Désambigüisation automatique d'homographes verbe / nom

Marc Hug

19, rue Oberlin – 67000 Strasbourg – France

(Professeur émérite à l'Université Marc Bloch, rue Descartes, 67000 Strasbourg)

Abstract

Disambiguating such ambiguous French types as *poste, alarme, laisse* etc., which can be interpreted in turn as noun forms or as verbal forms, is used as a touchstone for the value of an existing program that "categorizes", i.e. classifies into morpho-syntactical classes, each word of any text. It appears that even if one distinguishes between "sure" and "unsure" categorizations, the performances of such a program unescapably vary according to the individual word concerned and its specific properties. Consequently an automatic "categorization" should take into account a number of ad-hoc properties of the units to analyze in order to improve the results.

Résumé

La désambigüisation de formes ambigües telles que *poste, alarme, laisse* etc., qui peuvent recevoir soit une analyse de nom, soit une analyse de forme verbale, a été utilisée comme pierre de touche d'un programme de "catégorisation" informatique des formes d'un texte. Il apparait que même si l'on prend soin de distinguer les catégorisations "certaines" des catégorisations reconnues "incertaines", le rendement d'un tel programme ne peut manquer d'être très sensible à l'identité et aux caractères idiosyncratiques de l'unité concernée, ce qui devrait imposer la prise en compte d'un certain nombre de caractères "ad-hoc" de cette unité dans le traitement de ses occurrences.

Mots-clés : ambigüité lexicale, analyse automatique, analyseur, classes morpho-syntaxiques.

1. Introduction

La lemmatisation peut être une des étapes de l'indexation morpho-syntaxique des unités d'un texte, si toutefois une telle indexation est jugée nécessaire. Je pense, pour ma part, qu'avant de l'écarter, il est bon d'en mesurer certes le cout, mais aussi l'intérêt.

On se souvient de la controverse qui a opposé il y a déjà une vingtaine d'années Charles Muller à Pierre Lafon et à quelques autres à propos de cette question. Je crois qu'il faut se garder d'opter pour un travail sans lemmatisation simplement parce que c'est plus rapide, et de justifier après coup le procédé.

L'un des problèmes qui se posent si l'on veut opérer une lemmatisation de manière aussi correcte que possible, c'est de résoudre l'homographie d'un grand nombre de mots, en particulier ceux qui peuvent être à la fois des formes verbales conjuguées et des noms ou adjectifs. Parmi les mots présentant ce type d'ambigüité, je ne m'intéresserai ici qu'à la sous-classe de ceux qui se terminent dans la tradition écrite par un *–e* et dont l'interprétation verbale entre dans la conjugaison d'un verbe du premier groupe. Ces formes sont potentiellement au nombre de plus de 1250, avec des proportions des plus variables entre l'emploi de verbe et l'emploi de nom ou d'adjectif. On pourrait y ajouter des formes telles que *vis, as, est, fond, fait, part* etc., et quelques formes en *–e* de verbes du 3e groupe comme *faille,*

voie, sorte, bâtisse, offre etc. Je me limiterai en fait à des cas où l'interprétation d'adjectif n'entre pas en ligne de compte, en négligeant donc le cas de formes tells que *meuble, trouble, alerte, manifeste* etc.

