cul(s)-terreux, cumulo-nimbus, cumulo-stratus, cure-dent(s), cure-ongle(s), cure-oreille(s), cure-pipe(s), curriculum vitae, cuti-réaction,

D d'abord, d'ailleurs, d'antan, d'emblée, dame-pipi, de bric et de broc, de facto, de plano, de profundis, de visu, décret(s)-loi(s), décrochez-moi-ça, déjà-entendu, déjà-vécu, déjà-vu, delirium tremens, **demi** : demi-bas, demi-botte(s), demi-brigade (s), demi-cercle(s), demi-circulaire(s), demi-clef(s), demi-colonne(s), demi-deuil, demi-dieu(x), demi-douzaine(s), demi-droite(s), demi-enfant(s), demi-femme(s), demi-fin(s), demi-finale(s), demi-fine(s), demi-fond, demi-frère(s), demi-gros, demi-heure(s), demi-joue(s), demi-jour(s), demi-journée(s), demi-litre(s), demi-longueur(s), demi-lune, demi-mal, demi-mensonge(s), demi-mesure(s), demi-mondaine(s), demi-monde, demi-mort(e,s), demi-mot(s), demi-nu(e,s), demi-obscurité, demi-pause, demi-pension, demi-pensionnaire(s), demi-place(s), demi-portion(s), demi-produit(s), demi-quart, demi-queue, demi-reliure, demi-ronde, demi-saison(s), demi-sang, demi-sel, demi-semaine, demi-siècle, demi-soeur(s), demi-solde(s), demi-sommeil, demi-soupir(s), demi-sourire(s), demi-tarif(s), demi-teinte(s), demi-tige(s), demi-ton, demi-tour(s), demi-vérité(s), demi-vierge(s), démocrate(s)-chrétien(s), démocratie-chrétienne, dessous-de-plat, dessous-de-table, dessus-de-lit, dessus-de-plat, deus ex machina, deux-mâts, deux-pièces, deux-points, deux-ponts, deux-roues, dies irae, dix-septième, dix-huitième, dix-neuvième, don Quichotte(s),

don-quichottisme, double-bec(s), double(s)-blanc(s), double(s)-crème(s), double(s)-fond, double-jeu, duffel-coat(s),

E eau-de-vie, eau(x)-forte(s), eaux-vannes, ecce homo, électro-aimant(s), électro-encéphalogramme(s), électro-encéphalographie, elle(s)-même(s) en-cas, en-tête(s), entre-deux, entre-deux-guerres, ès lettres, ès qualité, ès sciences, esprit-de-sel, esprit-de-vin, essuie-glace(s), essuie-mains, essuie-pieds, essuie-verres, est-ouest, et cetera, état(s)-major(s), étouffe-chrétien, eux-mêmes, **Ex, extra** : ex abrupto, ex aequo, ex ante, ex cathedra, ex post, ex professo, ex-épouse, ex-époux, ex-femme, ex-mari, ex-ministre(s), expert-conseil, ex-voto, extra-légal(aux, le,les), extra-léger(*e,s*) extra-lucide(s), extra-muros, extra-parlementaire(s), extra-sensible(s), extra-sensoriel(le,les,s), extra-systole, extra-territorial(e), extra-territorialité, extra-territoriaux, extrême droite(s), extrême gauche(s), extrême-onction, extrême-oriental(le, les,aux), extrême-pointe,

F F-1, fac-similé(s), face-à-face face(s)-à-main(s), fair-play, faire-part(s), faire-valoir, fait-tout, faux-bourdon,faux-filet(s), faux-fuyant(s), faux-monnayeur(s), faux-saunier(s), faux-semblant(s), feed-back, feld-maréchal(aux), femme(s)-enfant(s), femme(s)-orchestre, ferry-boat(s), fesse-mathieu, fier-à-bras, fil-à-fil, flanc-garde, fou-fou, fourre-tout, fox-terrier(s), fox-trot, **Franc-** : franc-alleu, franc-bord, franc-bourgeois, franc-jeu, franc(s)-maçon(ne,nes,s), franc-maçonnerie(s), franc-maçonnique(s), franc-or, francs-or, franc-parler, franc-quartier, franc(s)-tireur(s),
Franco- (liste non limitative) : franco-africain(e,es,s), franco-algérien(ne,s), franco-allemand(e,es,s), franco-américain(ne,nes,s), franco-anglais(e,es), franco-arabe(s), franco-belge(s), franco-britannique(s), franco-canadien(ne,s), franco-espagnol(e,es,s), franco-iranien(e,s), franco-irlandais(e,s), franco-israélien(ne,s), franco-italien(ne,nes,s), franco-marocain(e,s), franco-mexicain(ne,nes,s), franco-portugais(e,s), franco-québécois(e,s), franco-russe(s), franco-soviétique(s), franco-suisse(s), fric-frac(s), frou-frou, fume-cigare, fume-cigarette, fusil-mitrailleur,

G gagne-pain, gagne-petit, **garde** : garde-à-vous, garde(s)-barrière, garde-boue, garde(s)-champêtre, garde-chasse, garde(s)-chiourme, garde-corps, garde-côte(s), garde-feu, garde-fort, garde-fou(s), garde-frein, garde(s)-malade, garde-manger, garde-meuble(s), garde-pêche, garde-place, garde-robe(s), garde-voie, garden-party, gas-oil, gastro-entérite(s), gastro-intestinal(e,s,aux), gâte-papier, gâte-sauce, gentleman-farmer, globe-trotter(s), gobe-mouches, gorge-de-pigeon, **grand-** : grand-place, grand-angle, grand-angulaire, grand-chose, grand-croix, grand(s)-duc(s), grand-ducal(e,s,aux), grand(s)-duché(s), grand-fête, grand-guignol, grand-guignolesque, grand(s)-mère(s), grand(s)-messe(s), grand(s)-oncle(s), grand-peine, grand(s)-père(s), grand-peur, grand-route, grand-rue, grand(s)-tante(s), grand(s)-voile(s), grands-parents, gras-double(s), gratte-ciel, gratte-cul(s), gratte-dos, gratte-papier(s), gréco-latin(e,s), gréco-romain(e,s), gri-gri, grill-room, grille-pain, grippe-sou, gris-bleu, gris-vert, gros-bec, gros-grain, grosso modo, guet-apens, gueule-de-loup, guide-âne, gulf-stream, gutta-percha,

H habeas corpus, hache-légumes, hache-paille, hand-ball, haut(s)-commissaire(s), haut-de-chausses, haut-de-forme(s), haut(s)-fond(s), haut(s)-fourneau(x), haut-le-coeur, haut-le-corps, haut-parleur(s), haut-relief(s), héroï-comique(s), hi-han, hic et nunc, hold-up, homme(s)-grenouille(s), homme(s)-orchestre(s), homme(s)-sandwich(es), honoris causa, hors-bord, hors-concours, hors-d'oeuvre, hors-jeu, hors-la-loi, hors-ligne, hors-piste, hors-texte, hôtel(s)-restaurant(s), huit-reflets, hydro-électricité, hydro-électrique(s),

I : idée(s)-force, import-export, in extenso, in extremis, in fine, in pace, in partibus, in petto, in situ, in vitro, in vivo, in-folio(s), in-octavo(s), in-plano, in-quarto(s), inch Allah, indépendance-

association, indo-européen(ne,nes,s), indo-hellénique(s), indo-paskinais(e,es), infra-son(s), ingénieur(s)-conseil(s), intra-muros, intra-utérin(e,s), intuitu personae, ipso facto, israélo-arabe(s).

J, K je-m'en-fichisme, je-m'en-foutisme, je-m'en-foutiste, jean-foutre, jet-set, jet-stream, jiu-jitsu, joli-coeur, jordano-palestinien(ne,nes,s), judéo-allemand(e,s), judéo-chrétien(ne,s), judéo-christianisme, juke-box, jupe(s)-culotte(s), jusqu'au-boutisme, jusqu'au-boutiste(s), jusque-là, juste-milieu, kif-kif, kilogramme(s)-force, kilomètre(s)-heure, knock-out,

L là-bas, là-dedans, là-dessus, là-dessous, là-haut, lacryma-christi, laissé(e,es,s)-pour-compte, laisser-aller, laissez-passer, lance-amarre(s), lance-bombe(s), lance-flamme(s), lance-fusée(s), lance-grenade(s), lance-pierre(s), lance-roquette(s), lance-torpille(s), langue-de-boeuf, langue-de-chat, latino-américain(ne,nes,s), laurier(s)-rose(s), lave-glace(s), lave-linge(s), lave-main(s), lave-vaisselle(s), lèche-cul, lèche-vitrines, lève-nez, libre-échange, libre-échangiste(s), libre-service, lie-de-vin, lieu-dit, lieutenant(s)-colonel(s), lit(s)-cage(s), lit(s)-divan(s), living-room, livret(s) A, lock-out, loi(s)-cadre(s), loi(s)-programme(s), long-courrier(s), longue-vue(s), louis-philipard(e,es,s), loup garou, loup(s)-cervier(s), lui-même,

M m'as-tu-vu(e,s), machine(s)-outil(s), ma-jong, main-forte, maison-mère, maître-autel, maître-à-danser, maître-fromager, maître-nageur, mal-jugé, mal-loti(s), mandat(s)-carte(s), mandat(s)-lettre(s), mange-tout, manu militari, marie(s)-salope(s), marteau(x)-pilon(s), martin(s)-pêcheur(s), marxisme-léninisme, marxiste-léniniste(s), mass média(s), mater dolorosa, médecin(s)-chef(s), médecin(s)-conseil(s), médico-légal(e), méli-mélo, mère-grand, meurt-de-faim, mezza-voce, **mi** : mi-bas, mi-bas, mi-carême, mi-certain(e,s), mi-certitude(s), mi-chemin, mi-clos, mi-close, mi-closes, mi-corps, mi-côte, mi-envieux(se,ses), mi-gai(e,es,s) mi-ironique(s), mi-moi(s, nom du mois), mini-informatique, mini-ordinateur(s), mi-parti, mi-partie(s), mi-semaine(s), mi-sérieux(se,ses), mi-temps, mi-triste, mi-voix ; **micro** : micro-analyse, micro-économie, micro-informaticien, micro-informatique, micro-onde(s),(s), micro-ordinateur(s), micro-organisme(s), micro-processeur(s), mieux-être, milk-bar(s), mille-feuilles, mille-pattes, mille-pertuis, mini-informatique, minus habens, missi dominici, modus vivendi, moi-même, moins-perçu, moins-value(s), moitié-moitié, mont-de-piété, monte-charge(s), monte-en-l'air, monte-plat(s), monte-sac(s), montre(s)-bracelet(s), mort-aux-rats, mort(s)-gage(s), mort-né(e,es,s), morte-eau, morte(s)-saison(s), mot-à-mot, mot-clef, moto-cross, moyen-âge, moyen-âgeu(se,ses,x), moyen-courrier(s), mule-jenny,

N national-socialisme, national-socialiste, nationaux-socialistes, ne varietur, negro-spiritual(s), **néo** : néo-calédonien(ne,s), néo-capitalisme, néo-capitaliste(s), néo-classicisme, néo-classique(s), néo-colonial(le,les, aux), néo-colonialisme(s), néo-colonialiste(s), néo-criticisme, néo-darwinisme, néo-gothique(s), néo-libéral(le,les,aux), néo-libéralisme, néo-mauresque, néo-mauresques, néo-platonicien(ne,s), néo-platonisme, néo-positivisme, néo-réalisme, néo-réaliste(s), néo-socialiste(s), néo-thomisme, néo-zélandais(e,s), nez-cassé, no man's land, **non** : non-activité, non-agression, non-aligné(e,s), non-alignement, non-assistance, non-belligérance, non-belligérant(s), non-combattant(te,s), non-comparution, non-conciliation, non-conformisme, non-conformiste(s), non-conformité, non-contradiction, non-contradictoire, non-engagé(e,s), non-être, non-euclidien, non-euclidienne, non-exécution, non-existence, non-figuratif(ve,s), non-ingérence, non-intervention, non-interventionnisme, non-interventionniste(s), non-jouissance, non-lieu, non-moi, non-pareil, non-recevoir, non-satisfaction, non-sens, non-stop, non-usage, non-valeur, non-violence, non-violent(e,s), nord-africain(e,s), nord-américain(e,s), nord-coréen(ne,s), nord-est, nord-ouest, nord-sud, nord-vietnamien(ne,s), nota bene, notre-dame, nous-même(s), nouveau-né(e,s), nouveau(x)-riche(s), nu-pieds, nu-tête, nue-propriété, numerus clausus,

