# Réflexions sur le traitement automatique des langues

Gaston Gross

LLI - UMR 7546 – Université Paris 13
Av. J-B Clément – 93430 Villetaneuse – France
gross@lli.univ-paris13.fr

The aim of this article is to examine two properties of natural languages which both constitute a hindrance to automatic processing and to propose solutions in each case. The first hindrance to any statistical processing of the vocabulary of a text, set phrases, is a constitutive feature of natural language and which occurs on a massive scale and which clearly complicates cutting a text up into units. The most natural way of solving this problem is to make an electronic dictionary with the same information given as for single words. The second difficulty is polysemy. All word categories are syntactically polysemous. It is therefore impossible to even start to analyse a sentence without being able to recognise meaning in context. For this, we put forward the idea of usage, which for a predicate means giving its argument pattern, which thereby defines meaning, actualisation (tense and mode). To describe argument patterns we put forward the idea of object classes, which is a perfectly natural way of factorizing analyses and taking inheritance into account.

## Résumé

Cet article a pour objet d'examiner deux propriétés des langues naturelles qui constituent autant d'obstacles au traitement automatique et de proposer des solutions possibles dans chaque cas. Le premier obstacle à un traitement statistique du vocabulaire d'un texte est le problème du figement, qui est une des propriétés constitutives des langues naturelles et qui constitue un phénomène massif. C'est ici le découpage d'un texte en unités qui est en cause. La façon la plus naturelle de régler ce problème consiste à élaborer des dictionnaires électroniques portant les mêmes indications que pour les mots simples. La polysémie est une seconde difficulté. Toutes les catégories sont syntaxiquement polysémiques. Aussi est-il impossible d'envisager l'analyse d'une phrase sans être en mesure de reconnaître le sens en contexte. Nous proposons à cet effet, la notion *d'emploi* qui constitue, pour un prédicat, à donner son schéma d'arguments, qui définit à son tour le sens, son actualisation (temps et aspect). Pour la description des schémas d'arguments nous proposons la notion de *classes d'objets*, qui permet de façon très naturelle de factoriser les analyses et de rendre compte des héritages.

## Abstract

**Mots-clés :** traitement automatique, figement, polysémie, classes d'objets, héritage.

## 1. Examen critique des conditions linguistiques du traitement statistique des textes

Le traitement statistique des textes pose en premier lieu le problème de la reconnaissance des unités lexicales constitutives. Compter le nombre de « mots » n'est pas une activité triviale. Ecartons d'abord comme peu problématique une particularité des langues naturelles qui influe sur le dénombrement des mots, à savoir l'amalgame : dans la suite *loi du moindre effort* faut-il considérer *du* comme un seul ou deux mots différents ? La syntaxe permet la plupart du temps de résoudre la problème. Deux propriétés des langues naturelles constituent en revanche un obstacle de taille à tout comptage : le figement et la polysémie. Voyons tout d'abord le figement. La définition du mot comme suite de caractères figurant entre deux blancs est une conception naïve du mot. Quel intérêt y aurait-il à décomposer *loi du moindre effort* en quatre unités constitutives et *preuve par neuf* en trois ? La question est d'autant plus importante que

20 à 30 % environ de la surface d'un texte (communication personnelle de M. Gross) est constituée par des mots complexes et que le nombre de mots composés est de l'ordre de plusieurs centaines de milliers. Les statistiques doivent donc porter non sur les éléments lexicaux formels mais sur les unités fonctionnelles, qui ont leur place dans les dictionnaires. Le système sera alors en mesure de les délimiter tout autant que les mots monolexicaux.

Mais, il y a plus important. Faire des statistiques sur les textes implique, par exemple, que l'on puisse déterminer si une forme comme *N de N* constitue une seule unité lexicale ou deux ou trois. La première décision que doit prendre un système en vue de l'analyse des groupes *N de N* consiste à déterminer si les suites ainsi formées sont des libres ou figées. Comme ces dernières doivent figurer dans le dictionnaire (elles ne peuvent pas faire l'objet d'un calcul, leur sens n'étant pas compositionnel), leur reconnaissance tient lieu d'analyse. C'est le cas, par exemple, des noms composés du type *pomme de terre, tour de vis, tête de pont.* Mais il faut dans ce domaine prendre en compte la suite la plus longue du figement. Il existe un composé *corps de garde* que le système doit reconnaître et qui est codé dans le dictionnaire à la fois comme un humain et un locatif. Mais on doit aussi être en mesure de regarder l'environnement immédiat pour éviter de disloquer un nom composé plus long comme *plaisanterie de corps de garde*, dont il est un élément constitutif. De telles analyses ne peuvent se faire que sur la base d'un dictionnaire indiquant les degrés de figement et les sous-ensembles figés de suites (figées) plus longues. Cette précaution permettra de reconnaître *station de sports d'hiver, bande d'arrêt d'urgence, huile de foie de morue, excédent de la balance des paiements, règle de non-affectation des recettes, chemin de fer de ceinture.*

## 2. La polysémie

Le problème posé dans la section précédente est connu depuis longtemps de ceux qui s'intéressent au traitement automatique des langues, même si l'importance fondamentale des dictionnaires dans le traitement automatique n'est pas partagée par tous. La polysémie constitue une autre difficulté théorique au traitement statistique des textes et pourrait représenter une forte limitation à l'intérêt que présente ce type de recherches. Dès les premiers travaux de statistiques, il était de tradition de séparer les formes lexicales appartenant à des catégories différentes (*le* article et pronom ; *que* relatif ou conjonctif).

