LE PESANT Denis. 1996. « Principes d'organisation des données lexicales dans un dictionnaire électronique ». *Sémiotiques* 14 : 35-54. Paris : INALF et Didier-Erudition.

Principes d'organisation des données lexicales dans un dictionnaire électronique

Denis Le Pesant

Laboratoire de Linguistique Informatique (LLI) CNRS : UMR 195
et Université d'Evry

Introduction

L'équipe *Lexique et Grammaire* du Laboratoire de Linguistique Informatique (LLI) consacre l'essentiel de ses activités, sous la direction de Gaston Gross, à élaborer une base informatisée de données lexicales destinée à servir de substrat à de futurs dictionnaires, traditionnels ou électroniques, et de formats divers (dictionnaires portant sur la langue générale ou sur les vocabulaires de spécialité, dictionnaires monolingues ou bilingues, dictionnaires didactiques destinés à des apprenants d'âge et de niveau divers, etc.). Après avoir indiqué quelles sont les exigences particulières qui s'imposent dans la mise en oeuvre d'une telle base de données, nous montrerons que le caractère global du projet, ainsi que la nature informatique du support, nécessitent tout à la fois une redéfinition de certaines notions lexicologiques traditionnelles et l'élaboration d'outils théoriques nouveaux.

1. Un impératif : construire un dictionnaire explicite

Des dictionnaires comme le Grand Robert ou le Trésor de la Langue Française, sont des dictionnaires non explicites. Ces ouvrages, que nous utilisons quotidiennement et que nous admirons sans réserves, ont été conçus pour être exploités par des êtres humains possédant la langue dans laquelle ils sont rédigés. Le cerveau humain a des aptitudes de mémorisation, de déduction et d'extrapolation telles, que les informations n'ont pas besoin d'être fournies sous une forme entièrement explicite. Du reste, une trop grande explicitation des informations rendrait ces livres impossibles à consulter.

Le cas des dictionnaires électroniques est différent. Ils sont aussi destinés à des cerveaux humains, mais à des cerveaux qui ont choisi de recourir à l'assistance de machines pour la réalisation de certaines tâches de recherche coûteuses en temps. Les ordinateurs rendent à la fois possible et nécessaire l'explicitation des informations linguistiques dans les dictionnaires. Aussi les dictionnaires électroniques visent-ils à être aussi explicites que possible.

Une information lexicale explicite est une information qui n'exige ni compétence linguistique, ni calculs inférentiels complexes pour être traitée. Un dictionnaire électronique explicitera notamment, autant que le permet l'état des connaissances linguistiques, les informations distributionnelles, syntaxiques et sémantiques.

1.1. Des informations distributionnelles explicites et présentées en extension

Un dictionnaire aussi complet que le Grand Robert n'a pas besoin, par exemple, d'expliciter le fait que le verbe *blâmer* sélectionne (les emplois métonymiques mis à part) un sujet humain. La définition proposée est : *porter un jugement défavorable sur quelqu'un ou quelque chose*. Une information sur le sujet figure-t-elle dans les entrées *jugement* ou *juger* ?

Pas davantage. S'agissant de la position de complément, il n'est pas précisé ce que recouvre exactement *« quelque chose »* (or il est clair que n'importe quel nom non humain ne peut pas être complément d'objet de *blâmer*). Certes, les constructions suivantes sont mentionnées plus loin : *blâmer la conduite, les agissements de quelqu'un*. Mais les noms de comportements humains sont-ils les seuls noms non humains à pouvoir être sélectionnés par *blâmer* ? Ce n'est pas précisé ; mais le lecteur, en découvrant dans l'exemple 10 la construction « *blâmer la curiosité, la finasserie de quelqu'un* », en infère que *blâmer* sélectionne aussi certains noms de qualités humaines. On pourrait dire à bon droit que si de telles informations sont inutiles dans ce dictionnaire, c'est parce qu'elles sont inférables au cours d'une consultation intelligente ; mais il est plus exact de dire qu'elles sont inutiles parce qu'elles sont *évidentes* pour n'importe quel lecteur du Grand Robert, ouvrage dont la lecture exige une compétence assez élevée en français.

Un dictionnaire électronique pourra et devra au contraire donner des informations explicites sur la distribution des prédicats. Par exemple, il devra expliciter le fait que *blâmer* sélectionne des noms humains tant en position de sujet qu'en position de complément d'objet. Il devra en outre préciser quelles sont les catégories de noms non humains (attitudes, comportements, qualités humaines, actions humaines) qui peuvent figurer en position de complément d'objet.