O, P oeil-de-boeuf, oeil-de-perdrix, oeil-de-pie, oiseau(x)-lyre, oiseau(x)-mouche, one-step, orang-outang(s), ouest-allemand(e,s), ouï-dire, outre-atlantique, outre-mer, outre-manche, outre-quiévain, outre-tombe, pale-ale, panier(s)-repas, papier(s)-calque, papier(s)-cuir, papier(s)-émeri, papier(s)-filtre, papier(s)-monnaie, papier(s)-peint(s), papier(s)-tenture, **par** : par-ci, par-dedans, par-dehors, par-delà, par-derrière, par-dessous, par-dessus, par-devant, par-devers, par-là, par-là-dessus, para-étatique(s), para-scientifique(s), parc(s)-auto, parce que, **pare** : pare-balles, pare-boue, pare-brise, pare-chocs, pare-éclats, pare-étincelles, pare-feu, pare-fumée, pare-soleil, **passe** : passe-boule(s), passe-crassane, passe-droit(s), passe-lacet(s), passe-montagne(s), passe-partout, passe-passe, passe-pied(s), passe-pierre(s), passe-plat(s), passe-temps, passe-thé, passe-velours, passe-volant, passing-shot, patte(s)-d'oie, peau(x)-rouge(s), peigne-cul, pêle-mêle, pense-bête, perce-muraille, perce-neige, perce-oreille, persona grata, persona non grata, pèse-bébé(s), pèse-lettre(s), pèse-sel(s), pèse-vin(s), pet-de-loup, pet-de-nonne, pet-en-l'air, pète-sec, **petit-** : petit(s)-beurre, petit(s)-bois, petit(e,es,s)-bourgeois(e,es), petit(e,es,s)-fils(le,s), petit-gris, petit-lait, petit(s)-maître(s), petit-nègre, petit(s)-neveu(x), petit(s)-suisse(s), petite-bourgeoisie, petite-maîtresse, petite(s)-nièce(s), petits-enfants, petits-pois, peu ou prou, peut-être, photo-électricité, photo-électrique(s), photo(s)-montage(s), photo(s)-souvenir(s), photo-synthèse, physico-chimie, physico-chimique(s), physico-mathématique(s), pick-up, pie(s)-mère(s), **pied** : pied-à-terre, pied-bot, pied-d'alouette, pied(s)-d'oiseau, pied(s)-de-biche, pied-de-cheval, pied-de-chèvre, pied-de-loup, pied(s)-de-poule, pied(s)-de-veau, pied(s)-fort(s), pied(s)-noir(s), pied(s)-plat(s), pin up, pin-pon, pince-fesse, pince-maille, pince-monseigneur, pince-nez, pince-sans-rire, ping-pong, pique-assiette(s), pique-boeuf(s), pique-feu, pique-nique(s), pique-niqueur(euse,euses,s), pis-aller, pisse-froid, pisse-vinaigre, pistolet-mitrailleur, plan-concave, plan-convexe, plat-bord, plate(s)-bande(s), plate(s)-forme(s), plateau-repas, play-boy(s), plein-emploi, plein-vent, pleure-misère, plum-pudding, plus-que-parfait, plus-value(s), poids lourd(s), poids mort(s), point(s)-virgule(s), poisson(s)-chat(s), poisson(s)-pilote(s), politico-financier(ère,ères,s), politico-judiciaire(s), politico-militaire(s), politico-mondain, pont(s)-levis, port-salut, **porte** : porte-à-faux, porte-à-porte, porte-affiche(s), porte-aiguille(s), porte-allumettes, porte-amarres, porte-avions, porte-bagages, porte-baïonnette, porte-balais, porte-bannière, porte-billets, porte-bonheur, porte-bouquet(s), porte-bouteilles, porte-brancard, porte-carte(s), porte-chapeaux, porte-cigares, porte-cigarettes, porte-clefs, porte-clés, porte-copie, porte-couteau(x), porte-crayons, porte-croix, porte-crosse, porte-documents, porte-drapeau(x), porte-enseigne, porte-épée, porte-étendard, porte-étriers, porte-étrivière, porte-faix, porte-fanion, porte(s)-fenêtre(s), porte-fort, porte-glaive, porte-greffe, porte-hauban(s), porte-jarretelles, porte-jupe(s), porte-lames, porte-malheur, porte-mine(s), porte-monnaie, porte-montre, porte-mors, porte-objet, porte-outil(s), porte-parapluies, porte-parole, porte-plume, porte-queue, porte-savon(s), porte-serviette(s), porte-vent, porte-voix, portrait(s)-robot(s), poste(s)-radio, post-scriptum, pot-au-feu, pot-bouille, pot(s)-de-vin, pot-pourri, potron-jaquet, potron-minet, pousse-café, pousse-cailloux, pousse-pied, pousse-pousse, prêchi-prêcha, premier-maître, premier-né, première-née, pré-retraité(s), président-directeur, président-directeur-général, presqu'île(s), presse-bouton, presse-citron, presse-étoupe, presse-fruits, presse-papiers, presse-purée, prêt-à-porter, pretium doloris, prie-Dieu, prima donna, pro forma, procès-verbal(aux), protège-cahier(s), protège-dents, protège-parapluie, protège-tibia(s), prud'homal(le,les,aux) prud'homme(s), public-relation, pur-sang,

Q qu'en-dira-t-on, quant-à-soi, quarante-huitard(e,es,s), quart-de-rond, quart-monde, quartier(s)-maître(s), **quasi** (liste non limitative) : quasi-agonie, quasi-certain(e,s), quasi-certitude(s), quasi-contrat(s), quasi-délit(s), quasi-mise, quasi-mort(e,s), quasi-réduction, quasi-sûr(e,s), quasi-totalité, quatre-mâts, quatre-quarts, quatre-saisons, quatre-temps, quelqu'un, quelqu'une, quelques-unes, quelques-uns, queue-de-cheval, queue-de-cochon, queue-de-morue, queue-de-pie, queue-de-poisson, queue-de-rat, queue-de-renard, qui-va-là, qui-vive, quote-part,

R rabat-joie, radio(s)-télévision(s), radio(s)-télévisé(e,es,s), rag-time, ramasse-miettes, ramasse-monnaie, ramasse-poussière, rase-mottes, ray-grass, rayon(s) x, reine-claude(s), reine-des-prés, reine-marguerite, remonte-pente(s), remue-ménage, remue-méninges, rendez-vous, réveille-matin, revenez-y, rez-de-chaussée, rince-bouche, rince-bouteilles, rince-doigts, risque-tout, rock and roll, rocking-chair, roman-feuilleton, roman-fleuve, rond-de-cuir, rond-point, ronds-de-cuir, rose-croix, rouge-queue, roulé(s)-boulé(s),

S sacré-coeur, sacro-saint(e,es,s), sage(s)-femme(s), **saint** saint-bernard, saint-cyrien(s), saint-esprit, saint-frusquin, saint-office, saint-paulin, saint-père, saint-siège, saint-simonien(ne,nes,s), saint-simonisme, sainte nitouche, saisie(s)-arrêt, saisie(s)-exécution, sang-froid, sang(s)-mêlé(s), **sans** : sans embages, sans conteste, sans encombre, sans-abri, sans-coeur, sans-culotte, sans-façon, sans-fil, sans-gêne, sans-le-sou, sans-logis, sans-parti, sans-souci, sans-travail, sapeur(s)-pompier(s), saut-de-lit, saut-de-loup, saut(s)-de-mouton, saute-ruisseau, sauve-qui-peut, savoir-faire, savoir-vivre, script-girl, seconde-main, self-control, self-made-man, self-service, **semi** : semi-aride(s), semi-automatique(s), semi-auxiliaire(s), semi-circulaire(s), semi-conducteur(,ce,ces,s), semi-désertique(s), semi-fini(e,es,s), semi-liberté, semi-nomade(s), semi-précieux(se,ses), semi-public(s), semi-publique(s), semi-remorque(s), semi-rigide(s), senatus-consulte, serbo-croate(s), sergent(s)-chef(s), serre-file(s), serre-fils, serre-frein(s), serre-joint(s), serre-livres, serre-nez, serre-papiers, serre-tête(s), sex-appeal, show-biz, side-car, sine die, sine qua non, social-démocrate(s), social-démocratie(s), sociaux-démocrates, socio-professionnel(le,s), soi-disant, soi-même, soixante-huitard(e,es,s), songe-creux, sourd(e,es,s)-muet(te,s), **sous** (liste non limitative) : sous-admissible, sous-alimentation, sous-alimenté(e,s), sous-amendement(s), sous-arrondissement(s), sous-barbe, sous-bas, sous-bibliothécaire, sous-bois, sous-brigadier(s), sous-chef(s), sous-classe, sous-commission(s), sous-comptoir, sous-consommation, sous-développé(e,s), sous-développement, sous-diaconat, sous-diacre, sous-directeur(s), sous-directrice(s), sous-emploi, sous-entendu(s), sous-entrepreneur(s), sous-équipé(e,s), sous-équipement, sous-estimation(s), sous-exposition(s), sous-faîte, sous-fifre(s), sous-garde, sous-gorge, sous-gouverneur(s), sous-homme(s), sous-ingénieur(s), sous-inspecteur (s,rice,s), sous-intendant(s), sous-jacent(e,s), sous-lieutenant(s), sous-locataire(s), sous-location(s), sous-main, sous-maître(s), sous-maîtresse(s), sous-marin(s), sous-marine, sous-marinier(s), sous-maxillaire, sous-multiple(s), sous-nappe, sous-normale, sous-oeuvre, sous-officier(s), sous-ordre(s), sous-pieds, sous-préfecture(s), sous-préfet(e,s), sous-produit(s), sous-production, sous-prolétaire(s), sous-prolétariat, sous-secrétariat, sous-seing, sous-sol(s), sous-station, sous-tangente, sous-tension, sous-titre(s), sous-traitant(e,s), sous-ventrière, sous-verge, sous-verre, sous-vêtement(s), soutien(s)-gorge, spatio-temporel(le,s), starting-block(s), station(s)-service, statu quo, steack-frites, steeple-chase, stock-car, strato-cumulus, strip tease, sud-africain(e,s), sud-américain(e,es,s), sud-coréen(ne,s), sud-est, sud-ouest, sud-viêtnamien(ne,s), super-grand(e,es,s), super-étendard(s), super-puissance(s), super-puissant(e,s), sur-le-champ, surprise-partie(s), système d.