Mais des distinctions doivent être établies aussi à l'intérieur des catégories elles-mêmes. Le comptage des verbes est une opération assez simple puisque leur morphologie spécifique permet leur reconnaissance, exception faite des formes de la troisième personne du singulier de l'indicatif qui sont souvent homographes des formes nominales. Mais même les dictionnaires actuels les plus détaillés, comme par exemple Bescherelle, ne sont pas entièrement satisfaisants. Dans ces ouvrages, les verbes sont regroupés selon leur conjugaison. C'est oublier qu'une même forme infinitive peut appartenir à deux types de conjugaisons différents en fonction de ses emplois. Le verbe *pleuvoir* appartient à la table 47 de Bescherelle, qui recense les verbes défectifs n'ayant que le 3$^{e}$ personne du singulier : *il pleut il pleuvait, il pleuvra*. Mais il existe un autre verbe *pleuvoir*, qui est un verbe support d'occurrence et que l'on trouve dans des phrases comme *Les coups pleuvaient sur la tête de Paul*. On voit que si on n'est pas en mesure de séparer les emplois, l'analyse n'a guère d'intérêt puisqu'on ne voit rien de commun à ces deux emplois. Donc, pas de morphologie sans syntaxe.

De façon plus générale, l'étude de la fréquence des verbes d'un texte ne peut se faire que si on est en mesure de reconnaître le statut syntaxique de chaque forme verbale. Or, il existe au moins six types de verbes différents.

a) Les plus fréquents sont évidemment les verbes prédicatifs, qui ont des arguments et que la tradition assimile à la catégorie.

b) Les verbes figurant dans les expressions verbales figées et qui ne doivent pas être confondus avec les précédents, dans la mesure où on ne peut pas parler d'arguments à leur propos : dans *prendre la poudre d'escampette*, le substantif ne peut pas être assimilé à un argument du verbe *prendre*. Pour éviter de leur attribuer des propriétés distributionnelles, il faut là encore les lister et les traiter comme des blocs inanalysables qui doivent être traduits comme tels.

c) Les pro-verbes qui fonctionnent comme des substituts de classes de verbes assez générales. C'est le cas de *faire* qui sert d'anaphore aux prédicats d'action.

d) L'emploi causatif de *faire (faire travailler les enfants)*, qu'on ne doit pas confondre avec le précédent. Voici quelques autres causatifs : *provoquer (un incendie), rendre (rendre fou), mettre (mettre en difficulté)*. Les causatifs sont des verbes qui opèrent sur d'autres prédicats. Ils figurent donc dans des phrases complexes.

e) Les deux derniers types de verbes représentent des auxiliaires. Les plus connus sont les auxiliaires verbaux, qui traduisent le temps mais plus souvent l'aspect (*aller, venir de, être sur le point de*).

f) Enfin, il existe des auxiliaires de prédicats nominaux, les verbes supports, qui conjuguent ces prédicats. Ces verbes sont donc syntaxiquement très différents, ils ne sont jamais prédicatifs.

Le simple décompte des formes verbales dans un texte ne rendrait pas compte de ces différences et n'apporterait aucune information linguistique pertinente.

## 3. Propositions : la notion d'emploi

Les observations que nous venons de faire montrent que le travail de description ne peut pas prendre pour argent comptant la notion de catégories grammaticales, car toutes peuvent être syntaxiquement polysémiques. Mais on doit aller plus loin. Si on examine les verbes prédicatifs eux-mêmes, force est de constater que tous sont polysémiques et qu'il est donc illusoire par exemple de parler dans un texte « du » verbe *tenir*. Une rapide description de ce verbe permet de montrer un grand nombre d'emplois différents, de sorte que des statistiques portant sur la forme morphologique elle-même n'ont qu'un intérêt réduit. L'idée d'un classement sémantique des lexèmes a été envisagée par Ch. Muller (Préface à Lafon, 1984) « (Le classement sémantique) est une question de temps et de moyens plus que de théorie ; on ne voit guère quelle objection de principe pourrait être opposée à des pratiques aussi obligeantes ». Abordons l'étude de *tenir* par les emplois de supports, c'est-à-dire d'actualisateurs de prédicats nominaux :

*Paul tient un discours politique*

*Paul tient une bonne cuite*

Il y a des emplois d'opérateurs à lien :

*Cet enfant est sous l'autorité du juge*

*Le juge tient cet enfant sous son autorité*

Parmi les emplois prédicatifs, on peut distinguer d'abord les emplois de type « locatif » :

*Paul tient cet enfant par la main*

*Paul tient ce bijou dans sa main*

*Ce bagage ne tient pas dans cette malle*

*Cela tiendrait dans le creux de ma main*

D'autres sont appropriés à certaines activités commerciales :