D'autre part, les informations sur la distribution des prédicats seront données *en extension*, ce qui veut dire que les noms pouvant figurer en telle ou telle position syntaxique pourront être énumérés sur demande de l'utilisateur. Par exemple, le dictionnaire devra être capable d'énumérer *tous* les noms de qualités humaines pouvant figurer en position de complément d'objet du verbe *blâmer*. Nous montrerons plus loin que la notion de *classe d'objets* permet de réaliser un tel objectif.

1.2. Des informations syntaxiques explicites et présentées de façon compacte

Dans les grands dictionnaires de langue générale, les citations ont pour fonction, entre autres, de fournir des informations syntaxiques implicites. Par exemple le Grand Robert, en mentionnant la construction *blâmer quelqu'un de* (ou *pour*) *son attitude*, indique implicitement que le verbe *blâmer* admet un deuxième complément nominal. Le fait que le deuxième complément puisse être aussi de forme *Groupe Verbal à l'infinitif* est signalé implicitement dans l'exemple 11 : «*Ils l'avaient blâmé de s'attacher à une maîtresse* ».

Certaines lacunes lexicales s'expliquent tout autant par l'état des connaissances linguistiques à l'époque où le dictionnaire a été rédigé, que par le parti pris de compter sur l'intuition linguistique du lecteur. C'est à ce dernier que revient par exemple, à partir des deux mentions suivantes : *blâmer la conduite de quelqu'un* et *blâmer quelqu'un de* (ou *pour*) *son attitude*, d'inférer que la phrase « *Ils l'avaient blâmé de s'attacher à une maîtresse* » a pour variante : « *Ils avaient blâmé son attachement à une maîtresse* ».

Beaucoup de propriétés transformationnelles des prédicats et des arguments sont ainsi passées sous silence : nominalisation, adjectivation, passivation, interrogation, etc. On pourrait être tenté de penser qu'elles sont trop générales pour figurer dans les entrées d'un dictionnaire. Deux exemples montreront qu'il n'en est rien. Comparons les deux locutions verbales *prendre le départ* et *prendre la fuite* ; aucune règle générale ne permet de prédire les différences de comportement du prédicat nominal : (*prendre un départ rapide ; le départ que j'ai pris*) versus (* *prendre une fuite rapide ; * la fuite que j'ai prise*). L'autre exemple concerne la transformation de nominalisation du verbe *porter*. On constate qu'il y a deux formes nominalisées : *port* et *portage (à dos d'homme , 0)*. Mais en rester à cette seule information serait trompeur, car des contraintes distributionnelles limitent l'emploi de l'une et l'autre forme : ainsi, *porter* n'est nominalisable en *port* qu'à condition que certaines classes

bien circonscrites de noms concrets figurent en position de complément d'objet : noms d'armes, de parures et coiffures, et surtout noms de vêtements.

Mais il serait tout à fait redondant d'énumérer toutes les propriétés syntaxiques de chaque prédicat sans prendre en compte le fait qu'il existe des classes de prédicats partageant les mêmes propriétés syntaxiques. Par exemple, *blâmer* appartient à une classe syntactico-sémantique de prédicats à trois arguments qui, comme *complimenter, critiquer*, *louer*, *maudire,* ont en commun un certain nombre de propriétés transformationnelles comme celles qui viennent d'être mentionnées :

Ils l'avaient (blâmé, complimenté, critiqué, loué, maudit) de s'attacher à une maîtresse

Ils avaient (blâmé, complimenté, critiqué, loué, maudit) son attachement à une maîtresse

Voici un autre exemple qui concerne les prédicats nominaux : étant donné leur très grand nombre, il serait absurde de ne pas indiquer une fois pour toutes que leurs verbes supports ont la propriété d'être effaçables, à certaines conditions, au sein d'une construction relative (e.g. *le balayage qu'a fait Paul de la maison; le balayage de Paul ; le balayage de la maison par Paul*). La constatation de telles régularités syntaxiques évite qu'on ait à répéter un grand nombre de fois la même information ; il suffit qu'on indique la classe syntactico-sémantique à laquelle appartient tel ou tel prédicat. C'est ce type de présentation que nous qualifions de *compacte.*

1.3. Des informations sémantiques explicites

Certaines informations sémantiques, comme celles qui concernent l'aspect, sont d'un grand intérêt pour l'utilisateur d'un dictionnaire. Le fait, par exemple, de signaler que le mot *blâme*, comme tous les prédicats nominaux qui ont *donner* comme verbes supports, a un aspect ponctuel implique qu'on ne pourra pas lui attribuer de supports aspectuels comme

entamer ou *esquisser*, et qu'il refuse un modifieur comme *interminable* ou un complément de temps de forme *pendant Nom de temps*.