T tac au tac, taille-crayon(s), taille-douce, taille-racines, talky-walky, tam-tam, tambour(s)-major(s), tampon-buvard, tandis que, tape-à-l'oeil, taste-vin, tensio-actif(ve,ves,s), terre-à-terre, terre(s)-neuvas, terre-neuve, terre-plein, tête-à-queue, tête-à-tête, tête-bêche, tête-de-loup, tête-de-maure, tête-de-mort, tête-de-nègre, teuf-teuf, tic-tac, tiers-monde, timbre(s)-poste, tire-au-cul, tire-au-flan, tire-botte, tire-bouchon(s), tire-clou, tire-fesse, tire-fond, tire-jus, tire-laine, tire-lait, tire-ligne(s), tire-pied, tire-veilles, tiroir(s)-caisse(s), tohu-bohu, toi-même, torche-cul, tord-boyaux, tord-nez, touche-à-tout, tourne-feuille, tout-à-l'égout, tout-fait, tout-fou(s), tout-petit(s), tout(e)-puissant(e,es,s), tout-venant, toute-puissance, trachée-artère, trade-union(s), trade-unionisme, tragi-comédie(s), tragi-comique(s), train-train, traîne-la-jambe, traîne-malheur, traîne-misère, traîne-patin(s), tranche-montagne, tranchée-abri, trench-coat, trois-mâts, trois-

points, trois-quarts, trompe-l'oeil, trompe-la-mort, trop-plein, trotte-menu, trou-madame(s), trousse-pied, trousse-queue, tsé-tsé, tue-chien, tue-diable, tue-mouche(s), turbo-alternateur,

U, V, W ultra-chic, ultra-conservateur(trice,s), ultra-court(s,e), ultra-rapide(s), ultra-royaliste(s), ultra-sensible(s), ultra-son(s), ultra-violet(te,s), va-et-vient, va-t-en guerre, va-tout, va-vite, vaso-dilatateur(trice,s), vaso-moteur(s), vélo-pousse, vert-de-gris, vert-de-grisé(e,s), **vice** : vice-amiral(aux), vice-chancelier(s), vice-consul(s), vice-légat(s), vice-légation, vice-présidence(s), vice-président(e,s), vice-recteur(s), vice-reine(s), vice-roi(s), vice-royauté, vice-versa vide-gousset, vide-ordures, vide-poches, vide-pomme, vidéo-disque(s), vidéo-cassette(s), vidéo-clip(s), vif-argent, vingt-et-unième, vis-à-vis, volley-ball, volte-face, vous-même(s), wagon-bar, wagon-lit, wagon-restaurant, wagons-lits, water-closets, water-polo, week-end

X, Y Z yacht-club.

ANNEXE 2
TABLE DES DESINENCES VERBALES

1. CLASSIFICATION

La classification des désinences du verbe dans les tableaux ci-dessous a été faite en fonction de quatre dimensions :

- Les temps et modes :
- La personne ou le genre et le nombre (participe passé)
- La classe de la conjugaison au sein du groupe (seul le troisième groupe de classes comporte plus de dix classes)

Table des désinences verbales

Paradigmes		Personnes	Nombre de classes
1.	A. Futur	1 à 6	3
	B. Conditionnel	1 à 6	3
2.	C. Présent de l'indicatif	4 et 5	8
	D. Présent du subjonctif	4 et 5	
	E. Imparfait de l'indicatif	1 à 6	
	H. Impératif	4 et 5	
	J. Participe présent		
3.	C. Présent de l'indicatif	1 à 3 et 6	30
	D. Présent du subjonctif	1 à 3 et 6	
	H. Impératif	2	
4.	I. Participe passé	1 à 4	9
5.	G. Passé simple	1 à 6	8
	F. Imparfait du subjonctif	1 à 6	
6.	K. Infinitif		6

Groupe de classes n° 1

Temps	Personne	1	2	3
A	1	erai	irai	rai
	2	eras	iras	ras
	3	era	ira	ra
	4	erons	irons	rons
	5	erez	irez	rez
	6	eront	iront	ront
B	1	erais	irais	rais
	2	erais	irais	rais
	3	erait	irait	rait
	4	erions	irions	rions
	5	eriez	iriez	riez
	6	eraient	iraient	raient

Groupe de classes n° 2

Temps	Personne	1	2	3	4	5	6	7
C	4	ons	isons	isons	vons	vons	eons	çons
	5	ez	Ites	ites	vez	vez	ez	cez
D	4	ions	isions	ssions	chions	yons	ions	cions
	5	iez	Isiez	ssiez	chiez	yez	iez	ciez
E	1	ais	Isais	isais	vais	vais	eais	çais
	2	ais	Isais	isais	vais	vais	eais	çais
	3	ait	isait	isait	vait	vait	eait	çait
	4	ions	isions	isions	vions	vions	ions	cions
	5	iez	isiez	isiez	viez	viez	iez	ciez
	6	aient	isaient	isaient	vaient	vaient	eaient	çaient
H	4	ons	isons	isons	chons	yons	eons	çons
	5	ez	îtes	ites	chez	yez	ez	cez
J	0	ant	isant	isant	chant	yant	eant	çant

Groupe de classes n° 3

Temps	Pers.	1	2	3	4	5	6	7	8	9	10
C	1	e	s	ds	us	s	ieds	s	s	s	-
	2	es	s	ds	us	s	ieds	s	s	s	-
	3	e	t	d	ut	t	ied	t	t	t	-
	6	ent	ssent	dent	lvent	ent	eyent	vent	illent	sent	-
D	1	e	sse	de	lve	e	eye	ve	ille	se	-
	2	es	sses	des	lves	es	eyes	ves	illes	ses	-
	3	e	sse	de	lve	e	eye	ve	ille	se	-
	6	ent	ssent	dent	lvent	ent	eyent	vent	illent	sent	-
H	2	e	s	ds	us	s	ieds	s	s	s	-

(Groupe de classes n° 3, suite)

Temps	Pers.	11	12	13	14	15	16	17	18	19	20
C	1	ds	ns	îs	s	ais	us	is	s	ds	is
	2	ds	ns	îs	s	ais	us	is	s	ds	is
	3	d	nt	ît	t	ait	ut	it	t	d	ît
	6	sent	gnent	issent	ment	ont	illent	ïssent	tent	lent	isent
D	1	se	gne	isse	me	asse	ille	ïsse	te	le	ise
	2	ses	gnes	isses	mes	asses	illes	ïsses	tes	les	ises
	3	se	gne	isse	me	asse	ille	ïsse	te	le	ise
	6	sent	gnent	issent	ment	assent	illent	ïssent	tent	lent	isent
H	2	ds	ns	is	s	ais	us	is	s	ds	is

(Groupe de classes n° 3, suite)

Temps	Pers.	21	22	23	24	25	26	27	28	29	30
C	1	eux	ds	is	s	ux	ts	ts	x	cs	is
	2	eux	ds	is	s	ux	ts	ts	x	cs	is
	3	eut	d	it	t	ut	t	t	t	c	ît
	6	euvent	nent	vent	nent	lent	ttent	tent	lent	quent	issent
D	1	uisse	ne	che	ne	ille	tte	te	ille	que	isse
	2	uisses	nes	ches	nes	illes	ttes	tes	illes	ques	isses
	3	uisse	ne	che	ne	ille	tte	te	ille	que	isse
	6	uissent	nent	chent	nent	illent	ttent	tent	illent	quent	issent
H	2	eux	ds	che	s	ux	ts	ts	ille	cs	is

Groupe de classes n° 4

Temps	Pers.	1	2	3	4	5	6	7	8	9
I	1	é	i	u	û	s	s	t	û	ï
	2	és	is	us	us	s	s	ts	ûs	ïs
	3	ée	ie	ue	ue	te	se	te	ûe	ïe
	4	ées	ies	ues	ues	tes	ses	tes	ûes	ïes

Groupe de classes n° 5

Temps	Personne	1	2	3	4	5	6	7	8
G	1	ai	is	us	ûs	ïs	ins	eai	çai
	2	as	is	us	ûs	ïs	ins	eas	ças
	3	a	it	ut	ût	ït	int	ea	ça
	4	âmes	îmes	ûmes	ûmes	ïmes	înmes	eâmes	çâmes
	5	âtes	îtes	ûtes	ûtes	ïtes	întes	eâtes	çâtes
	6	èrent	irent	urent	ûrent	ïrent	inrent	èrent	çâtes
F	1	asse	isse	usse	ûsse	ïsse	insse	easse	çasse
	2	asses	isses	usses	ûsses	ïsses	insses	easses	çasses
	3	ât	ît	ût	ût	ït	înt	eât	çât
	4	assions	issions	ussions	ûssions	ïssions	inssions	eassions	çassions
	5	assiez	issiez	ussiez	ûssiez	ïssiez	inssiez	eassiez	çassiez
	6	assent	issent	ussent	ûssent	ïssent	inssent	eassent	çassent

Groupe de classes n° 6

Temps	1	2	3	4	5	6
K	er	ir	re	oir	eoir	ïr

ANNEXE 3
LES HOMOGRAPHIES DU PARTICIPE PASSE.
(Classement alphabétique)

Cette liste ne prétend pas être exhaustive. Normalement, l'homographie est signalée sous la forme canonique de l'adjectif ou du substantif (masculin singulier pour les adjectifs, masculin ou féminin singulier pour les substantifs) sauf quand elle concerne également un autre mode du même verbe (verbes se terminant en 'is' notamment) ou lorsqu'elle englobe un substantif féminin (verbes se terminant en 'ées' notamment). Généralement, pour les adjectifs les homographies concernent également le féminin singulier ainsi que les masculin et féminin singulier et pluriel. Pour les substantifs, le pluriel est également concerné. Ces flexions ne sont pas mentionnées pour ne pas allonger la liste.
Les formes placées en italiques signalent une homographie avec un substantif.

abasourdi(s), abâtardi(s), abatti(s), abêti(s), aboli(s), abruti(s), absout, accompli(s), accroupi(s), *accidenté*, accueilli(s), *accusé*, *acqui(s), adjoint*, admi(s), adouci(s), affadi(s), affaibli(s), affairé, affecté, affermi(s), affiché, affranchi(s), agencé, agrandi(s), aguerri(s), ahuri(s), aidé, aigri(s), aiguillonné, alarmé, alerté, *allée*, allégé, *allié*, allongé, alourdi(s), amaigri(s), amalgamé, amarré, amené, amendé, aminci(s), amnistié, amoindri(s), amolli(s), amorcé, amorti(s), *amputé*, analysé, ancré, anéanti(s), anémié, anesthésié, angoissé, ankylosé, annexé, annoncé, anobli(s), antidaté, *aperçus*, aplani(s), aplati(s), apostrophé, apparenté, appauvri(s), *appelé*, appesanti(s), appliqué, appri(s), apprêté, approfondi(s), archivé, *armée*, *arrivée*, arrondi(s), aspergé, asphyxié, assagi(s), assaini(s), assassiné, *assemblée*, asservi(s), assi(s), *associé*, assombri(s), assorti(s), assoupi(s), assoupli(s), assourdi(s), assouvi(s), assujetti(s), atrophié, *attaché*, attaqué, atteint, *atteinte*, attellé, attendri(s), attrapé, auréolé, avachi(s), *avancée*, avantagé, averti(s), aveuglé, avili(s), avisé, avoué, axé

bâché, bafoué, bagué, baissé, baladé, balafré, balancé, balisé, balotté, bandé, banni(s), bardé, barré, barricadé, basculé, basé, *bâti*, bâti(s), *battue*, bée, béni(s), bercé, berné, beurré, biaisé, bichonné, biffé, bitumé, blâmé, blanchi(s), *blasé*, blémi(s), *blessé*, bleui(s), *blindé*, blondi(s), blotti(s), bombardé, bombé, borné, botté, bouché, *bouchée*, bouclé, *bouffée*, bougé, bougonné, boulé, boulotté, bourré, *bourrée*, boxé, braillé, branché, brandi(s), braqué, brassé, *brassée*, bredouillé, bricolé, bridé, brigué, briqué, brisé, *brisée*, broché, bronzé, brossé, brouillé, bruni(s), brusqué, *bus*, buté, *butée*

câblé, cabossé, caché, cachetté, cadencé, cadré, caillé, calé, calibré, calligraphié, calmé, calomnié, calotté, calqué, canné, capsulé, captivé, capturé, cardé, caréné, caressé, cargué, caricaturé, carié, carossé, *carré*, casé, caserné, casqué, cassé, catalogué, catapulté, catastrophé, causé, ceinturé, célébré, censuré, centré, cerclé, cerné, chagriné, chambré, *chambrée*, changé, chapitré, *chargé*, charmé, chassé, chatouillé, chauffé, chaussé, *chaussée*, *cheminée*, chemisé, chéri(s), *chevauchée*, chevillé, chiffré, chiqué, chloroformé, choisi(s), chromé, chronométré, *cinglé*, cintré, circonscrit, circonsi(s), *ciré*, cisaillé, *cité*, claqué, classé, *cliché*, civilisé, cloîtré, cloqué, *clos*, clôturé, coalisé, coché, codé, cogné, *cognée*, coiffé, collé, collecté, *colonisé*, combatu, *combiné*, comblé, commandé, *commi(s),* complété, compri(s), *compromi(s),* compté, conclus, concurrencé, conçus, *condamné*, *conduit*, *conduite*, confessé, *confit*, connus, conqui(s), consacré, consenti(s), conservé, considéré, consigné, consolé, construit, conté, *contenu*, contenté, contesté, contingenté, continué, contracté, contraint, *contrainte*, contrasté, contré, *contrée*, contrebalancé, contredit, contrefait, contrôlé, controversé, *converti(s),* copié, *corrigé*, *corrompu*, corsé, costumé, *coté*, couché, coudé, *coudée*, *coulée*, *coupé*, *coupée*, couplé, courbé, couronné, couru, *couvée*, *couvert*, craint, *crainte*, cravaché, *crépi*(s), creusé, crevassé, criblé,