*Paul tient un commerce de voitures*

*Paul tient un stand au marché de Noël*

Il y a un emploi « passif » de prédicats :

de <don> ou de <transmission> *:*

*Paul tient cette nouvelle de son voisin*

*Paul tient cette maison de son grand-père*

de sentiments :

*Paul tient à ce voyage*

de « ressemblance » :

*Cet enfant tient de sa mère*

*Ce discours tient du délire*

aspectuels :

*La promesse tient*

*La colle tient*

*Le pont tient*

Ces emplois sont si différents que toute manipulation qui les assimilerait ne rendrait pas compte de leur fonctionnement linguistique. Il faut donc être en mesure de reconnaître à quel ensemble appartient un mot, c'est-à-dire de préciser son emploi. Quand nous parlons d'emploi, nous entendons plusieurs paramètres intégrés. Un emploi prédicatif est défini par :

a) Un domaine d'arguments (ces arguments, dont on note la suite la plus longue, sont définis à l'aide des classes d'objets) ;

b) Un sens, d'où par conséquent un ou plusieurs synonymes, un antonyme et une traduction ;

c) Une forme morphologique : *verbe*, *nom*, *adjectif* pour les prédicats des phrases simples (prédicats du premier ordre) ; *prépositions* et *conjonctions* pour les prédicats du second ordre ;

d) Une actualisation : conjugaison pour les verbes, verbes supports pour les prédicats nominaux. Certains prédicats sont défectifs du point de vue de leur actualisation : *regarder* (concerner) n'a pas de passé composé ;

e) Un système aspectuel. Il doit y avoir compatibilité aspectuelle entre les différents éléments porteurs d'aspect dans la phrase (déterminants des arguments, adverbes, adjectifs, etc.). *Peur* et *peureux* n'ont pas le même aspect et ne peuvent pas être considérés comme constituant le même prédicat : *peur* est un sentiment qui peut être ponctuel ; *peureux* ne désigne pas un sentiment mais un trait de caractère et est nécessairement duratif. En revanche, *désirer* et *désireux* ont le même système aspectuel ;

f) Des restructurations (transformations) qui lui sont propres : passivation, thématisations différentes, pronominalisations, etc.

L'information la plus importante est constituée par le schéma d'arguments, car c'est de lui que dépendent les autres propriétés. Or, pour mettre en évidence ces schémas, il faut être en mesure de décrire les arguments avec précision.

## 4. Les classes d'objets

C'est un fait admis qu'un prédicat est d'abord défini par son domaine d'arguments, c'est-à-dire son sujet et ses compléments. Nous avons vu que la plupart des prédicats sont polysémiques. C'est une observation empirique que tout changement de sens d'un prédicat est corrélé à un changement de son schéma d'arguments. Soit la phrase *Vous suivrez ce chemin*. Si on remplace l'objet *chemin* par des substantifs comme *route, rue, voie, sentier* le verbe *suivre* garde le même sens. On regroupera ces mots sous le terme générique de <voies>. Si en revanche, on remplace le mot *chemin* par *cours*, alors on a affaire à un autre emploi et le substantif *cours* peut être remplacé par *séminaire, stage, formation, cycle d'étude*, etc., qu'on rangera sous le classifieur <enseignement>. Le sens du verbe *suivre* serait encore différent si le complément était *recommandation, suggestion, avis* qu'on classerait comme <conseil>, ou encore *cure, médication, régime , thérapeutique* qui relèverait de la classe des <traitements>. La mise au point du sens exige que l'on soit à même de préciser la nature sémantique des arguments que prend un emploi donné de prédicat. Les ensembles lexicaux représentant les arguments en compréhension s'appellent *classes d'objets*. Il est donc indispensable qu'une information comme <voies >, <enseignement>, <conseil >, <traitement> correspondant à des *classes d'objets*, figure dans le dictionnaire comme classifieurs de substantifs décrivant des positions argumentales.

## 5. Opérations sur la base de ces descriptions

Tous les substantifs figurant dans une même position argumentale pour un sens déterminé appartiennent à la même classe d'objets. Dès lors que l'on a séparé les différents emplois d'un prédicat à l'aide des classes d'objets, on est en mesure de reconnaître ou de générer toutes les phrases correspondant à chacun des emplois. Le sens d'un prédicat est donc fonction du schéma d'arguments. Ainsi, le verbe *régler* peut avoir comme arguments *régler/N0 :hum/N1 :<facture>/N2 : à hum*. Appartiennent à la classe <*facture*> les éléments suivants : *addition, compte, état de frais, facture, note, relevé de compte*. À cette liste, il faut ajouter des termes populaires (*douloureuse*) ainsi qu'un très grand nombre de noms composés, où le complément en *de N* spécifie l'objet du règlement : *note d'électricité, note d'honoraires, note d'hôtel, note d'un artisan, note d'un entrepreneur, note de blanchisseuse, note de crédit, note de droit d'auteur, note de frais de transport, note de frais, note de gaz, note de manucure, note de pressing, note de restaurant, note de téléphone***.**

L'ensemble de ces mots forme une des classes sémantiques possibles en position d'objet du verbe *régler*. Cette classe ne doit pas être confondue avec une autre qui désigne un reçu représentant une attestation de paiement : *attestation, bon de caisse, bulletin de bagages, bulletin de consigne, justificatif, récépissé, reçu, vignette, reçu de carte bancaire, ticket de caisse*.