2. Problématique générale de la lemmatisation

2.1. Intérêt de l'opération

L'intérêt de cette désambigüisation peut être variable selon les unités dont il s'agit et selon le but qu'on poursuit. Si le but du chercheur est purement linguistique, la désambigüisation est forcément intéressante. Si c'est au contraire l'analyse du contenu sémantique qui l'intéresse, son intérêt est variable : pour les types *cause*, *analyse*, *bagarre*, *contraste* ou *couronne*, on pourra se désintéresser de la question de savoir si ce sont des occurrences de la forme verbale ou du nom, puisque le contenu sémantique des deux mots est proche ; il sera cependant intéressant de regrouper les fréquences des formes fléchies d'une même unité du lexique. Mais lorsqu'il s'agit de formes telles que *laisse*, *ferme* ou *bouche*, le contenu sémantique est radicalement différent entre la forme verbale et le nom, et la distinction parait devoir aller de soi. Il paraitrait paradoxal qu'une recherche de nature sémantique ou thématique se désintéresse d'oppositions de sens aussi radicales. Entre les deux situations, celle de *bagarre* ou *analyse* d'un côté, celle de *bouche* de l'autre, impossible de tracer une limite nette. Les avis peuvent diverger lorsqu'on se demande s'il y a en français moderne un rapport de sens clairement perçu entre *la branche* et *brancher*, entre *une alerte, un vieillard alerte* et *alerter*, entre un *libelle* et *libeller*, entre *un lustre* et *lustrer* etc. Il s'y ajoute bien entendu dans certains des noms ou des verbes une polysémie qui peut être interprétée en termes d'homonymie. Je ne veux pas entrer ici dans cette discussion.

2.2. Méthodes : les deux pôles envisageables

Pour désambigüiser les occurrences de ces formes dans un corpus textuel, on peut procéder soit par intuition, en utilisant sa compétence de locuteur, soit par un programme informatique, soit en combinant les deux méthodes, la première étant plus sûre, mais demandant énormément de temps, la seconde étant très rapide, mais comportant nécessairement une proportion plus ou moins importante d'analyses erronées.

De manière générale, on peut postuler que plus un corpus est étendu, plus la part de la logique informatique doit être grande, et plus par conséquent il est nécessaire d'accepter une part d'analyses erronées. Cette question a été évoquée entre autres par D. Labbé : dans la classification morpho-syntaxique des unités, peut-on accepter qu'il y ait des erreurs, et si oui, quelles proportions d'analyses erronées considèrera-t-on comme acceptables ? Alors que D. Labbé souhaite arriver à un résultat dépourvu de toute erreur, ce qui implique nécessairement l'intervention d'un contrôle intuitif, l'équipe d'informaticiens de FRANTEXT a opté pour une procédure totalement automatisée, ce qui est compréhensible au vu des dimensions du corpus à traiter ; ce qui est plus surprenant, c'est que cette équipe défend explicitement le point de vue suivant lequel le rendement d'un logiciel de catégorisation peut être défendable même s'il produit neuf fois sur dix – dans un cas particulier – une analyse fausse. L'idée plus ou moins explicite qui est développée dans leur système d'aide en ligne est que la proportion d'erreurs importe beaucoup moins que la présence effective, dans les sorties, des occurrences recherchées. Cela présuppose toutefois que ce ne soit pas précisément l'indication quantitative qui soit au centre de l'intérêt du chercheur. Il y a de toute façon là un paralogisme, car on pourrait prétendre de façon tout aussi plausible, avec des a-priori

différents, que peu importe qu'il manque certaines des occurrences recherchées, pourvu que toutes celles qui sont proposées soient correctes. Il faut se méfier de ce type de raisonnements.

3. Le corpus extrait de FRANTEXT

3.1. Nature des données

Les formes examinées ont été sélectionnées dans la liste supposée complète que proposait une étudiante de maitrise, Mme Laurence Philipps[i]. Plutôt que d'extraire de ce corpus un échantillon représentatif, j'ai préféré opérer sur un petit nombre d'unités librement choisies, en analysant la totalité des occurrences. Les proportions indiquées ne pourront donc pas être considérées comme représentatives de l'ensemble du corpus. J'espère toutefois que les problèmes soulevés sont représentatifs, eux, de ceux que soulève le corpus dans son ensemble. Ce corpus est celui de la totalité des textes présents dans la base FRANTEXT catégorisée et datés de 1960 et après. Dans cette base dite catégorisée, chaque mot ou "entité" composée de plus d'un mot a reçu une étiquette de classe morpho-syntaxique ; cette étiquette est considérée par le logiciel lui-même tantôt comme "certaine", tantôt comme "incertaine". J'ai analysé intuitivement les occurrences des formes du tableau 1 avec l'aide d'un petit programme informatique.