critiqué, *croisé*, *croisée*, *crue*, crus, crûs, cueilli(s), *cuirassé*, cuisiné, cuit, *cuite*, culbuté, *curé*, *curée*, cuvé, *cuvée*

dactylographié, damé, daté, *débauché*, *débilité*,*débouché*, décalqué, déchargé, décharné, déchu, décompté, découpé, *découvert*, *découverte*, décrit, décuplé, déçu, dédicacé, *dédit*, déduit, défait, *défaite*, défavorisé,*défilé*, défini(s), défraîchi(s), *défroqué*, dégainé, dégarni(s), dégluti(s), dégourdi(s), *dégradé*, dégrossi(s), délabré, *délégué*, *délibéré*, demandé, démarqué, *démenti(s),* demeuré, démi(s), démoli(s), démuni(s), dépêché, dépeint, dépensé, dépoli(s), dépouillé, dépri(s), *dérivé*, dérobé, *dérobée*, dérouté, désabusé, désaccordé, désaffecté, désagrégé, désappointé, désappri(s), désarmé, désarticulé, désassorti(s), désavantagé, *désaxé*, désempli(s), *déséquilibré*, déserté, *désespéré*, désintéressé désobéi(s), dessaisi(s), desservi(s), *destinée*, désuni(s), détaxé, déteint, *détenu*, déterminé, détruit, *dictée*, *différé*, diffusé, discipliné, discontinu, disjoint, *disparu*, *dispensé*, disposé, disputé, dissolu, distancé, distrait, dit, diverti(s), *divorcé*, domestiqué, donnée, dosé, doublé, douché, dragué, drainé, *drogué*, dupé, durci(s), *durée*, *dus*, dynamité

ébahi(s), ébauché, ébloui(s), ébranlé, écaillé, échangé,échappée, écharpé, *éclaircie*, éclairci(s), éclipsé, éconduit, écouté, *écrit*, écumé, édité, effaré, égalé, élargi(s), élevé, *élu*, *embauché*, *embellie*, embelli(s), emblavé, embouché, embouti(s), embrassé, embué, *émigré*, émi(s), empli(s), empoigné, empreint, *empreinte*, emprunté, émus, enchéri(s), enclavé, enclenché, encombré, encourus, encré, endormi(s), *enduit*, endurci(s), enfoui(s), enfui(s), englouti(s), enhardi(s), *enjambée*, enjoint, enlaidi(s), ennobli(s), enqui(s), *enragé*, enrichi(s), enseigné, enseveli(s), entaillé, entendu, enthousiasmé, entraperçus, entraidé, entravé, *entrée*, *entremise*, *entreprise*, entretenu, *entrevue*, envahi(s), enveloppé, envié, *envoyé*, épaissi(s), épanoui(s), épargné, épicé, épinglé, épongé, épousé, épouvanté, épri(s), épuré, équarri(s), équilibré, *équipée*, escaladé, escompté, escorté, espacé, espionné, esquissé, esquivé, estampillé, estimé, estompé, estourbi(s), *estropié*, *établi(s),* étagé, étalé, *été*, éteint, étendu, *étendue*, étiquetté, étoffé, étoilé, étouffé, étouffée, étourdi(s), étreint, *étreinte*, eus, évacué,*évadé*, évanoui(s), *exclus*, excusé, *exempté*, *exilé*, *expatrié*, explicité, *exploité, exposé*,*extrait*

fabriqué, facturé, faibli(s), *failli*,*fait*, fané, farci(s), fatigué, fauché, faussé, fécondé, *fédéré*, feint, *feinte*, fêlé, fermé, *fessée*, fêté, feuilletté, feutré, *fiancé*, *fiancée*, ficellé, fiché, figuré, filé, filtré, financé, fini(s), fissuré, fixé, *flambée*, fléché, fléchi(s), flétri(s), fleuri(s), fondé, forcé, forci(s), forgé, formé, formulé, fouillé, foulé, *foulée*, fourbi(s), fourni(s), fourré, *fusillé*, fracturé, fraîchi(s), franchi(s), frappé, fréquenté, *fricassée*, fringué, frisé, frit, *frite*, froncé, frustré, *fumée*, *fusée*

gâché, gagé, gainé, galonné, *galvanisé*, galvaudé, gangrené, garanti(s), *garantie*, gardé, garé, *garni(s)*, gâté, gauchi(s), gavé, gazé, *gelée*, gêné, gercé, germé, giflé, givré, glacé, gommé, gondolé, gorgé, *gorgée*, goupillé, goutté, gradué, graissé, grandi(s), gratiné,*gratinée*, gratté, gravé, greffé, grêlé, grêvé, gribouillé, griffé, grimacé, grimé, grippé, grisé, grossi(s), groupé, guéri(s), gueulé, guidé, guigné, guillotiné

habitué, haché, hâlé, hébergé, hérissé, homologué, honni(s), *huée*, huilé, hypertrophié, hypothéqué,

idolâtré, *illustré*, *immigré*, imparti(s), importuné, incendié, inclus, incommodé, indigné, induit, infecté, infléchi(s), influencé, informé, infusé, injecté, innocenté, inquiété, *inscrit*, inspecté, *instruit*, insulté, issu, *issue*, *insurgé*, intercepté, *interdit*, *interné*, interprété, interverti(s), *intitulé*, intoxiqué, intrigué, introduit, inventé, inversé, investi(s), *invité*

jailli(s), jalousé, jaugé, jauni(s), *jetée*, *joint*, jugé, jumellé, juré,

lâché, laissé, *lampée*, *lancée*, lardé, largué, lassé, lavé, léché, légitimé, lésé, lessivé, lesté, leurré, *levée*, lézardé, *libellé*, ligué, limé, limité, liquidé, lissé, livré, *livrée*, logé, loti(s), loupé, lu, lustré, luxé

mâché, maculé, maigri(s), maîtrisé, manié, manipulé, manoeuvré, manqué, marbré, *marché*, *marié*, marqué, masqué, massacré, massé, maté, *matinée*, matraqué, *maudit*, méconnus, médit, médité, médusé, mélangé, *mêlée*, menacé, *menée*, méprisé, mérité, mesuré, métamorphosé, *meublé*, meurtri(s), miné, mi(s), *mise*, mitraillé, *mobilisé*, *modelé*, moisi(s), molli(s), *montée*, montré, moqué, mouché, mouillé, moulé, moulus, mugi(s), muni(s), mûri(s), murmuré, mus, *mutiné*

nanti(s), *naufragé*, né, *nécessité*,*nichée*, noirci(s), noté, noué, nourri(s), *noyé*, nuancé, *nuit*

obéi(s), *obsédé*, obscurci(s), offensé, oint, ombragé, ombré, omi(s), orchestré, orné, orthographié, ouaté, oublié, ourdi(s), outragé, outré, oxydé, oxygéné,

pâli(s), palpé, pansé, parachuté, parafé, paraffiné, parasité, parcouru, *parfait*, parodié, parqué, partagé, *parti*, *partie*, *parti(s)*, parus, *parvenu*, *passé*, patiné, patronné, paumé, *pêché*, peigné, peiné, peint, *pensée*, percé, *percée*, perché, perçus, perlé, *permi(s),* persécuté, perverti(s), pétri(s), peuplé, photocopié, piloté, *pincée*, pipé, *piqué*, piquetté, placé, plaint, *plainte*, *planqué*, planté, plaqué, plâtré, plébiscité, plié, *plongée*, plumé, poché, policé, poli(s), *polycopié*, poncé, *portée*, posé, posté, poudré, pourri(s), *poussée*, pratiqué, prêché, prédit, préfacé, *préjugé*, prémuni(s), prescrit, présenté, pressé, pressenti(s), prêté, prétexté, prévalus, primé, pri(s), *prise*, prisé, *privilégié*, *procédé*, prodigué, *produit*, profané, professé, programmé, *promi(s)*, *promise*, *promu*, prôné, *proscrit*, *puni(s)*, purgé

quadrillé, quadruplé, quintuplé, quitté,

rabougri(s), raccourci(s), *raclée*, racorni(s), radiodiffusé, radiographié, radouci(s), raffermi(s), rafraîchi(s), ragaillardi(s), raidi(s), rajeuni(s), *ralenti(s)*, rallongé, ramolli(s), *randonnée*, rangé, *rangée*, *rapatrié*, *râpé*, rassi(s), raté, *ratée*, ravagé, raviné, ravi(s), réadmi(s), réassorti(s), rebâti(s), recherché, récolté, récompensé, reconduit, reconnu(s), reconstruit, reconverti(s), recoupé, recueilli(s), recuit, reçu(s), redit, *redite*, *réduit*, réélu, refait, réfléchi(s), réformé, *réfugié*, refroidi(s), regarni(s), régi(s), réglé, rejailli(s), rejeté, rejoint, réjoui(s), relâché, relancé, relaxé, relevé, relu(s), remarqué, rembruni(s), remi(s), *remise*, remisé, *remontée*, remorqué, rempli(s), renchéri(s), rencontré, rengainé, *renommée*, *rentrée*, renversé, réparti(s), *répartie*, *repenti*(s), repéré, réprimandé, *repri(s)*, *reprise*, repus, requi(s), réservé, résolu, ressaisi(s), ressenti(s), resserré, resservi(s), ressorti(s), restreint, *résumé*, rétabli(s), retombé, *retombée*, retouché, retransmi(s), rétréci(s), retrempé, rétrogradé, réuni(s), réussi(s), rêvé, revécu(s), reverdi(s), *révolté*, *revue*, ridé, rimé, risqué, rôti(s), roué, rougi(s), rouillé, roussi(s), *ruée*, ruiné, rusé, rythmé

sablé, sabré, saccagé, sacré, *saisie*, saisi(s), *salarié*, salé, sali(s), salopé, saoulé, sapé, satisfait, *sauté*, sauvé, sauvegardé, scié, sclérosé, séché, secondé, séduit, semé, semoncé, senti(s), séparé, séquestré, seriné, serré, serti(s), servi(s), sévi(s), signé, singé, soldé, sondé, sorti(s), *sortie*, soudé, *soufflé*, souillé, soumi(s), souscrit, soustrait, stimulé, *stratifié*, structuré, *subordonné*, sucré, suicidé, suivis, surchargé, surchauffé, surfait, surpri(s), *surprise*, survécus, sus

tablée, taché, taillé, tanné, tapé, taquiné, tari(s), taxé, teint, *teinte*, teinté, télécommandé, télégraphié, téléphoné, télévisé, tenaillé, *tenue*, tenté, terrassé, ternis, terré, tiédi(s), tiré, titré, toisé, *tombée*, toqué, torpillé, tortillé, torturé, touché, tourmenté, *tournée*, tracé, traduit, trahi(s),

traînée, *trait*, *traite*, *traîté*, tramé, *tranchée*, transcrit, transfusé, transgressé, transi(s), transmi(s), traqué, *traversée*, *travesti(s)*, tremblé, trempé, tressailli(s), triplé, trompé, troublé, troussé, truffé, tu, tuméfié, tu(s), *tué*

ulcéré, unis, usiné, vacciné, validé, valus, vécus, *veillée*, vendangé, verdis, *vernis*, versé, vidé, vieillis, violé, visité, *visée*, vitré, volée, vomis, voté, voulus, voûté, *vue*

zébré.