Inversement, on peut aussi se demander quels sont, pour une classe donnée, les opérateurs qui lui sont appropriés. Dans la démarche homographique que nous adoptons et qui considère que les différents emplois d'un prédicat polysémique constituent des ensembles disjoints, un argument n'est pas essentiellement défini par ses traits sémantiques inhérents mais par l'ensemble des prédicats qui lui sont strictement appropriés.

*acquitter/N0:<acheteur>/N1:<facture>/N2:*

*annuler/N0:<acheteur>/N1:<facture>/N2:*

*augmenter/N0:<vendeur>/N1:<facture>*
*baisser/N0:<vendeur>/N1:<facture>/N2:*
*corser/N0:<vendeur>/N1:<facture>/N2:*
*demander/N0:<acheteur> /N1:<facture>/N2:à <vendeur>*
*dresser/N0:<vendeur>/N1:<facture>/N2: à l'ordre de <acheteur>*
*établir/N0:<vendeur>/N1:<facture>/N2: à l'ordre de <acheteur>*
*fournir/N0:<vendeur>/N1:<facture>/N2:à <acheteur>*
*grossir/N0:<vendeur>/N1:<facture>/N2:*
*honorer/N0:<acheteur>/N1:<facture>*
*payer/N0:<acheteur>/N1:<facture>/N2:à <vendeur>*
*présenter/N0:<vendeur>/N1:<facture>/N2:à <acheteur>*
*recevoir/N0:<acheteur>/N1:<facture>/N2: de <vendeur>*
*rédiger/N0:<vendeur>/N1:<facture>/N2:*
*réduire/N0:<vendeur>/N1:<facture>/N2: de < %>*
*régler/N0:<acheteur>/N1:<facture>/N2:à <vendeur>*
*s'élever/N0:<facture>/N1: à Card<unité monétaire>*
*solder/N0:<acheteur>/N1:<facture>/N2:*

Les prédicats appropriés de *<reçu>* sont entre autres les suivants :

*antidater/N0:<vendeur>/N1:<reçu>/N2:*
*dater/N0:<vendeur>/N1:<reçu>/N2:*
*délivrer/N0:<vendeur>/N1:<reçu>/N2:à <acheteur>*
*dupliquer/N0:<vendeur>/N1:<reçu>/N2:*
*exiger/N0:<acheteur>/N1:<reçu>/N2: de <vendeur>*
*fournir/N0:<vendeur>/N1:<reçu>/N2: à <acheteur>*
*recevoir/N0:<acheteur>/N1:<reçu>/N2: de <vendeur>*
*remettre/N0:<vendeur>/N1:<reçu>/N2:à <acheteur>*
*valider/N0:<vendeur>/N1:<reçu>/N2:*
*viser/N0:<vendeur>/N1:<reçu>/N2:*

Au regard des verbes que nous venons d'énumérer les substantifs de la classe des <factures> se comportent de la même façon. Les distinctions qu'on peut établir ne sont pas de nature syntaxique mais pragmatique. Une *addition* est une <facture> que l'on établit dans un restaurant, une *note* est la <facture> qu'on paie dans un hôtel, un *état de frais* est établi par un employé à l'intention de son employeur, un *relevé (de compte)* représente la <facture> que l'on reçoit d'une administration prestatrice de services (EDF, GDF). Ces informations peuvent être ajoutées dans une base de données en ouvrant un champ indiquant les domaines.

## 6. Classes linguistiques et classes référentielles

La notion de classes (et d'hyperclasses) a comme premier avantage de pouvoir factoriser et par là de décrire de façon compacte tous les éléments d'une classe d'objets donnée. Ainsi tous les éléments d'une classe héritent de l'ensemble des prédicats appropriés de celle-ci, comme nous venons de le voir. Nous abordons ici la notion générale d'héritage et nous commençons par des cas où les ensembles sont homogènes. Le remplacement des éléments par leur classe

représente une grande simplification de la description. Si d'une part, on attribue dans un dictionnaire électronique à chaque substantif le code de la classe à laquelle il appartient et si, d'autre part, on décrit les schémas d'arguments des prédicats à l'aide de ces classes, alors on est en mesure de reconnaître ou de générer l'ensemble des phrases appartenant à un emploi donné. Mais cela n'est possible que si l'on a pris soin de créer des classes « linguistiques », c'est-à-dire des classes qui comprennent des substantifs ayant exactement les mêmes propriétés sémantiques et syntaxiques. Dans notre démarche, un mot est défini par son environnement et non en lui-même, comme c'était le cas dans l'analyse sémique traditionnelle. Nos classes représentent donc des ensembles lexicaux, l'ensemble des <factures>, l'ensemble des <habitations>, l'ensemble des <unités monétaires>.