Je me suis intéressé aux variations auxquelles sont sujettes les proportions suivantes :

- proportions d'occurrences, pour une forme donnée, de l'emploi nominal et de l'emploi verbal ;

- proportions d'occurrences que le logiciel a catégorisées comme verbe ou comme substantif, ce qui n'est pas la même chose ; on pourra tirer de là les proportions d'erreurs ;

- proportions de formes dont les catégorisations sont considérées comme "certaines" ou comme "incertaines", et proportions corespondantes d'erreurs de catégorisation.

- proportion d'occurrences qui peuvent être désambigüisées de façon à peu près infaillible par un certain nombre de contextes aisés à définir ; cette proportion ne se confond pas avec celle des catégorisations "certaines".

Il doit être entendu que dans tout ce développement, le but poursuivi n'est pas de faire le procès ou l'apologie de la logique de catégorisation mise en œuvre dans FRANTEXT. L'intérêt évident de ce programme de catégorisation est qu'il a été appliqué à un corpus tellement important qu'on ne peut pas en soupçonner la logique d'avoir été taillée sur mesure pour un corpus particulier. Je ne m'interdirai pas toutefois d'apprécier sur certains points les performances de cette logique.

3.2. Proportions de noms et de verbes

Le tableau 1 montre comment se présentent les proportions de substantifs et de verbes parmi les occurrences d'une dizaine de formes ambigües.

---

i. En fait, la notion même de complétude est dans un tel cas illusoire, tout autant que peut l'être toute idée d'une liste "complète" du lexique d'une langue. Il m'est arrivé par hasard de rencontrer en cours de route une ou deux occurrences d'un verbe *suiter* que les dictionnaires ne donnent pas en général. De ce fait, la forme *suite* pourrait être ajoutée à la liste de Mme Philipps.

Tableau 1

	"Certains"		"Incertains"		Total Liste		Total réel	
	Subst	Verbes	Subst	Verbes	Subst	Verbes	Subst	Verbes
aide	1843	255	73	24	1916	279	1869	326
cause	1451	140	60	6	1511	146	1493	163
cesse	1118	397	112	139	1230	536	1167	599
couronne	359	21	33	7	392	28	391	29
garde	1892	542	264	67	2156	609	2020	739
laisse	180	1947	321	182	501	2129	116	2513
oublie	16	520	15	39	31	559	0	590
rayonne	27	14	12	15	39	29	34	34
réclame	79	134	40	38	119	172	91	200
taille	1399	49	79	8	1478	57	1476	59
Total	8364	4019	1009	525	9373	4544	8657	5252

N.B. Les six premières colonnes indiquent les effectifs fournis par le logiciel d'interrogation de FRANTEXT ; les deux dernières remplacent les totaux par les totaux observés après analyse intuitive. Les effectifs ne sont pas strictement identiques, pour les raisons suivantes. Une des occurrences de *cause* est en fait la forme brève de l'anglais *because*, et se trouve dans une citation en anglais. Parmi les occurrences de *garde*, il y en a d'une part dix qui ne sont pas à ranger dans les deux classes prévues (graphie *r'garde* pour *regarde*, par exemple, ou *lac de Garde*), et d'autre part on a intégré trois occurrences que le logiciel avait rangé fautivement dans d'autres classes que celles de verbe ou de substantif. De même une occurrence de *laisse* a été omise (simple graphie "phonétique" d'un nom, ininterprétable).

Les proportions de noms et de verbes sont très variables. Comme le choix des types n'est cependant pas un tirage au hasard, rien ne dit que les proportions globales, par exemple celles qu'on trouve à la ligne "Total", soient représentatives de ce qu'on trouverait dans l'ensemble des formes ambigües.