Voici un autre exemple d'information sémantique : le fait de qualifier d'*événement fortuit* un prédicat nominal comme *orage* implique la possibilité de lui appliquer non seulement les supports *il y a, avoir lieu,* mais aussi *se produire* (qui serait impossible avec un nom d'événement « organisé » comme *défilé militaire*).

Notre dernier exemple concerne une portion du vocabulaire médical. On peut montrer que les noms d'*infections* (e.g. *tuberculose, méningite*) ainsi que les noms de *symptômes* (e.g. *fièvre*) et d*'effets secondaires* (e.g. *réaction cutanée*) sont des prédicats nominaux d'*événements*. Il suffit de constater que leurs verbes-supports sont des verbes d'occurrence typiques des prédicats d'événements, comme *se produire, survenir, s'observer*, avec des variantes aspectuelles telles que *s'installer, présenter un pic, céder spontanément, être fréquent dans* :

> *Dans le traitement de cette affection, une forte fièvre (se produit, survient, s'observe, s'installe) fréquemment*

D'autre part, le fait que de tels prédicats puissent être sélectionnés par des prédicats comme *durer, se prolonger, évoluer, être bref, être chronique, être tenace* est une preuve qu'ils ont une aspect duratif.

Nou*s* montrerons plus loin que la notion de classe d'objets permet d'intégrer au dictionnaire ce type d'informations sémantiques en relation avec les propriétés distributionnelles et syntaxiques.

2. Redéfinition de la notion d'*emploi* d'une unité lexicale

Les finalités propres d'un dictionnaire informatisé imposent une redéfinition de la notion traditionnelle d'*emploi* d'une unité lexicale. Avant d'y venir, énonçons deux principes essentiels: l'unité syntaxique de base est la phrase, et le lexique doit offrir une description *intégrée* des unités lexicales. Ils s'appliquent plus ou moins librement à tous les dictionnaires ; mais dans le cas des dictionnaires électroniques, ce sont des impératifs catégoriques.

2.1. L'unité syntaxique de base est la phrase

La conception suivant laquelle l'unité syntaxique de base est la phrase est implicite dans tout dictionnaire de la langue générale. La plupart des prédicats étant polysémiques, il n'y a pas d'autre moyen, pour lever les ambiguïtés, que de stipuler quelles catégories d'arguments correspondent aux différents emplois. Soit par exemple le verbe *peser,* dans son emploi comme verbe transitif : tous les dictionnaires suggèrent qu'il sélectionne un sujet humain et distinguent entre un premier sous-emploi, dans lequel *peser* sélectionne un nom concret (e.g. *peser une enveloppe*) , et un deuxième sous-emploi, dans lequel *peser* sélectionne un nom abstrait (e.g. *peser des circonstances*). Définir l'emploi d'un prédicat (verbe, prédicat nominal, prédicat adjectival) en spécifiant la nature de son environnement, c'est un façon d'admettre que, même dans un dictionnaire, l'unité syntaxique de base n'est pas le mot, mais la phrase.

D'autre part, il apparaît qu'il est impossible d'indiquer les propriétés transformationnelles comme celles auxquelles nous avons fait allusion au paragraphe 1.2. sans se placer dans le cadre de la phrase. L'exemple de la transformation de nominalisation du verbe *porter* nous a montré que ses conditions ne peuvent être déterminées sans une définition précise des classes de noms pouvant figurer en position d'arguments. Un exemple de propriété transformationnelle relevant du niveau de la phrase, la passivation, serait encore plus probant. Soient les verbes transitifs *concerner* et *regarder* dans leur emploi avec un complément

humain : *Ceci me (concerne, regarde)*. Une de leurs différences d'emploi concerne la forme de la phrase dans laquelle il sont employés, puisque seul le premier est passivable : *Je suis concerné par ceci* versus * *Je suis regardé par ceci*.

Beaucoup d'informations de nature sémantique sur une unité lexicale particulière mettent en jeu tous les éléments de la phrase : déterminants, temps de la conjugaison, adverbes ou adjectifs, et compléments circonstanciels. C'est en particulier le cas des informations concernant l'aspect des prédicats, comme le montre l'exemple de l'analyse de l'aspect du prédicat *blâme*, que nous avons esquissée dans la section 1.3. On a vu dans le même paragraphe que le fait de qualifier *défilé militaire* de « nom d'événement organisé » implique, au niveau de la phrase, l'utilisation des supports *il y a* et *avoir lieu*, mais exclut celle du support *se produire*.