Mais tous les ensembles ne constituent pas des classes au sens où nous l'entendons ici. Prenons un terme collectif comme les *effets personnels*, défini ainsi par le Nouveau Grand Robert « 3. (XVII^e^). *Cour. Le linge et les vêtements. – Affaire (affaires), défroque, fringue, frusque, habit, harde, nippe, trousseau, vêtement. Mettre ses effets dans une valise. Ballot d'effets. – Bagage. Les effets d'un militaire. – Paquetage. Effets civils, militaires.* ». On voit que ce terme désigne des éléments qui n'ont pas le même comportement syntaxique : les prédicats appropriés au mot <linge> ne sont pas ceux qui s'appliquent aux <vêtements> : on met un vêtement mais non un linge. Il n'est pas clair non plus si le terme « bagage » est compris dans la définition. On a affaire ici à des classes référentielles mais non linguistiques.

La même observation peut être faite avec un terme comme *vaisselle* que le NGR définit « 2. (XIX^e^). Ensemble des plats, assiettes, ustensiles de table, etc., qui sont à laver. *Laver, faire la vaisselle. – Nettoyer, relaver* (régional). Laver les vaisselles. Écurer (vx), égoutter, rincer, essuyer la vaisselle. Laveur de vaisselle. – Plongeur. Machine à laver la vaisselle. – Lave-vaisselle. Laisser s'entasser la vaisselle. – Bac à vaisselle ». Il est clair que les contenants (*plat, assiette, saucier*, etc.) n'ont pas les mêmes opérateurs appropriés que les ustensiles de table (*cuiller, fourchette, couteau*). Ces derniers ne constituent pas non plus une classe « linguistique ». Il va de soi, d'autre part, que les classes doivent pouvoir faire l'objet d'une énumération objective et donc d'un consensus. Un ensemble comme les « choses écœurantes » ne constitue pas une classe linguistique, car une description en extension serait aléatoire.

Prenons un autre exemple. Les <sports> ont comme opérateur approprié le verbe *pratiquer*, les <mouvements> *effectuer*, les <matières scolaires> *faire (du)*. Il serait raisonnable de penser que ces termes ont comme hyperonyme le mot *activité*, dont le verbe approprié est *exercer : il exerce une activité débordante*. Or, les hyponymes n'héritent pas cet opérateur **exercer du foot, *exercer une promenade, *exercer le latin.* On ne classera donc pas ces diverses actions sous le terme générique d'activité.

## 7. Classes et sous-classes

Nous avons vu que les substantifs de la classe des <factures> ont tous le même comportement syntaxique : ils prennent les mêmes prédicats appropriés. Il n'y a donc aucune raison de les sous-catégoriser du point de vue linguistique. Les différences observées relèvent de considérations pragmatiques, comme nous l'avons vu : telle facture est propre aux restaurants (*addition*) , telle autre aux hôtels (*note*) ou encore aux relations professionnelles (*état de frais*). Mais la description adéquate d'ensembles lexicaux nécessite la plupart du temps que l'on subdivise les classes en sous-ensembles. Prenons l'exemple de la classe des <boissons>. Observons d'abord que les notions de *boire* et de *boisson* ne sont pas à mettre sur le même plan. Au sens strict du mot, *boire* a comme objet un <liquide>, car on peut boire par inadver-

tance des liquides non destinés à cet effet : produits pharmaceutiques, carburants, etc. Beaucoup de produits d'entretien portent l'indication « ne pas avaler ». Le terme de <boisson> ne doit donc pas être confondu avec les compléments possibles du verbe en position d'objet. Une boisson est un <liquide> destiné à être bu. Cette classe exclut les liquides dont nous avons parlé. Cela dit, on peut alors faire le recensement de tous les prédicats appropriés à cette classe dans son ensemble :

Verbes : *boire, siroter, siffler*

Sont communs aux aliments : *absorber, avaler, ingurgiter*

Prédicat nominal (aspect itératif) *: être buveur de*

Adjectifs : *froid, glacé, tiède, chaud, brûlant, doux, insipide, fade, buvable, imbuvable, potable, non potable*

Les opérateurs que nous venons de donner sont communs à toutes les boissons. Il faut ensuite mettre au point des sous-classes. Une première grande division est celle qui sépare les boissons alcoolisées des autres. On procédera pour les boisson alcoolisées à la même factorisation en mettant en évidence les opérateurs généraux puis en se servant de nouveaux opérateurs pour créer des sous-classes. Parmi les opérateurs appropriés aux <boissons alcoolisées> on trouve en position de sujet, entre autres :

*titrer /N0:<alcool>/N1 :[*Card] *degré(s),*
*enivrer/N0:<alcool>/N1:hum,*
*saouler/N0:<alcool>/N1:hum*

En position d'objet avec un sujet humain, on peut relever : *cuver, frelater, picoler, pinter, pitancher, trafiquer*. Les prédicats nominaux sont entre autres : *avoir* [Card] *degré(s), avoir une teneur en alcool de* [Card] *degrés*. Et les prédicats adjectivaux : *âpre, doux, léger, lourd, moelleux, raide, sec.*

À l'aide d'autres opérateurs on peut établir des sous-classes. Les <alcools et spiritueux> prennent le verbe *distiller (N0:hum/N1:<alcools et spiritueux)*. Les prédicats appropriés aux <vins> sont plus nombreux :