ANNEXE 4
LES HOMOGRAPHIES DU PARTICIPE PRESENT.
(Classement alphabétique)

Les formes qui sortent du cas normal (verbes/adjectifs) sont placées en italiques. Il s'agit des homographies entre les verbes et les{adjectif-substantif} ou enventuellement des prépositions, des adverbes…

abattant, abondant, *aboutissant*, accablant, accaparant, *absorbant*, *accédant*, accommodant, accueillant, *acidifiant*, *activant*, adonnant,*adoucissant*, affaiblissant, affligeant, affolant, *africanisant*, agaçant, agglomérant, agglutinant, aggravant, agissant, *agonisant*, ahurissant, aiguillonnant, *aimant*, alarmant, aliénant, allaitant, *allant*, alléchant, altérant, alternant, *amaigrissant*, *américanisant*, *amincissant*, amplifiant, amolissant, amusant, *anesthésiant*, *anglicisant*, angoissant, apaisant, apitoyant, appauvrissant, appelant, approchant, *arabisant*, *arc-boutant*, arrangeant, arrivant, *aspirant*, *assaillant*, assainissant,*assiégeant*, *assistant*, assommant, *assouplissant*, assourdissant, astreignant, attachant, *attaquant*, attendrissant, atténuant, atterrant, attirant, attristant, *avenant*, aveuglant, avilissant, asphyxiant, *azurant*

babillant, badinant, bafouillant, baillant, baissant, balançant, balayant, balbutiant, *ballant*, ballonnant, ballotant, bandant, baragouinant, barbant, barbotant, basculant, *battant*, bavant, béant, bedonnant, bégayant, bêlant, berçant, *beuglant*, bidonnant, blâmant, blanchissant, blêmissant, blessant, bleuissant, blondissant, bluffant, boitillant, bombant, bottant, bouffant, bougonnant, bouillant, bouillonnant, bouleversant, bourdonnant, bourgeonnant, bourrant, branlant, brasillant, bredouillant, *brillant*, brimant, brinqueballant, *brisant*, bronzant, bruissant, brûlant, brunissant

cafouillant, cahotant, cajolant, câlinant,*calmant*, captivant, caquetant, caracolant, *carburant*, caressant, carillonnant, cascadant, cassant, causant, *cédant*, *célébrant*, cessant, chagrinant, chancelant, changeant, chantant, charmant, chatouillant, chatoyant, chauffant, chaussant, chavirant, chevrotant, chiant, chiffonnant, chipotant, choquant, chuchotant, chuintant, *cicatrisant*, cinglant, claironnant, clapotant, claquant, claudiquant, *clignotant*, cliquetant, *coagulant*, coassant, coiffant, *collant*, *colorant*, *combattant*, *commandant*, commençant, *commerçant*, *commettant*, commotionnant, *communiant*, compatissant, complaisant, complexant, *composant*, compromettant, *comptant*, *concédant*, *concélébrant*, concertant, conciliant, concluant, concordant, condescendant, confiant, confondant, congestionnant, *conquérant*, consentant, *considérant*, consistant, consolant, consternant, *constipant*, *constituant*, *consultant*, contaminant, *contenant*, *contractant*, contraignant, contrariant, *contre-attaquant*, *contremanifestant*, *contrevenant*, contristant, convainquant, convenant, convergeant, convulsant, *coopérant*, *correspondant*, corrodant, *cotisant*, *couchant*, couinant, coulant, coulissant, coupant, *courant*, crachant, crachotant, craquant, craquelant, *crémant*, crépitant, crevant, criant, crispant, crissant, *croissant*, *croquant*, *croulant*, croupissant, croustillant, *croyant*, cuisant, culbutant, culminant, culpabilisant

dandinant, dansant, débectant, débilitant, *débitant*, débordant, déboussolant, *débroussaillant*, *débutant*, *décalaminant*, *décapant*, décevant, déchirant, déclamant, déclinant, *décolorant*, décomplexant, déconcertant, *décongestionnant*, *déconstipant*, décourageant, décroissant, déculpabilisant, défaillant, déferlant, déformant, défoulant, défrisant, *dégivrant*, dégoulinant, dégoûtant, dégradant, dégraissant, dégrisant, délassant, délirant, démangeant, *démaquillant*,

démêlant, démobilisant, démoralisant, démystifiant, dénigrant, dépendant, dépersonnalisant, dépitant, déplaisant, *dépliant*, dépolluant, *déposant*, déprimant, déraillant, dérangeant, dérapant, déroulant, déroutant, désaltérant, désappointant, désarçonnant, désarmant, désaxant, *descendant*, désenivrant, *désensibilisant*, déséquilibrant, désespérant, *désherbant*, déshonorant, déshumanisant, *déshydratant*, *désinfectant*, désirant, désobéissant, désobligeant, désolant, désopilant, desséchant, déstructurant, *détachant*, *détaillant*, *détartrant*, détendant, *déterminant*, détonant, détraquant, dévalorisant, déversant, *déviant*, dévorant, diffractant, *diluant*, *dirigeant*, disant, disconvenant, discordant, *discriminant*, *discutant*, dispersant, disqualifiant, *dissolvant*, dissonant, distrayant, divaguant, divergeant, divertissant, dodelinant, *dominant*, donnant, dopant, dormant, douchant, drapant, *durant*

éblouissant, ébouriffant, écaillant, échauffant, éclairant, éclatant, écoeurant, écorchant, écrasant, écumant, édifiant, *édulcorant*, effarant, effarouchant, effrayant, égarant, égayant, emballant, embarassant, embaumant, embellissant, embêtant, émergeant, émerveillant, *émigrant*, emmerdant, émotionnant, émoustillant, émouvant, empestant, empirant, empoisonnant, *émulsifiant*, encerclant, enchérissant, encombrant, encourageant, endiablant, endormant, endurant, endurcissant, énervant, enflammant, engageant, englobant, engourdissant, enivrant, enrichissant, enrobant, *enseignant*, ensorcelant, entendant, entêtant, enthousiasmant, entortillant, entraînant, entreprenant, entrant, envahissant, enveloppant, environnant, envoûtant, épaississant, épanouissant, *épargnant*, épatant, *épilant*, époustouflant, éprouvant, épuisant, équilibrant, ergotant, errant, éructant, esquintant, essouflant, *estivant*, étamant, éternuant, étincelant, étonnant, étouffant, étourdissant, *étudiant*, exaltant, exaspérant, *excitant*, *exécutant*, exigeant, *exploitant*, *exposant* , exténuant, exultant,

fabriquant, fâchant, faiblissant, facinant, fascisant, faseillant, fatiguant, feignant, fendant, ferraillant, *fertilisant*, *figurant*, filant, filochant, filtrant, flambant, flamboyant, fléchissant, florissant, flottant, fluctuant, foisonnant, *fondant*, formant, *fortifiant*, foudroyant, foulant, fourmillant, fracassant, fraîchissant, frappant, fredonnant, frémissant, frétillant, frigorifiant, frisant, frisottant, frissonnant, froissant, frustrant, fuyant, fulgurant, fumant,

gagnant, galopant, gambadant, *garant*, gargouillant, gazouillant, gémissant, gênant, gérant, gesticulant, gisant, givrant, glaçant, glissant, gloussant, gondolant, gonflant, *gouvernant*, grandissant, gratifiant, grésillant, grimaçant, grimpant, grinçant, grisonnant, grondant, grouillant, gueulant,

habitant, haletant, hallucinant, harassant, hésitant, hibernant, horrifiant, horripilant, humiliant, hurlant, hydratant,

identifiant, *ignorant*, *immigrant*, imperméabilisant, implorant, *important*, imposant, impressionnant, incommodant, indisposant, insdustrialisant, infantilisant, infectant, inhibant, innovant, inquiétant, insensibilisant, insinuant, insistant, insultant, intégrant, intéressant, *intervenant*, intimidant, invalidant, invitant, irradiant, irritant, *isolant*

jacassant, jaillissant, jargonnant, jaunissant, jouissant,

lancinant, languissant, larmoyant, lassant, lénifiant, *levant*, *liant*, louchant, louvoyant, *lubrifiant*, luisant, lustrant,

maintenant, malfaisant, *mandant*, *manifestant*, manoeuvrant, manquant, maquillant, marchant, marmonnant, marquant, marrant, maugréant, médisant, médusant, méfiant, menaçant, *mendiant*, méprisant, méritant, meuglant, *migrant*, *militant*, mimaudant, minant, miroitant, *montant*,

mordant, mortifiant, motivant, moulant, *mourant*, moussant, moutonnant, mouvant, moyennant, mugissant, mûrissant, murmurant, *mutant*, mystifiant,

naîssant, nasillant, naviguant, navrant, *négociant*, *nettoyant*, nourrissant

obéissant, obligeant, obnubilant, obsédant, *occupant*, offensant, *officiant*, ondoyant, ondulant, *opposant*, oppressant, oscillant, outrageant, *oxydant*,

palpitant, paniquant, pantelant, papillonnant, papotant, paradant, paralysant, *participant*, *partant*, *passant*, passionnant, patoisant, payant, pénalisant, *penchant*, pendouillant, *pendant*, pénétrant, pensant, perçant, percutant, *perdant*, pérorant, persévérant, persistant, perturbant, pesant, pétaradant, pétant, pétillant, pétrifiant, piaffant, *piquant*, pivotant, *plaignant*, plaisant, planant, *pliant*, plongeant, poignant, *polluant*, pompant, pontifiant, *portant*, *possédant*, *postulant*, pouffant, pourrissant, *poursuivant*, *pratiquant*, préexistant, prenant, préoccupant, pressant, *prétendant*, prévalant, prévenant, prévoyant, proliférant, *protestant*, provoquant, puant, pullulant,

rabattant, radotant, rafraîchissant, rageant, ragoûtant, râlant, ramollissant, *rampant*, rasant, rassérénant, rassurant, ravigotant, ravissant, rayonnant, rebondissant, rebutant, réchauffant, *récitant*, *récoltant*, *réconfortant*, reconnaissant, *reconstituant*, récriminant, *récurant*, *redoublant*, réfléchissant, réfrigérant, refroidissant, regardant, régénérant, régnant, réjouissant, relaxant, relevant, reluisant, remettant, *remontant*, *remplaçant*, remuant, renaissant, rentrant, renversant, *repentant*, *répondant*, reposant, repoussant, *représentant*, répugnant, *requérant*, résidant, *résistant*, resplendissant, ressemblant, *ressortissant*, *restaurant*, *restant*, résultant, retentissant, rétrécissant, rétrocédant, *revenant*, revigorant, révoltant, révulsant, ricanant, riant, ronchonnant, ronflant, ronronnant, rosissant, roucoulant, rougeoyant, rougissant, roulant, roussissant, rugissant, ruisselant, *ruminant*

saignant, *saillant*, saisissant, salissant, sanglant, sanglotant, saoulant, satisfaisant, sautillant, scintillant, séduisant, *semblant*, seyant, *servant*, sidérant, sifflant, sifflotant, *signifiant*, *soignant*, sommeillant, sonnant, sortant, souffrant, soûlant, soumettant, *soupirant*, souriant, *sous-traitant*, stérilisant, *stimulant*, *stupéfiant*, subjuguant, suant, suffisant, suffoquant, suintant, *suivant*, *suppléant*, suppliant, suppurant, surabondant, surplombant, surprenant, *surveillant*, *survivant*, *sympathisant*,

tambourinant, tapant, tâtonnant, tenaillant, *tenant*, tentant, terrifiant, terrorisant, tiédissant, *tirant*, titillant, titubant, tolérant, tombant, tonifiant, tonitruant, tonnant, tordant, touchant, tourbillonnant, tourmentant, *tournant*, tournoyant, tracassant, traçant, traînant, traitant, *tranchant*, *tranquilisant*, transcendant, transhumant, traumatisant, trébuchant, tremblant, tremblotant, trépidant, trépignant, triomphant, trottinant, troublant, tuant,

usant,
vacillant, valorisant, vannant, vasouillant, *versant*, vibrant, vieillissant, *vitrifiant*, vivifiant, *vivant*, *voyant*, *volant*, *votant*,

ANNEXE 5
LES AUTRES HOMOGRAPHIES DU VERBE.
(liste alphabétique)

Pour les verbes du premier groupe (du type "abîme, abîmes"), seule la première personne est mentionnée.