Ce qu'on remarque, c'est que les proportions de substantifs sont assez systématiquement surévaluées par le logiciel de catégorisation, mais dans une mesure variable d'une unité à l'autre. On ne peut même pas dire que ce soit la difficulté syntaxique qui soit à l'origine de ces variations, car le mot *taille* a quelques emplois délicats à traiter, et pourtant le nombre de substantifs indiqué par le logiciel est presque identique à l'effectif réel. En revanche le nombre d'erreurs de catégorisation est inversement proportionnel à la part des substantifs : lorsque l'essentiel des occurrences est constitué par des formes verbales, il y a beaucoup d'erreurs (c'est le cas de *laisse*) ; lorsqu'il y a essentiellement des occurrences du nom, il y en a peu ; prenons les deux grandeurs suivantes :

(a) la proportion de substantifs effectivement constatée parmi l'ensemble des occurrences de la forme ambigüe ;

(b) parmi les occurrences considérées par le logiciel comme des substantifs, la proportion de ceux qui en sont réellement.

Ces deux grandeurs sont indépendantes l'une de l'autre du point de vue strictement arithmétique. Sur les dix formes ambigües du tableau 1, la corrélation entre ces deux

grandeurs est de 0,914 pour 8 degrés de liberté : plus il y a de substantifs en proportion, moins on trouve d'erreurs de catégorisation.

En réalité on n'a travaillé que sur le nombre de substantifs indiqué par la liste et le nombre réellement trouvé ; s'il y avait autant d'erreurs dans un sens que dans l'autre, les effectifs globaux pourraient être corrects malgré un grand nombre d'erreurs de catégorisation. En fait, les occurrences considérées à tort comme des substantifs sont nombreuses, celles qui sont considérées à tort comme des verbes sont très rares ; c'est pourquoi le test évoqué peut effectivement avoir une signification.

Tableau 2. Récapitation des erreurs de catégorisation

Forme ambiguë	Substantif analysé comme verbe	Verbe analysé comme substantif
aide	3	50
cause	20	143
cesse	3	66
couronne	2	3
garde	8	132
laisse	0	384
oublie	0	31
rayonne	3	8
réclame	8	36
taille	6	8

Mais regardons le tableau 2, qui nous indique combien de fois un substantif a été à tort catégorisé comme verbe, et combien de fois un verbe a été catégorisé comme nom. Nous voyons que les situations changent considérablement d'une unité à l'autre, même si c'est toujours la deuxième colonne qui offre l'effectif le plus important. La forme *taille*, un des types très fréquents, occasionne nettement moins d'erreurs que la forme *réclame*, qui est une des moins fréquentes de la liste. Pourtant ces deux mots offrent des particularités syntaxiques assez analogues au premier regard, essentiellement les emplois en juxtaposition après un autre nom, comme dans *ballon-réclame* ou *cendrier-réclame* dans un cas, comme *une jupe taille 42* dans l'autre. Mais il se trouve que malgré sa grande polysémie, *taille* s'emploie de façon très prédominante dans un ensemble de tournures faciles à repérer, par exemple après préposition, ou après une suite préposition + adjectif ou préposition + article.

De tout cet ensemble d'informations, on tire l'impression que des paramètres ad-hoc demandent à être pris en compte pour certaines au moins de ces unités.

3.3. Catégorisations "certaines" et "incertaines"

C'est ici que se révèle la plus grande faiblesse du système d'analyse appliqué par FRANTEXT, car l'utilisateur non prévenu pourrait croire que la catégorisation dite certaine est réellement fiable ; or il n'en est rien. En ce qui concerne les catégorisations de verbe, les erreurs sont suffisamment rares pour qu'on puisse considérer le résultat comme bon en ce sens que les formes étiquetées comme verbes en sont réellement. Dans le cas des noms au contraire, il peut arriver que les formes verbales soient plus nombreuses que les emplois de noms, même parmi ce qui est étiqueté "Substantif certain". La forme *oublie* est considérée 31 fois comme le substantif, dont 16 fois comme "certain", et il n'y a pas une seule occurrence du substantif dans le corpus ; ce n'est pas le seul cas de ce type, même si c'est tout de même l'exception. Le mot *laisse* est analysé 501 fois comme un nom, alors que 116 emplois seulement de nom se rencontrent effectivement. Bien sûr c'est surtout dans les noms dits "incertains" que cela se manifeste : sur 321 occurrences de *laisse* "substantif incertain", on ne trouve plus en fait que 2 emplois de nom, le reste étant constitué par des formes verbales.