Verbes :

*accompagner/N0: <vin>/N1 :<plat>*
*débourber/N0:hum/N1: <vin>*
*décanter/N0:hum/N1: <vin>*
*se madériser/N0: <vin>*
*tirer/N0: <vin>*

Adjectifs :

*aigre/N0:<vin>*
*aigrelet/N0:<vin>*
*âpre/N0:<vin>*
*astringent/N0:<vin>*
*bourru/N0:<vin>*
*capiteux/N0:<vin>*
*charnu/N0:<vin>*
*charpenté/N0:<vin>*
*corsé/N0:<vin>*
*équilibré/N0:<vin>*
*gouleyant/N0:<vin>*

*grêle/N0:<vin>*
*harmonieux/N0:<vin>*
*pétillant/N0:<vin>*
*piquant*
*piqué/N0:<vin>*
*tuilé/N0:<vin>*
*vert/N0:<vin>*
*vieux/N0:<vin>*

Un verbe comme *brasser* s'applique aux <bières>, qui prennent aussi des adjectifs comme :

*aigre/N0:<bière>*
*amère/N0:<bière>*
*filante/N0:<bière>*
*forte/N0:<bière>*
*plate/N0:<bière>*

## 8. Relations d'héritage

Un élément lexical donné prend, bien entendu, les opérateurs qui sont strictement appropriés à sa classe. Ainsi, un <vin> est défini par des prédicats verbaux comme *madériser* ou adjectivaux comme *gouleyant* ou *charnu*. Mais comme relevant de l'hyperclasse <alcool>, il hérite de l'ensemble des opérateurs qui définissent ce niveau : il peut *enivrer, monter à la tête, saouler* ; on peut le *cuver*, le *tenir*. Il peut avoir une *teneur de* [Card] *degrés*. Au niveau supérieur, comme tout alcool est une boisson, le vin peut *se boire*, s'*absorber*, *se siroter*, *se siffler* par un *buveur de vin*. Il peut être *froid, glacé, tiède, chaud, doux, insipide, fade, buvable, imbuvable, potable, non potable*. Mais on peut aussi subdiviser les vins à l'aide des adjectifs *rouge, blanc, rosé, vert, jaune*. Ces adjectifs ne sont pas des qualificatifs mais des désignatifs caractérisant différents types de vins. Cette sous-classification est référentielle mais non linguistique, car elle ne permet pas de mettre en évidence des opérateurs qui seraient appropriés à chacun de ces types de vins. Les vins peuvent encore être subdivisés en appellations et noms de marques. Là non plus, nous n'avons pas affaire à de vraies sous-classes linguistiques, car elles ne génèrent pas non plus de syntaxe spécifique.

Si l'on devait faire une arborescence rendant compte de la syntaxe des noms de <vins>, on partirait au niveau la plus élevé de la notion de <concret>. De ce fait, ces substantifs hériteraient de toutes les propriétés générales des concrets : poids, volume, couleur, etc. Ensuite on peut hésiter sur l'ordre de deux traits : <artefact> ou <liquide>. Nous choisissons d'abord le trait <liquide>, ce qui permet de prédire des verbes comme : *verser, couler, déborder, imbiber, s'égoutter* ou des adjectifs comme *dense, fluide, huileux*. Ensuite, une subdivision séparera les liquides naturels (dont on vient de voir la syntaxe) des <(liquides) artefacts>, qui ont des verbes comme *fabriquer, réaliser, mettre au point* mais aussi *vendre, acheter, avoir tel ou tel prix* et tous les autres prédicats pouvant caractériser les produits commerciaux. Ensuite, on notera qu'il s'agit de <boissons>, puis de <boissons alcoolisées> et enfin de <vins> et on aura les opérateurs appropriés que nous avons notés plus haut. Un tel travail descriptif est d'abord un problème linguistique plus que de représentation informatique.

## 9. Héritage et métaphore

La définition du sens à l'aide de l'environnement permet de détecter des emplois métaphoriques. Prenons les moyens de transports. En tant que tels, ils ont des opérateurs généraux *se déplacer en, voyager en, aller en, arrêter, descendre de, monter en, prendre,* etc. Pour les décrire, il y a deux grands paramètres d'analyse. Le premier met en jeu le mode de transport : terrestre, maritime, ou aérien. Chacun de ces types de transports a des prédicats appropriés et dont la liste n'est pas difficile à établir : pour les avions : *atterrir, décoller, descendre en piqué, descendre en spirale, descendre en vrille, piquer, plafonner, planer, s'écraser, s'écraser au sol, se cabrer, se poser ;* pour les bateaux *appareiller, chavirer, démâter, dériver, faire*

*eau, gîter, lever l'ancre, mouiller, s'échouer, tanguer ;* pour les transports terrestres *caler, circuler, freiner, rouler, prendre la route de, verser, dépasser*.