abat, abîme, absous, acceptions, acquit, active, adjoint, admirent, adoptions, adresse, adultère, affaire, affections, affiche, affluent, agence, agis, agrafe, agressions, aide, aiguille, aiguillons, alarme, alerte, allège, alliez, allions, allonge, amalgame, amarre, ambre, amène, amende, amnistie, amorce, analyse, ancre, anémie, anesthésie, angoisse, ankylose, annexe, annonce, antidate, apostrophe, apparente, applique, approche, arbitre, archive, arme, arnaque, as, asperge, asphyxie, assassine, atrophie, attache, attaque, attelle, attente, attentions, attrape, augure, aura, auréole,avance, avantage, aventure, aveugle, avions, avoir, axe

bâche, badine, bafouille, bagarre, bague, baguenaude, baie, baille, bâillons, baise, baisse, balade, balafre, balance, balise, balotte, bamboche, bande, banque, barbe, barre, barricade, bascule, base, bassine, bataille, bâtisse, bâtons, bats, batte, bavarde, bave, bêche, berne, besogne, beurre, biaise, biche, bichonne, biffe, bigle, bigorne, bile, bise, bisons, bisque, bitume, blague, blâme, blasons, blasphème, blouse, blousons, bobine, bois, bombarde, bombe, borne, bosse, botte, bouche, boucher, bouchons, boucle, boue, bouffe, bouffons, bouge, bougonne, bouille, bouillons, boule, boulotte, bourre, bout, boutons, boxe, braille, braise, brame, branche, branle, braque, brasse, brave, bredouille, bricole, bride, bridge, brigue, brique, brise, brocante, broche, bronche, bronze, brosse, brouette, brouillasse, brouille, brouillons, bruine, bruit, brume, brusque, bûche, bûcher, bus, but, butte

câble, cabosse, cabriole, cache, cachette, cadence, cadre, caille, cale, calibre, câline, calligraphie, calme, calomnie, calotte, calque, cambriole, cane, canne, canons, canule, capitule, capote, capsule, captive, capture, caracole, carapace, carbure, carde, carence, carène, caresse, cargue, caricature, carie, carrosse, carotte, carre, cascade, case, caserne, casque, casse, castagne, catalogue, catalyse, catapulte, catastrophe, cause, cavalcade, cavale, ceinture, ceinturons, cela, célèbre, censure, centre, centuple, cercle, cerne, cesse, cessions, chagrine, chambre, change, chapitre, charge, charme, chasse, chatouille, chauffe, chaume, chausse, chaussons, chemise, cheville, chicane, chiffre, chine, chinoise, chique, cloque, chloroforme, choient, chope, choucroute, chrome, chronomètre, chute, cintre, cire, cisaille, claque, classe, clenche, cloche, cloître, cloque, clôture, coche, cocher, cochonne, cochons, code, coffre, cogne, coiffe, colle, collecte, combat, combine, comble, commande, commerce, commère, communions, comparais, complainte, complète, complexions, compte, concorde, concours, concurrence,confesse, confessions, conforme, conjecture, conseiller, conserve, consigne, console, conte, content, contente, contentions, conteste, contingente, continue, contractions, contraste, contre, contre-attaque, contrebalance, contrôle, controverse, convergent, conversions, convient, copie, copine, copule, corne, corse, costume, cote, cotons, couche, coude, coulisse, coupe, couple, coupons, courbe, couronne, cours, court, cousine, couvent, crâne, cravache, cravate, crème, crêpe, creuse, crevasse, crevassent, crible, critique, croît, crotte, croûte, crûmes, crûtes, cube, cueille, cuirasse, cuisine, culbute, cure, cuve, cylindre

dactylographie, dalle, dame, danse, date, daube, débat, débauche, débine, décalque, décharge, décompte, découpe, décroît, décuple, dédicace, défroque, dégaine, déjections, délire, demande, démarche, démarque, dément, démerde, démérite, demeure, départ, dépêche, dépense, dépouille, déprime, dérive, déroute, désavantage, déséquilibre, déserte, dessert, détaxe, devins, devise, diagnostique, dialogue, dictions, dictons, diffuse, diffusions, dîner, dingue, diphtongue, diplôme, discipline, discours, dispense, dispersions, dispose, dispute, distance, distrait, divergent, divisions, divorce, domestique, dominions, donne, dose, double, doublons, douche, doute, dragues, draine, drogue, drosse, daube, dupe, dure, dynamite

ébat, ébauche, écaille, échange, écharpe, échine, éclipse, écope, écossais, écoute, écume, éditions, égale, élève, embauche, emblave, embouche, embrasse, embrouille, embue, empire, empoigne, encaisse, enclave, encombre, encre, enquête, enseigne, entaille, entame, enthousiasme, entraide, entrave, entre, entremet, entretiens, enveloppe, envie, épargne, épaule, épice, épilogue, épingle, éponge, épouse, épouvante, épure, équilibre, équipe, équivalent, erre, est, escalade, escompte, escorte, escrime, espace, espionne, esquisse, esquive, est, estampe, estampille, estime, estompe, étage, étale, étalons, étanche, été, éteint, étincelle, étiquette, étoffe, étoile, étourdis, être, étreint, étrenne, évident, excellent, exclus, excuse, exécutions, exempte, exemptions, expertise, explicite, explosions, extrait,

fabrique, facture, faille, fane, fanfaronne, fatigue, fauche, fausse, faute, féconde, feignante, feinte, fêle, ferme, ferment, ferraille, fesse, fête, feuillette, feutre, ficelle, fiche, fiente, figure, file, filoche, filtre, finance, fissure, fixe, flèche, flibuste, flingue, flotte, foire, folâtre, fond, force, forge, forme, formule, fouille, fouine, foule, fourche, fourrage, fourrager, fous, fracture, fraise, frange, frappe, fraude, fréquente, frime, fringue, fripons frise, fronce, fronde, frusques, frustre, fugue, fût

gâche, gaffe, gage, gaine, gambade, gangrène, garantie, garde, gardons, gare, gargouille, garrotte, gaule, gave, gaze, geint, gendarme, gêne, gerbe, gerce, germe, gifle, gîte, givre, glace, glaçons, glande, godille, goinfre, gomme, gondole, gorge, gouaille, goupille, gouverne, gourmande, goutte, graisse, gratte, grave, greffe, grêle, genouille, grève, gribouille, griffe, griffons, grille, grimace, grime, grippe, grise, grogne, grognasse, grognonne, grognons, groupe, gueule, guide, guigne, guillotine

hache, hâle, harangue, hâte, hausse, hélas, hérissons, homologue, huile, hypertrophie, hypothèque

idolâtre, illustre, impatiente, importune, impulsions, incendie, incise, incommode, indigne, infecte, infirme, influence, influent, informe, infuse, infusions, injections, innocente, inquiète, inspections, insulte, intègre, intentions, interceptions, interne, interprète, intime, intrigue, invective, inventions, inverse, inversions, invite

jacasse, jalouse, jauge, jetons, jeûne, joue, jouis, juge, jumelle, jurons, jute

lâche, laisse, lambine, lame, lampe, lance, lancer, lançons, lanterne, laque, lardons, largue, lasse, lave, lèche, légitime, lésine, lésions, lessive, leste, leurre, lézarde, libelle, lie, ligature, ligue, lime, limite, lions, liquide, lisse, lit, lithographie, livre, loge, longe, louche, loupe, lustre, lutte, luxe

mâche, machine, macule, maintiens, maîtrise, manie, manifeste, manipule, manoeuvre, manque, maraude, marbre, marchande, marche, marine, marmotte, marque, marre, masque, massacre, masse, mate, matraque, mécontente, médaille, méduse, mélange, menace, ménage, ménager,

mentions, méprise, merde, mérite, mesure, métamorphose, mètre, met, meuble, millésime, mine, minerai, minute, mire, mitraille, modèle, module, monnaie, monologue, monte, montre, moque, moucharde, mouche, mouille, moule, mourons, mousse, moussons, murmure, musarde, mutine

nage, natte, naufrage, neige, niais, niche, nichons, nippe, nombre, note, notions, noue, nourrissons, nuance

objections, oblique, obvie, oeuvre, offense, officier, offre, oignons, ombrage, ombre, opéra, opinions, oppressions, orchestre, orne, orthographe, ouate, oublie, outrage, outre, oxyde, oxygène

pagaie, paie, palabre, panique, panse, parachute, parade, parafe, paraffine, paraphrase, parasite, parcours, parent, paresse, paria, parjure, parodie, parque, part, partage, participe, passe, passions, pastiche, patiente, patine, patronne, patrouille, paume, pavane, pêche, pécher, pédale, peigne, peine, pelote, peluche, pensions, pépie, perce, perche, perle, persécutions, peste, pétarade, peuple, philosophe, photocopie, photographie, piaule, pige, pigeonne, pigeons, pile, pilons, pilote, pince, pinçons, pinte, pioche, pipe, pique, piquette, pirate, pirouette, place, plaisante, planche, plancher, plante, plantons, plaque, plastique, plâtre, plébiscite, plie, plonge, plongeons, plume, plus, poche, point, pointe, poisse, poissons, polémique, police, polissonne, polissons, polycopie, pompe, ponce, porte, portions, pose, poste, potasse, poudre, pousse, pouvoir, pratique, prêche, précise, précisions, préface, prélude, présage, présente, président, presse, pressions, présure, prête, prétexte, prime, prisons, prodigue, profane, professions, programme, progressions, prône, prospections, prospère, psychanalyse, purge, pus,

quadrille, quadruple, querelle, quête, quintuple, quitte,

rabat, râble, racle, radiodiffusions, radiographie, rafle, rage, raie, râle, rallonge, rame, rampe, râpe, rase, rate, rations, ratons, rature, ravage, ravine, rayonne, rayons, rebelle, recensions, recherche, rechute, récidive, réclame, récolte, récompense, recoupe, recours, reculons, redoute, rééditions, réforme, régente, règle, règne, rejetons, relâche, relance, relations, relaxe, relève, remarque, remonte, remorque, rencontre, rengaine, renverse, repère, réplique, réprimande, reprise, reproche, réserve, résident, resquille, resserre, ressort, reste, retouche, retrempe, rétrograde, rêve, réveillons, révisions, révolte, ride, rigole, rime, riposte, rire, risque, rogne, rognons, rosse, rote, roue, rouille, rue, ruine, ruse, rythme

sable, sabre, saccage, sacre, sale, salive, salons, salope, saoule, sape, sauce, saute, sauve, sauvegarde, savoir, savons, scie, sclérose, sèche, seconde, secours, selle, sème, semonce, sens, séquestre, série, serine, serre, siège, signe, singe, solde, sombre, somme, somnolent, sonde, songe, sort, sorte, soude, souffle, souffre, souille, souillons, soupe, souper, sourd, souris, sourire, soutiens, statue, stimule, stipule, strie, structure, subdivisions, sublime, suçons, sucre, suicide, suis, surcharge, surchauffe, suspecte, syncope

table, tache, tâche, tâcherons, taille, taloche, tape, taquine, tasse, tâtons, taxe, télécommande, télécopie, télégraphie, téléphone, téléscope, télévisions, tempête, tenaille, tendre, tendons, tenons, tente, terrasse, terre, tétons, tiens, tiennes, timbre, tique, tire, titre, toile, toilette, toise, toisons, tombe, tonne, toque, torche, torchons, torpille, torsade, tortille, torture, touche, tourmente, trace, traîne, traite, trame, tranche, transfusions, transgressions, traque, traverse, tremble, tremblote, trempe, tresse, triche, triomphe, triple, trompe, trône, trotte, trouble, trousse, truffe, tube, type

ulcère, urine, usine, vaccine, vadrouille, vagabonde, vague, valide, valons, valse, vanne, véhicule, veille, vendange, verse, versions, vibrions, vide, vins, viole, violent, violente, violons, virevolte, vis, visions, visite, visons, vitre, vive, vogue, voie, voile, voisine, voiture, voltige, vote, voûte, voyage, vrille
zèbre, zone.