Mais même parmi les mots *laisse* classés "Substantifs certains", un peu plus du tiers sont en fait des verbes.

Le plus surprenant est cependant que la répartition entre catégorisations dites "certaines" et catégorisations reconnues" incertaines" ne semble pas suivre une logique claire. Les choses se présentent comme si les principes retenus n'avaient pas de base linguistique solide et comme si certaines données élémentaires avaient été négligées.

3.4. Principes à respecter

Des erreurs relevées, on peut extraire quelques principes d'analyse utiles.

1. Pour analyser les mots d'un corpus aussi énorme que FRANTEXT, on ne peut raisonnablement envisager qu'une grammaire locale, en acceptant qu'il y ait un certain nombre d'erreurs, si du moins on ne dispose pas des moyens matériels et humains d'un contrôle intuitif. C'est ce qui a été fait dans FRANTEXT, si bien que, par exemple, l'unité ambiguë est classée "Substantif" lorsqu'elle est immédiatement précédée d'une préposition ou d'un déterminant, alors qu'elle est analysée comme "Verbe" lorsqu'elle est précédée de *ne* ou d'un pronom clitique.

2. Il est nécessaire de prendre en compte de manière raisonnée les principes appliqués lors de la saisie des textes. Ces principes, personne ne les connait mieux que les informaticiens auteurs du programme de catégorisation. Or il y a là des lacunes. En voici deux dont l'impact quantitatif n'est pas énorme, mais qu'il aurait été facile de combler. La première concerne le tiret. Comme dans la plupart des saisies informatiques, le tiret et le trait d'union n'ont pas été distingués. Voici un cas où cela occasionne une erreur :

(a) {P325/GURVITCH.G/TRAITE DE SOCIOLOGIE T.1/1967}

42 P325,91-92 $ est d'une simplicité autoritaire, linéaire. *Halbwachs aussi qui classe, une fois pour toutes. *Marcel *Mauss est plus divers, mais nous ne le lisons guère-et pour <cause>-nous entendons sa pensée, répercutée par ses disciples et qui se mêle, ainsi, vivante au droit fil de la recherche actuelle.

Ici la présence de *pour* avant *cause* devait faire classer ce dernier mot comme "Substantif", alors que la séquence *-nous* qui suit devait le faire classer comme verbe ; compte tenu du fait que les traits d'union ne sont pas distingués des tirets, le premier critère devait être prioritaire. Cela n'a pas été le cas, d'où la classification erronée de "Verbe", qualifiée de "certaine", de surcroit. Ce type d'erreurs est particulièrement choquant pour l'utilisateur humain lorsque la suite de mots se trouvant de part et d'autre du tiret ne peut pas constituer une suite verbe + pronom, comme dans l'exemple suivant :

{R727/BAYON/LE LYCEEN/1987}

R727,132 $ Dominique prend un bref élan, pousse une pointe et, «Yaaaah», franchit la barre critique des 1,30 mètre - pas loin de sa <taille> - en saut périlleux. Incroyable. Au sol. Jamais vu...

Dans cette occurrence, où *taille* est considéré comme "verbe, certain", c'est parce qu'il est suivi de *-en*, alors qu'il y aurait *tailles-en* si on avait affaire au verbe ; là aussi, le possessif qui précède aurait dû être prioritaire.

Des erreurs ont été commises aussi avec les guillemets.

{S232/ROY.C/LA TRAVERSEE DU PONT DES ARTS/1979}

35 S232,117-118 $ proche de celle qu'il avait essayé de suggérer avec des sons, de traduire en musique dans ses Trois mouvements. Quand Charles disait que le temps intérieur le plus profond ne «

<coule> » pas, mais ressemble à une eau dormante, qu'aucun courant n'oriente, qu'aucune direction n'infléchit, Audoin répondait que […]

Comme tous les caractères non alphabétiques, le guillemet ouvrant ou fermant est considéré comme un "mot", ce qui a empêché de voir ici le *ne* qui précédait, et a fait considérer *coule* comme un nom ; or si les guillemets peuvent conduire à traiter comme un nom ce qu'ils entourent, c'est loin d'être leur seule fonction. Après un *ne,* il était évident qu'il devait y avoir un verbe.