À cela s'ajoute la distinction entre moyens de transports individuels et collectifs. Cette opposition est linguistique et pas seulement pragmatique. S'il est possible de *prendre* tout type de moyens de transports, *emprunter* n'est possible qu'avec les transports en commun. Ces derniers ont aussi comme particularité d'être suivis d'un horaire (*le train de midi, *la voiture de midi)* et d'une destination (*le train de Paris, *la voiture de Paris*). Ils ont aussi un grand nombre de prédicats appropriés. En position de sujet on trouve *desservir la destination de, accuser un retard de, annoncer un retard de, être à destination de, partir à n heures* et en position d'objets *attraper le dernier, embarquer dans, embarquer sur, emprunter, manquer son, louper, partir par, prendre le dernier, rater*.

Parmi les transports routiers, on peut isoler une sous-classe particulière, celle des <transports par animal> : *cheval, mulet, chameau, âne,* etc. On trouve alors des opérateurs appropriés : *voyager à dos de, faire une promenade à, monter N en amazone, monter, faire du, être à califourchon sur, se déplacer à dos de, tomber de, faire une chute de*. Il va de soi que ces animaux ne sont interprétés comme des moyens de transports qu'avec les opérateurs que nous avons mentionnés. D'autres verbes les feraient appartenir à la classe des animaux de traits : *brider, harnacher, bouchonner, ferrer, soigner, seller*.

Si maintenant on analyse le comportement des moyens de transports individuels appelés <deux-roues>, on observe qu'ils ont des opérateurs appropriés communs avec les <moyens de transports animaux>. À la différence des autres moyens de transports qui prennent la préposition *en (en bateau, en voiture, en train),* on a ici la préposition *à (être, aller, monter, faire un tour) à (cheval, vélo, moto*). On trouve aussi la préposition *sur : être perché sur (son vélo, son cheval)*. Ce sont des compléments naturels du verbe *enfourcher : enfourcher (son cheval, son vélo*). De plus, ils ont en commun des prédicats de mouvement : (*tomber, faire une chute) de (cheval, vélo, moto*). Observons encore qu'à la différence des autres moyens de transports, un <deux-roues> ne peut ni *partir* ni *arriver*. La métaphore est donc une particularité des langues naturelles qui interdit que l'on établisse des arborescences en dehors de la syntaxe.

## 10. Les unités lexicales complexes : héritages multiples ou autonomie ?

Nous avons vu avec les moyens de transports animaux que certains substantifs peuvent appartenir à plusieurs classes. Par exemple, un cheval appartient à la classe des <équidés> et par là des <mammifères>, on aura alors des opérateurs comme *pouliner, mettre bas*. Il peut aussi appartenir à l'ensemble des <animaux de traits> avec des verbes comme *atteler, dételer, harnacher* ou encore des <animaux de course> et on aura les verbes *monter, entraîner, jouer sur, miser sur*. Cela pose de façon générale le problème des arborescences et des conditions nécessaires à la création de nouvelles classes. Les substantifs qui sont compléments à la fois des verbes *porter, mettre, enfiler* et *ôter* sont des vêtements. Un substantif comme *cotte de maille* entre dans cette classe. Or, ce substantif peut être sujet d'un verbe comme *rouiller*. Faut-il de ce fait créer une nouvelle classe de <vêtements> ? La réponse est d'ordre statistique. On peut penser que les occurrences du verbe *rouiller* sont si rares qu'on peut le négliger. Mais cette remarque doit être étayée. Tout d'abord il existe d'autres noms de vêtements qui peuvent être en métal : *cuirasse, heaume, casque,* etc. D'autre part, d'autres verbes pourraient être appropriés à des objets métalliques *tinter, faire du bruit, résonner,* etc. Il y aurait donc intérêt à ouvrir une sous-classe de <vêtements de chevalerie>. D'autres cas vont en sens inverse. Le verbe *chausser* a comme compléments des substantifs de la classe des <chaussures> : *pantoufle, soulier, botte, espadrille, basket,* etc. Mais il existe un emploi où le verbe a

comme objet le mot *lunettes* (ou *bésicles*). Ces mots appartiennent à la classe des <prothèses> qui ont avec les <vêtements> beaucoup de verbes en commun : *porter, mettre, ôter* mais non *enfiler*. Faut-il créer une classe autonome <lunettes> du seul fait que l'on peut utiliser le verbe *chausser*, on peut hésiter. Cela dépend de l'objectif qu'on se fixe.

Dans d'autres cas, le problème se pose de façon plus sérieuse. La syntaxe du mot *livre* a été souvent examinée (cf. Kayser, 1987 et 1989 ; Kleiber et Riegel, 1989 ; Pujstejovsky, 1995 ; Pustejovsky et Bouillon, 1995). Comme nous définissons les arguments (i.e. les substantifs) par leurs prédicats appropriés, il est clair qu'un même mot entrera dans autant de classes qu'il sera défini par des séries prédicatives différentes. Cela est évident pour les homographes ou les mots polysémiques : chien (canidé : *aboyer*), chien (injure : *traiter qq de* ), chien (pièce coudée de certaines armes à feu : *abattre*). Voyons le cas du mot *livre*. Il est caractérisé par des séries prédicatives très diverses. Il peut être interprété comme :

a) un concret : tenir (dans ses mains), peser (tant), être adj de couleur, tomber