ANNEXE 6

Locutions comportant un verbe et un substantif homographe susceptible de ne pas être précédé d'un déterminant*

affaire : avoir, faire
aide : chercher, porter
atteinte : porter
bande : faire
barre : avoir
cause : être, faire, prendre
cercle : faire
compte : donner, rendre, tenir
cours : avoir, donner
cure : avoir
date : faire, prendre
échange : avoir, faire
envie : avoir, donner, faire
escorte : faire
fait : prendre
faute : faire
ferme : être, montrer, tenir
fête : faire
figure : faire
fond : faire
forme : avoir, donner, prendre
garde : avoir, prendre
garant : porter
gorge : rendre
goutte : comprendre, voir
hâte : avoir
lit : faire
maille : avoir
marche : faire
masse : faire
mine : faire
nuit : faire
œuvre : faire
ombrage : porter
part : avoir faire, prendre
parti : prendre
partie : être, faire
peine : avoir, faire, prendre
pendant : faire
place : avoir, faire, prendre
porte : tenir
prise : avoir, donner, faire
querelle : avoir, chercher
quitte : être, faire, tenir
recours : avoir
reproche : faire
semblant : faire
signe : donner, faire
table : faire, tenir
tache : faire
terre ; être, toucher
tort : avoir, donner, faire, porter
visite : faire, rendre
vue : avoir

* Pour une discussion : § 5.234 et, pour chaque cas, les éventuels renvois en liste alphabétique. Signalons qu'il manque à cette liste : "payer comptant" qui peut être interprété aussi comme un adjectif utilisé adverbialement.

ANNEXE 7

Pierre HUBERT
Ecole des Mines de Paris
Fontainebleau.

Dominique LABBE
CERAT-Institut d'études politiques
Grenoble

NOTE SUR L'INDICE DE REPARTITION UTILISE DANS L'INDEX DU VOCABULAIRE DE F. MITTERRAND

I. LA NOTION DE "REPARTITION" EN STATISTIQUE LEXICALE

La confection de l'index des interventions radiotélévisées du premier septennat de F. Mitterrand présentait une double difficulté : étant donné son volume, l'index ne peut être reproduit en entier dans un livre. En effet l'indication des numéros d'ordre de tous les textes contenant des occurrences de chacun des vocables aurait donné un tableau de plus de 250 pages. On aurait pu se contenter d'indiquer le nombre d'émissions dans lesquelles le mot a été prononcé mais la taille des textes variant considérablement (le texte le plus long contient presque 40 fois plus de mots que le plus petit…), ce chiffre n'aurait pas eu une grande signification.

Or, à la suite du travail de Gougenheim sur le français fondamental[1], il existe une sorte d'accord tacite pour considérer que la répartition d'un vocable dans un corpus désigne "le nombre de textes (ou de tranches) de ce corpus où ce vocable est attesté"[2]. Ainsi Lafon indique que l'"indice de répartition" "décompte les parties dans lesquelles la forme est présente"[3].

Alphonse Juilland propose de calculer un coefficient de dispersion qui est le rapport de l'écart type à la moyenne arithmétique (coefficient de variation relative)[4]. Ce coefficient est utilisé ensuite pour pondérer la fréquence. Mais cette démarche comporte une sorte de jugement de valeur : à fréquence égale, le vocable employé dans un grand nombre de textes est plus important que celui qui n'apparaît que dans un petit nombre de textes ou de tranches. Alors qu'il semble difficile d'établir une hiérarchie entre eux : l'un appartient au vocabulaire commun à l'ensemble du corpus et l'autre à un vocabulaire spécifique à l'une ou l'autre des parties de ce corpus. Il faut donc trouver un indice qui, à côté de la fréquence du vocable considéré, donne une information sur la distribution de ses occurrences, c'est-à-dire sur sa répartition.

C. Muller a bien vu que la répartition des vocables dans un corpus découpé en tranches, "naturelles" ou non, n'est qu'un cas particulier d'un problème plus général : comment juger de la distribution des apparitions d'un vocable dans un texte ou un ensemble de textes ? C'est pourquoi il propose une définition plus générale de la *répartition* comme étant "la façon dont les occurrences d'un vocable sont réparties dans l'étendue du texte" (ou du corpus)[5].

---

1 Georges GOUGENHEIM, René MICHEA, Aurélien SAUVAGEOT, *L'élaboration du français fondamental*, Paris, Didier, 1967.

2 Gunnel ENGWALL, Vocabulaire du roman français, Stockholm, Sture Allen, 1984, p XXXVIII.

3 Pierre LAFON, *Dépouillements et statistiques en lexicométrie*, Paris-Genève, Slatkine-Champion, 1983, p. 51. Egalement : "Sur la variabilité de la fréquence des formes dans un corpus", *Mots*, 1, octobre 1980, p. 127-165.

4 Alphonse JUILLAND, Dorothy BRODIN, Catherine DAVIDOVITCH, *Frequency Dictionnary of French Words*, La Haye, Mouton, 1970.

5 Charles MULLER, *Principes et méthodes de statistiques lexicale*, Paris, Hachette, 1977, p. 55. Voir également : "Sur les répartitions lexicales" et "La répartition lexicale : problèmes et solutions", *Langue française, linguistique quantitative, informatique*, Genève Paris, Slatkine-Champion, 1985, p. 87-101 et p. 103-113.

A sa suite, nous définirons *la répartition d'un vocable dans un texte comme l'ensemble des emplacements où ce vocable apparaît*, ou encore ses "adresses" comme disent les informaticiens. Quand l'apparition est unique, cette localisation est en elle-même significative et il n'est pas besoin d'information supplémentaire. En revanche, dès qu'un vocable revient en plusieurs endroits d'un texte, une question est soulevée : ce retour est-il régulier et, dans le cas contraire, peut-on considérer que le vocable est caractéristique de tel ou tel passage ?

Nous proposons, dans la seconde partie de cette note, un indice de répartition qui quantifie le degré de régularité d'un vocable à l'intérieur d'un corpus. Nous l'avons appliqué aux vocables les plus fréquents de l'index du vocabulaire de F. Mitterrand[1].

II. CALCUL DE L'INDICE DE REPARTITION.

Soit un texte composé de N mots dont V vocables différents. Considérons un vocable ayant F occurrences dans le texte (fréquence). On associe à ce vocable une dimension caractéristique T égale à l'inverse de sa fréquence relative :

$$T = N/F$$

Puis on calcule les intervalles séparant chaque occurrence du vocable. Soit d_i le nombre de mots séparant les i^e et $(i+1)^e$ occurrences de ce vocable[2]. Les bornes des intervalles étant comprises dans le calcul, la somme de ceux-ci sera :

$$\sum_{i=1}^{i=F} d_i = N$$

Les F intervalles d_i sont classés par tailles croissantes (i variant de 1 à F).

L'indice i est incrémenté de 1 à (k-1) tant que d_i est inférieur ou égal à T. Soit k la valeur de i lorsque d_i devient supérieur à T [3]. L'intervalle d_k contient un certain nombre de fragments de longueur T dans lequel le vocable considéré ne figure pas. Leur nombre est égal à $(d_k - T)$. Il y a $(F - k) + 1$ intervalles[4] où une telle situation est possible. Le nombre total de fragments de longueur T d'où le vocable est absent est égal à :

$$\sum_{i=k}^{i=F} (d_i - T) = \sum_{i=k}^{i=F} d_i - [((F - k) + 1) * T]$$

Soit N' le nombre de segments de longueur T contenant le vocable considéré. On a :

$$N' = N - \left(\sum_{i=k}^{i=F} d_i - [((F - k) + 1) * T] \right)$$

L'indice de répartition repose sur la comparaison de N et de N'.

Il variera entre 0 et 1 :

---

1 Voir également Pierre HUBERT Dominique LABBE, "La répartition des mots dans le vocabulaire présidentiel", *Mots*, 22, mars 1990.

2 Pour la première et de la dernière occurrence, on "boucle le texte" c'est-à-dire que l'on additionne le nombre de mots les séparant respectivement du début et de la fin du texte.

3 Généralement T n'est pas entier. Il convient alors de prendre comme longueur de d_k l'entier immédiatement supérieur à T.

4 Et non pas (F-k) comme nous l'avions écrit dans la première version de ce texte. Nous remercions Monique Becue de nous avoir signalé cette erreur.

- si F=k alors N'=N. Tous les segments possibles de longueur T contiendront une occurrence du vocable. Celui-ci connaît un retour périodique, une régularité parfaite. L'indice prend sa valeur maximum, soit : 1 ;

- si N'=F alors toutes les occurrences sont contiguës et sont contenues dans un intervalle de F mots. Dans ce cas l'indice sera égal à 0.

On en déduit l'indice de répartition :

$$\mathbf{R = \frac{N' - F}{N - F}}$$

Par construction, l'indice est égal à 1 si N'=N, c'est-à-dire dans l'hypothèse d'une répartition du vocable parfaitement régulière. Il est égal à 0 si N'=F, c'est-à-dire lorsque toutes les occurrences du vocable sont contiguës. En pratique, d'après les observations réalisées sur les interventions de F. Mitterrand durant son premier septennat - soit un peu plus de 300 000 vocables -, R semble varier entre 0,8 et 0,1.

III SIGNIFICATION DE L'INDICE DE REPARTITION

On remarquera que l'indice de répartition est une bonne approximation de la *probabilité pour qu'un segment, de T mots contigus, prélevé aléatoirement dans le corpus, contienne le vocable considéré*. Pour le vérifier, nous avons effectué un test sur les propos de F. Mitterrand lors de son face-à-face d'avril 1988 avec J. Chirac. Ce test a porté sur les vocables de fréquence supérieure à 9. Pour chacun d'eux, nous avons prélevé aléatoirement 10.000 tranches de longueur T et nous avons compté le nombre de tranches dans lesquelles apparaissait le vocable considéré. Puis nous avons calculé l'indice R. Les résultats sont donnés dans le tableau ci-dessous pour les vocables les plus fréquents. Le nombre relatif de tranches de taille T prélevées aléatoirement et où la recherche du vocable a été positive (R') se voit associer un intervalle de confiance de deux écarts types.

On peut constater que les résultats du calcul de l'indice tombent généralement dans cet intervalle de confiance. Il est possible d'en conclure que le modèle donne la description assez exacte du phénomène estimé grâce à nos tirages aléatoires. C'est-à-dire la probabilité de rencontre du vocable étudié dans une tranche quelconque de T mots prélevée dans le corpus.

L'indice s'approche d'autant plus de cette probabilité que F sera très petit par rapport à N et N'. Dans ce cas, on peut écrire :

$$R = \frac{N' - F}{N - F} \ \# \ \frac{N'}{N} \ \text{(avec N>>F)}$$

En effet, le rapport N'/N est le nombre de cas favorables (rencontre du vocable dans un segment quelconque de longueur T) sur le nombre total de cas possibles (autant de combinaisons possibles que de mots contenus dans le texte). Ce rapport tend vers une limite inférieure égale à F/N (dans le cas d'occurrences contiguës du vocable). La comparaison n'est donc possible qu'entre les vocables de même fréquence alors que R ne présente pas cet inconvénient.

L'indice sert à resserrer la présentation de l'index (une seule ligne par vocable) sans perdre totalement l'information concernant la localisation des occurrences. De plus, il permet d'isoler les vocables les plus réguliers et les plus irréguliers, ce qui ajoute à la fréquence une indication dont des études ultérieures devront déterminer la portée et l'intérêt exacts.