3. Le traitement des locutions plus ou moins lexicalisées demande une attention particulière. Il est entendu que toute liste de locutions lexicalisées établie a priori peut donner lieu à contestation. Dans FRANTEXT, la locution *à cause de* a été codée globalement comme une seule unité, alors que, par exemple, *sans cesse* ne l'a pas été. On aurait pu faire l'inverse. Les deux locutions ont des fréquences comparables, mais *à cause de* est assez couramment dissocié, par exemple en *à cause en particulier de…, à cause aussi de, à cause même de…, à cause sans doute de…, à cause surtout de…* etc., alors que les variantes de *sans cesse* sont rarissimes (*sans trêve ni cesse, sans fin ni cesse*). Mais la principale critique qu'il y a lieu de faire, c'est qu'il faudrait éviter la dichotomie brutale entre d'une part les locutions lexicalisées, qui sont codées en bloc, et les autres locutions, qui sont soumises aux règles générales. Il existe une foule de locutions très courantes, aisées à repérer, mais dont le figement n'est pas suffisant pour justifier un codage global ; le repérage ad-hoc de ces locutions permettrait de coder convenablement les unités dont elles se composent ; ainsi, pour en rester à la forme *cause*, des locutions verbales comme *mettre en cause* ou *remettre en cause* ; hors de ces locutions ou d'autres semblables, la séquence *en cause* ne peut pas être d'emblée considérée comme une suite préposition + nom, puisqu'on trouve *on en cause*, avec *en* pronom et *cause* verbe (il y a pas mal d'erreurs sur ce point dans FRANTEXT). Un traitement analogue pourrait concerner des locutions telles que *venir en aide à, mettre en garde, prendre garde à, monter la garde, garde à vous*, etc. ou bien *les* N *en cause*. Donc : trouver un traitement intermédiaire entre la locution considérée comme figée et la syntaxe considérée comme totalement libre. J'ai volontairement laissé de côté ici les locutions du type de *être de taille à, à l'aide de,* dans lesquelles la présence de la préposition permet la catégorisation "substantif'" sans risque d'erreur.

4. La distinction entre analyses "certaines" et analyses "incertaines" demande à être soigneusement fondée ; en fait, il n'y a que peu de règles de grammaire locale qui soient d'une fiabilité absolue, mais il y en a qui sont presque totalement fiables, et ce sont celles-là seulement dont l'application devrait être labélisée "certaine". On a évoqué précédemment les cas des formes *laisse* et *oublie*. Il faudrait se contenter d'afficher le degré de fiabilité qu'on peut effectivement garantir, quitte à annoncer des résultats moins brillants.

On l'a vu, les analyses dites "certaines" de verbe sont presque toujours correctes, alors que celles de nom le sont moins souvent. L'identification certaine du nom est effectivement plus difficile par grammaire locale que celle du verbe ; mais il faudrait alors accepter que la proportion des étiquettes "certain" soit plus faible d'autant.

5. On devrait prendre en compte les caractères connus des unités analysées, en particulier le genre et le nombre des noms. Cela présuppose évidemment que le texte analysé soit dépourvu de coquilles - c'est là un pari risqué dans le cas de FRANTEXT, où les manipulations successives de la base ont introduit progressivement un nombre assez important de fautes qui s'ajoutent à celles qui, présentes au départ, n'ont pas toujours été corrigées. Mais si on accepte ce présupposé, des suites telles que *le cause, les cause* seront automatiquement analysées

comme Pronom + Verbe, alors que *la cause* aurait besoin d'un contexte plus large pour être désambigüisé. Il ne faut pas s'illusionner sur le rendement de cette prise en compte. Voici les fréquences pour quelques séquences (tableau 3a et 3b).