La nécessité de situer l'analyse lexicale dans le cadre de la phrase ne s'applique pas aux seuls prédicats ; il en va de même avec les arguments. En effet, bien des noms figurant en position d'argument sont, comme les prédicats, polysémiques. Là encore, on ne peut déterminer les différents emplois qu'en examinant l'environnement phrastique. Deux exemples simples suffiront. Le mot *livre* a un emploi concret quand il figure en position de complément de verbes comme *relier, assembler* ; il a un emploi abstrait quand il figure en position de complément de verbes comme *écrire, plagier*. Le nom *mètre* a un emploi concret (nom d'instrument de mesure) quand il figure en position de complément circonstanciel de moyen du verbe *mesurer (N0<h> ; N1<concret>)* ; mais c'est un nom abstrait (nom d'unité de mesure) quand il figure en position de deuxième argument d'un autre emploi du verbe *mesurer (N0<h><concret> ; N1: cardinal <unité de mesure>)*.

2.2. Le lexique doit offrir une description *intégrée* des unités lexicales

Par description intégrée, nous entendons une description qui combine différents aspects linguistiques, notamment la syntaxe, la sémantique et le lexique. L'impossibilité de séparer ces trois niveaux de l'analyse linguistique s'est toujours imposée aux lexicographes. Conçoit-on qu'on puisse traiter sérieusement des unités lexicales sans analyser leur sens et leur construction (ni également leur morphologie, leur prononciation, leur orthographe, etc.) ?

Prenons l'exemple d'un des emplois du verbe *commettre* (dans *commettre une faute, un délit, un crime*). La première chose à relever est que, dans cet emploi, *commettre* n'est pas un prédicat, mais un verbe support. Cela implique entre autres choses qu'il est effaçable au sein d'une construction relative ; mais il faudra affiner l'analyse en fonction des différents prédicats, comme le montrent ces différences de construction : *le crime d'Oswald* (*contre la personne de Kennedy,* 0) et *l'assassinat d'Oswald* (*par Ruby*, 0). Mais de quelles catégories de prédicats nominaux le verbe *commettre* est-il le support approprié ? De catégories de prédicats signifiant certaines fautes ou délits (sur la notion nouvelle de *support approprié,* voir [Gross, 1996]). Il est à remarquer que ces prédicats ont un aspect ponctuel, ce qui interdit l'emploi de certains auxiliaires aspectuels comme *commencer à* ou *continuer à* (il existe en revanche une variante aspectuelle terminative du support : il s'agit de *consommer*, qui est approprié aux prédicats signifiant des crimes). On peut encore relever que *commettre*, quand il est approprié à des prédicats signifiant de simples fautes comme *gaffe, faute de goût,* est une variante de *faire*, qui ne convient en revanche que maladroitement à des prédicats de délits (? *faire un vol,* ? *faire un assassinat*). Devant des prédicats de crimes, *commettre* a pour variante *perpétrer*. Il est donc nécessaire de subdiviser les prédicats nominaux dont *commettre* est le support en un certain nombre de classes. Ces classes de prédicats seront définies en extension. Elle feront à leur tour l'objet d'une analyse linguistique intégrée : des sous-classes seront déterminées en fonction du nombre des arguments, des propriétés transformationnelles, des particularités aspectuelles, etc.

On notera que toutes ces informations ont un haut degré d'évidence pour n'importe quel francophone. La tâche des lexicographes du LLI consiste à expliciter les informations qui font partie de la compétence linguistique et à les présenter de telle sorte qu'elles soient accessibles au traitement par machine.

2.3. La notion d'*emploi* d'une unité lexicale

La notion d'emploi est rien moins que nouvelle. Il ne s'agit ici que d'en renouveler la définition, qui découle de tout ce qui vient d'être exposé. Rappelons que séparer les emplois d'une unité lexicale est crucial en lexicographie, puisque ce n'est pas faire autre chose que de rendre compte des ambiguïtés.

L'emploi d'un prédicat (verbe, prédicat nominal, adjectif), c'est la relation qu'il entretient avec les différentes classes de noms qui peuvent figurer en telle ou telle position argumentale ; ces classes de noms sont définies aussi bien sémantiquement qu'en extension ; la combinaison du prédicat et des classes d'arguments est par ailleurs caractérisée par un certain nombre de propriétés aspectuelles et transformationnelles. Un prédicat morphologique aura autant d'emplois qu'il a de classes d'arguments corrélées à des propriétés transformationnelles spécifiques et à une actualisation aspectuelle propre.