Etude de la répartition des vocables les plus fréquents dans les interventions de F. Mitterrand face à J. Chirac (avril 1988). Calcul sur des tranches de texte prélevées aléatoirement (R') puis simulation à l'aide du modèle de partition (R)

	fréquence	Echantillons aléatoires		Indice de répartition
	F	R'	2 écarts types	R
le (article)	852	0,673	0,016	0,670
de	561	0,680	0,018	0,684
être (verbe)	349	0,654	0,016	0,650
avoir (verbe)	338	0,653	0,014	0,643
je	299	0,553	0,018	0,556
à	210	0,684	0,017	0,682
et	180	0,665	0,016	0,661
il	173	0,595	0,013	0,593
que (conj.)	159	0,583	0,012	0,595
vous	157	0,498	0,019	0,497
ne	144	0,607	0,012	0,601
un (article)	144	0,612	0,022	0,607
ce (pronom)	143	0,609	0,017	0,611
pas (adverbe)	126	0,595	0,013	0,607
qui	123	0,615	0,017	0,609
ce (article)	89	0,659	0,016	0,657
en (préposition)	86	0,618	0,014	0,616
dans	78	0,615	0,010	0,618
le (pronom)	76	0,662	0,017	0,655
dire	73	0,671	0,017	0,687
cent	72	0,427	0,015	0,429
pour	70	0,579	0,015	0,584
que (pronom)	66	0,644	0,016	0,647
se	62	0,600	0,017	0,602
nous	57	0,590	0,015	0,576
faire	56	0,576	0,015	0,574
y	55	0,568	0,013	0,569
mille	53	0,411	0,014	0,413
falloir	52	0,476	0,015	0,469
quatre	52	0,503	0,009	0,513
on	51	0,541	0,021	0,538
vouloir	50	0,651	0,013	0,650
vingt	49	0,524	0,015	0,524
neuf (numéral)	46	0,495	0,017	0,497
plus	46	0,629	0,016	0,628
mais	45	0,615	0,016	0,613
par	45	0,569	0,018	0,573
bien (adverbe)	44	0,624	0,019	0,620
monsieur	43	0,573	0,016	0,561
premier	41	0,525	0,013	0,526
avec	39	0,611	0,020	0,607
sur	38	0,537	0,018	0,542
ministre	36	0,495	0,015	0,497
pouvoir (verbe)	36	0,641	0,016	0,638
tout (déterminant)	35	0,647	0,015	0,639
là	34	0,673	0,013	0,678
ils	32	0,444	0,016	0,450
très	32	0,626	0,013	0,634
cela	31	0,597	0,011	0,590
si (conj.)	31	0,548	0,010	0,549
en (pronom)	30	0,615	0,014	0,614

ANNEXE 8
LES PRINCIPALES HOMOGRAPHIES
(CLASSEMENT ALPHABETIQUE)

Les numéros renvoient :
- premier chiffre : chapitre
- deuxième chiffe : section
- troisième chiffre : sous-section
- quatrième chiffre : paragraphe

Acquis 5.222, 5.231, 5.34, 6.32 {substantif-adjectif}-{verbe (acquérir, passé simple et participe passé)}
Affaire 5.234, 5.311 substantif-verbe (affairer)
Aide 6.124, 5.311 {substantif masc et féminin}-verbe (aider).
Ailleurs 7.132
Air 6.123 (atmosphère-aspect, physionomie)
Allié 5.222, 6.32, {adj-substif}-{verbes (allier, aller)}
Après 7.224 (préposition-adverbe de temps)
Arrière 7.132 (substantif-adverbe)
Attendu 5.223, 7.322 {substantif masculin- verbe (attendre)}-préposition
Au 7.326 (préposition-déterminant)
Aucun (e) 6.431, 6.573 (déterminant-pronom)
Autre (s) 6.431, 6.573 {adjectif-déterminant}-pronom
Avant 6.331, 7.322 (substantif masculin)-{préposition-adverbe de temps ou de lieu}
Bas 6.233, 6.32, 7.131, 7.132 {adjectif-substantif}-adverbe
Beau 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Bien 6.32, 7.133 {adverbe-préposition-interjection}-{adjectif-substantif}
Bon 6.32, 7.131, 7.132 {adverbe-interjection}-{substantif-adjectif}
Bref 7.131 {abverbe-interjection}-adjectif
C'est-à-dire {locution figée} 2.231
Calme 5.311, 6.32{adjectif-substantif}-verbe (calmer)
Car 7.233 (substantif masculin)-conjonction
Ce 6.532 (déterminant-pronom démonstratif)
Certain (s,e,es) 6.431, 6.573 {déterminant-adjectif}-pronom
Cesse 5.311 (substantif féminin-verbe)
Cher 7.131 (adverbe-adjectif)
Clair 6.233, 6.32, 7.131 {adjectif-substantif}-adverbe
Compte(s) 5.234, 5.311 (substantif masculin)-verbe (compter)
Contre 5.311, 7.321, 7.322 {adverbe-préposition}-{substantif-verbe (contrer)}
Cours 5.31, 6.124, 6.125 {substantif féminin-substantif masculin}-verbe (courir)
Court 5.31, 6.32 {substantif masculin-adjectif}-verbe (courir)
D'abord 2.134, 7113 (substantif-locution adverbiale)
D'accord 2.134, 7113 (substantif-locution adverbiale)
D'ailleurs 7113 (adverbe-substantif-locution adverbiale)
D'emblée 2.134 (locution adverbiale)
De 7.325 préposition-déterminant
Dehors 7.132 adverbe-substantif
Demi 6.233, 6.32, 7.131 7.132 {adjectif-substantif masculin}-adverbe

Derrière 7.323 (préposition-substantif masculin)
Désert 6.32 (adjectif-substantif masculin)
Déserte 5.311 adjectif (désert)-verbe (déserter)
Dessous 7.132 adverbe-substantif
Dessus 7.321, 7.323 {préposition-adverbe}-substantif
Devant 5.334, 6.331, 7.322 {préposition-adverbe}-{verbe (devoir)-substantif}
Devoir 5.32 substantif-verbe (devoir)
Différent(s) 6.433 (déterminant-adjectif)
Dire 5.32 substantif masculin-verbe (dire)
Divers(es) 6.433 (déterminant-adjectif)
Doute(s) 5.311 substantif masculin-verbe (douter)
Droit 6.233, 6.32, 7.131 {adjectif-substantif masculin}-adverbe
Droite 6.32 (adjectif-substantif féminin)
Dur 6.233, 6.32, 7.131 {adjectif-substantif masculin}-adverbe
Dure 5.311, 6.32 {adjectif-substantif féminin}-verbe (durer).
Durant 5.334, 7.322 {préposition-adverbe}-verbe (devoir)
Echéant (le cas) 5.334 adjectif-verbe (échoir)
En 7.324 (préposition-pronom)
Enfant 6.127 (substantif masculin et féminin)
Ensemble 6.331, 7.132 (adverbe-substantif masculin)
Entre 7.322 préposition-verbe (entrer)
Entreprise 5.233, 5.34 {substantif-adjectif}-verbe (entreprendre).
Envers 7.323 (préposition-substantif masculin)
Etat-état 1.22 (substantifs homographes)
Etudiant 5.232, 5.33 {substantif masculin-adjectif}-verbe (étudier)
Excepté 5.242,7.322 {adjectif-préposition}-verbe (supposer, participe passé)
Fait 5.231, 5.34, 6.32 {substantif-adjectif}-{verbe (faire, passé simple et participe passé)}
Faut 5.221 verbe(falloir)-verbe (faillir)
Faux 6.233, 6.32, 7.131 {substantif-adjectif}-adverbe
Ferme 5.311, 6.32{adjectif-substantif}-verbe (fermer)
Fier 5.31, 6.32{adjectif-substantif}-verbe (fier)
Fils 6.125 (substantif-substantif)
Fonds 6.126 (substantif-substantif)
Force 7.323 {substantif-adverbe}-verbe
Fort 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Froid 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Garde(s) 5.311, 6.124 {substantif masculin-substantif féminin}-verbe (garder)
Grand 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Haut 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Immigré 5.233 5.341 {adjectif-substantif}-verbe (immigrer, participe passé)
Juste 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Le 6.572 (déterminant-pronom)
Leur 6.572 (déterminant-pronom)
Livre 5.311, 6.124 {substantif masculin-substantif féminin}-verbe (garder)
Lutte 5.311, 6.32{adjectif-substantif}-verbe (lutter)
Maintenant 7.135 adverbe-verbe (maintenir au participe présent)
Mal 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Mauvais 6.233, 6.32, 7.131 (adjectif-adverbe)
Manche 6.124 (substantif masculin-substantif féminin)
Manifeste 5.311, 6.32{adjectif-substantif}-verbe (manifester)
Même 6.573 {déterminant-pronom}-adverbe

Mise 5.221 {substantif masc-féminin}-{verbe-verbe (mettre, miser)}
Mode 6.127 (substantif masculin- substantif féminin)
Mort 5.233, 5.34, 6.124 {(substantif masculin-substantif féminin)-adjectif}-verbe (mourrir)
Naturel 6.32 (adjectif-substantif masculin)
Net 6.233, 6.32, 7.131 (adjectif-adverbe)
Neuf 6.433 adjectif-déterminant (numéral)
Nul 6.431, 6.573 (déterminant-pronom)
Office 6.124 (substantif masculin-substantif féminin)
Or 7.233 (substantif masculin)-conjonction
Où 7.136 (pronom-adv)
Outre 7.323 (substantif féminin-préposition)
Parer 5.221 (verbe-verbe)
Pas 7.134 (substantif-adverbe)
Passager 6.32 (substantif masculin-adjectif)
Penchant 5.334, 6.331, 7.322 {verbe (pencher)-substantif}
Pendant 5.334, 6.331, 7.322 {préposition-adverbe}-{verbe (pendre)-substantif}
Personne 6.574 (substantif-pronom)
Petit 6.233, 6.32, 7.131 {adjectif-substantif}-adverbe
Plein 6.32, 7.323 {substantif-adjectif}-préposition
Plus 7.135, 7.321{adverbe-préposition}-verbe (plaire)
Point 7.134 (substantif masculin-adverbe)
Port 6.123 (d'un objet ou de tête,…de mer)
Porte 5.311, 6.124 substantif-verbe et préfixe
Poste(s) 5.311, 6.124 {substantif masculin-substantif féminin}-verbe (poster)
Pouvoir 5.32 substantif masculin-verbe (pouvoir)
Pourvu 5.233, 5.242, 7.322 {adjectif-préposition}-verbe (pourvoir, participe passé)
Premier 6.432 {adjectif-substantif masculin}-déterminant
Première 6.432 {adjectif-substantif féminin}-déterminant
Proche 6.32, 7.323 (substantif masculin-préposition)
Puis 7.221 conjonction-verbe (pouvoir)
Quand 7.224 (conjonction-préposition)
Que 7.222 {conjonction-adverbe}-pronom
Quel (quels, quelle, quelles) 6.56 (pronom-déterminant)
Quelque 6.573 (déterminant-pronom)
Quitte 7.135
Reprise 5.221 {substantif masc-féminin}-{verbe-verbe (reprise, reprendre)}
Rien 6.574 (pronom-substantif masculin)
S' 7.223, 7.224 {conjonction-adverbe}-pronom réfléchi
Sauf 7.322 (adjectif-préposition)
Savoir 5.32 substantif masculin-verbe (savoir)
Second 6.432 {adjectif-substantif masculin}-déterminant
Seconde 5.311, 6.432 {adjectif-déterminant}-verbe (seconder)
Si 7.223, 7.224 {conjonction-adverbe}-substantif
Soit 7.211 conjonction-verbe (être première personne du subjonctif présent)
Somme 5.311, 6.124 {substantif masculin-féminin}-verbe (sommer)
Son 6.432 (déterminant-substantif)
Sous 6.125, 7.323 (substantif masculin singulier et pluriel)-préposition
Suis 5.235 verbe (être)-verbe(suivre)
Suivant 5.334, 7.322 {(substantif-adjectif)-préposition}-verbe (suivre, participe présent)
Supposé 5.223, 5.242 7.322 {adjectif-préposition}-verbe (supposer, participe passé)
Tel 6.573 (déterminant-pronom)

Tiens, tiennes 6.55 pronom-verbe (tenir)
Ton 6.432 (déterminant-substantif)
Tout (e,s,es) 7.137 {déterminant-pronom-adverbe}-substantif
Tu 6.532 pronom-verbe (taire participe passé)
Un (e,es,s) 6.413, 6.572 (déterminant-pronom)
Vers : 6.125, 7.323 {substantif masculin singulier et pluriel}-préposition.
Vis 5.221, 5.31 {verbes(voir)-verbe (vivre)}-substantif féminin
Voile 5.311, 6.124 {substantif masculin-substantif féminin}-verbe (voiler)
Vrai 6.233, 6.32, 7.131 {adjectif- substantif}-adverbe
Vu 7.322 préposition-verbe (voir)
Vue 5.233, 5.34 {substantif féminin-adjectif}-verbe (voir, participe passé)
Y 7.136 (adverbe-pronom)