Ces quelques effectifs suffisent pour montrer

Tableau 3 a

laisse	*le* V	*les* V	*la* N	*la* V	*en* N	*en* V
Effectif	132	52	31	65	51	37

Tableau 3b

aide	*les* V	*l'* N	*l'* V	*en* N	*en* V
Effectif	9	302	76	67	0

- que les effectifs dépendent énormément, non seulement des fréquences respectives des emplois de nom et de verbe, mais aussi de la fréquence d'emploi du nom dans des tours où il est précédé d'une préposition, du caractère transitif ou intransitif du verbe, etc., par conséquent de caractères syntaxiques propres aux unités individuelles concernées ;

- que le rendement de la prise en compte du genre et du nombre varie énormément, non seulement selon que le mot ambigu commence par une voyelle ou une consonne, mais aussi selon l'emploi plus ou moins fréquent du nom avec l'article défini et selon la construction habituelle du verbe ; ajoutons que dans la majorité des occurrences, la nature verbale du mot *laisse* précédé de *la* serait indiquée de toute façon par le mot précédent ; il ne faut donc pas surestimer le rendement quantitatif de ce critère, surtout si l'on pense à des cas de noms comme *garde* ou *aide*, où le nom peut être des deux genres, ce qui diminue encore ce rendement ; mais la prise en compte du genre et du nombre est nécessaire parce que là où ce critère s'applique, il est sûr.

4. Conclusion

Devant les observations très contrastées qu'on peut faire à propos des catégorisations automatiques opérées dans la base catégorisée FRANTEXT, les remarques suivantes me paraissent s'imposer :

- En plus des règles de grammaire locale d'application générale, qui sont nécessaires en particulier dans la mesure où elles suivent des lois distributionnelles rigides et donnent par conséquent des résultats sûrs, il faut recourir dans bien des cas à des traitements ad-hoc qui demandent que le lexique à la base du logiciel de catégorisation ne se contente pas de nommer les classes morpho-syntaxiques, mais inclue aussi des catégories telles que le genre et le nombre, la personne des formes verbales, les constructions les plus fréquentes, les tours fréquents en cours de lexicalisation.

- La recherche d'un résultat totalement dépouvu d'erreurs est une utopie lorsqu'il s'agit d'une base aussi étendue et aussi composite que FRANTEXT ; toutefois les étiquettes de "catégorisation certaine" et de "catégorisation incertaine" demanderaient à être plus solidement assurées dans la réalité des analyses produites.

Ces deux remarques ont en commun de souligner la nécessité de revenir plus souvent à la réalité des textes et de ne pas se contenter de règles énoncées et programmées a priori ou sur un sous-sorpus trop restreint. Je sais que l'analyse manuelle ou semi-manuelle de près de 14000 occurrences de formes ambigües a quelque chose de monstrueux, mais c'est un exercice plein d'enseignements, et dont le fruit est loin de se limiter aux quelques indications que j'ai pu donner ici.

Références

Catach, N. (1984) Phonétisation automatique du français, Paris, CNRS.

Dister, A. (2000) Réflexions sur l'homographie et la désambigüisation des formes les plus fréquentes, JADT 2000 : 131-138, Lausanne

FRANTEXT, base de données textuelles, son logiciel de catégorisation et son système d'aide, 44, avenue de la Libération, 54000 NANCY

Labbé, Dominique (1990), Normes de saisie et de dépouillement des textes politiques. Cahier n° 7 du CERAT, Grenoble.

Montibus, Marie-Jeanne (1990), Lemmatisation semi-automatique du français (Document de travail non publié, issu d'un travail fait avec l'INaLF, Nancy)

Muller, Charles (1982)De la lemmatisation, préface à P. Lafon, Dépouillements et statistiques en lexicométrie, Genève-Paris, Slatkine, repris dans Langue française, linguistique quantitative, informatique, Genève-Paris, Slatkine, 1985.