b) un abstrait : être obscur, difficile, indéchiffrable

c) un humain : prétendre, affirmer, exposer, révéler

d) un locatif : contenir (n chapitres), comprend (des erreurs), parcourir (un livre)

e) un événement : *paraître, sortir*

Ici, il est difficile de dire que l'une des séries est plus fréquente ou plus naturelle ou disponible que les autres. Théoriquement, les substantifs qui relèvent de cette classe pourraient avoir des héritages multiples. Le problème posé ici n'est pas de savoir comment on peut représenter informatiquement les héritages multiples mais comment on doit rendre compte des faits linguistiques. Le mot *livre* doit être décrit de façon plus précise. Il est clair qu'un livre représente un <texte> et que de ce fait il est compatible avec des adjectifs comme *obscur, long, indéchiffrable* et des adjectifs comme *écrire* et *lire*. Mais un livre est aussi un <support de textes> comme *journal, revue, périodique*. Rappelons d'abord qu'on ne doit pas confondre les <supports de textes> avec les <supports d'écriture> : *cahier, ardoise, calepin*. Il y a une différence entre ces deux classes : on peut lire un livre ou un journal mais non un cahier ou une ardoise.

Dans l'interprétation du mot *livre* on est en présence d'une série de métonymies. Un texte peut par métonymie être assimilé à un humain, d'où des adjectifs communs : *obscur, incompréhensible (texte, auteur)*, mais *indéchiffrable* ne semble s'appliquer qu'à des textes. Le constitution d'un livre en chapitres en fait métaphoriquement un lieu. L'ensemble des prédicats que nous venons de donner ne s'appliquent qu'à la classe des <livres>. Il y a donc intérêt à considérer des classes de ce type comme autonomes et sans lien avec des hyperclasses, en quelque sorte comme des entités « autocéphales ». Cela pourrait multiplier les classes mais aurait l'avantage de la précision.

## Conclusion

Dans ces pages, nous avons essayé de soulever quelques difficultés que présentent les langues au traitement automatique et en particulier aux lemmatiseurs. Nous nous sommes placés dans la perspective de la reconnaissance automatique et nous avons montré que chaque forme appartient à un emploi, au sens technique du mot que nous avons décrit au paragraphe 3. Cette notion d'emploi insère les éléments lexicaux dans des phrases où les schémas d'arguments spécifient le sens en contexte des prédicats. D'autres informations spécifiques à chaque emploi sont notées, de telle façon que le levée de la polysémie et en général des ambiguïtés

soit possible sur la base des propriétés, qui figurent dans un lexique électronique. Celles-ci peuvent être considérées dans les textes comme des indices permettant de reconnaître l'emploi effectif parmi un grand nombre d'autres possibles. Notre visée est donc celle d'une automatisation des procédures de reconnaissance.

## Bibliographie

Bescherelle. (1997). *La conjugaison pour tous*. Hatier.

Gross G. (1989). *Les constructions converses du français*. Droz.

Gross G. et Clas A. (1997). Synonymie, polysémie et classes d'objets. *Meta*, vol. (42/1). Presses de l'Université de Montréal : 147-155.

Gross G. (1998). Pour une véritable fonction *Synonymie* dans un traitement de texte. *Langages*, vol. (131). Larousse : 103-114.

Gross G. (1999). Élaboration d'un dictionnaire électronique. *Bulletin de la Société de Linguistique de Paris*, Tome (XCIV/1). Peeters : 113-138.

Gross G. (1997). Les classes d'objets et le désambiguïsation des synonymes. *Cahiers de Lexicologie*, vol. (70). Didier Erudition : 27-40.

Gross G. (1998). Pour une typologie des prédicats. *Prédication, assertion, information*. *Studia Romanica Upsaliensis*, vol. (56). Uppsala : 221-230.

Gross G. (1999). La notion d'emploi dans le traitement automatique. *La pensée et la langue*. *Wydawnictowo Naukowe AP :* 24-35.

Gross G. et Guenthner Fr. (1999). Traitement automatique des domaines. *Revue Française de Linguistique Appliquée*, vol. (III/2) : 47-56.

Kayser D. (1987). Une sémantique qui n'a pas de sens. *Langages*, vol. (87) : 33-45.

Kayser D. (1989). Réponse à Kleiber et Riegel. *Lingvisticae Investigationes*, vol. (XIII/2) : 419-422.

Kleiber G. (1984). Polysémie et référence : la polysémie, un phénomène pragmatique ? *Cahiers de lexicologie*, vol. (44/1) :85-103.

Kleiber G. et Riegel M. (1989). Une sémantique qui n'a pas de sens n'a pas de sens. *Lingvisticae Investigationes*, vol. (XIII/2) : 405-417.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Pustejovsy J. et Bouillon P. (1995). Aspectual Coercion and Logical Polysemy. *Journal of Semantics*, vol. (12) : 133-162.

Pustejovsky J. (1995). *The Generative Lexicon*. The MIT Press.

Tournier M. (1985). Sur quoi pouvons-nous compter ? Réponse à Charles Muller. *Verbum, Hommage à Hélène Nais* : 481-492.