Quant à l'emploi d'un nom-argument, c'est son appartenance à une classe donnée d'arguments, définie par sa classe sémantique, et associée à un certain ensemble de prédicats. Par exemple, l'emploi du nom *autorail* est caractérisé de la façon suivante : il appartient à la classe des noms de *moyens de transport en commun ferroviaires* ; cette classe est à son tour définie par le fait que les noms qui la composent ont, entre autres propriétés, celle d'être en position d'argument sujet de verbes tels que *desservir (une gare, une localité), être en gare, être à quai* (cf. [Gross, 1994]). Un nom-argument aura autant d'emplois qu'il existe de classes d'arguments qui le contiennent. Par exemple, il existe au moins deux emplois du

morphème *forêt*, puisqu'il appartient d'une part à la classe des noms collectifs d'*arbres* (reliée à des prédicats comme *jaunir, perdre ses feuilles*), d'autre part à la classe des noms de *lieux* (reliée à des prédicats comme *avoir une superficie de*, *être peuplé de*).

Les classes de noms que nous venons d'évoquer constituent soit des classes très générales correspondant à un trait syntactico-sémantique (c'est le cas des noms d'*arbres* et des noms de *lieux*), soit, comme les noms de *moyens de transport en commun ferroviaires*, des classes plus particulières appelées « classes d'objets ». Ces deux types de classes font l'objet des deux sections suivantes de cet article.

3. Redéfinition de la notion de trait syntactico-sémantique

La notion de *trait syntactico-sémantique*, comme celle de *classe d'objets* qui sera présentée en quatrième partie, sert à séparer les emplois tant des arguments que des prédicats, c'est-à-dire à lever les ambiguïtés. Les traits syntactico-sémantiques définissent les classes d'unités lexicales les plus générales, comme celles des *humains*, des *inanimés concrets*. On verra dans la partie 4 que les classes d'objets ne sont rien d'autre que les subdivisions de ces classes très générales. Les classes correspondant aux traits syntactico-sémantiques sont définies par leurs relations avec des verbes appelés *opérateurs généraux*. On mettra en évidence l'importance des traits, mais aussi leurs limites.

3.1. Traits syntactico-sémantiques

Les traits syntactico-sémantiques sont utilisés couramment dans de nombreuses disciplines de la linguistique, notamment en syntaxe transformationnelle (Noam Chomsky, Zellig Harris, Maurice Gross). Ils se révèlent indispensables. Ainsi, il est commode et efficace de préciser que le verbe *blâmer* sélectionne un humain en position de sujet et, en position de complément,

soit un humain, soit une action humaine, soit une qualité humaine. Mais on a pu leur reprocher ne pas avoir été jusqu'à présent définis linguistiquement et de reposer sur la seule intuition. Par exemple, les noms de *bruits* et d'*odeurs* ont pu passer pour des concrets parce que les réalités qu'ils signifient ont, comme les entités massives, des qualités sensibles. Or il apparaît que ces noms figurent avec des verbes qui, tels *provoquer, s'élever*, ne sauraient caractériser des noms inanimés concrets comme *brique* ou *caillou* ; du reste leur aptitude à admettre des modifieurs aspectuels comme *subit* ou *persistant*, ainsi que l'existence des variantes adjectivales morphologiquement reliées *bruyant* et *odorant,* indiquent assez que ces noms sont des prédicats nominaux et que ce ne sont pas des noms concrets, mais des noms d'événements.

De même, un nom comme *maison* pourrait passer pour un nom concret. Or, à y regarder de près, on constate qu'il refuse la plupart des prédicats qui sélectionnent les noms concrets, notamment les prédicats de *grandeurs* appartenant à la catégorie de la masse (*avoir un poids, une masse*), les prédicats de qualités sensibles en rapport avec la masse comme *être (lourd, léger, pesant, dense),* les prédicats de qualités sensibles en rapport avec le sens du toucher comme *être (dur, mou, moelleux, rugueux, lisse, raide*)*,* les prédicats d'action en rapport avec le toucher, comme *palper, toucher, tâter, caresser,* et le prédicat de mesure *peser* :

* *Je (pèse, palpe) une maison*

* *Cette maison est (lourde, lisse)*

* *Cette maison pèse mille tonnes*

Le prédicat *maison* n'est donc pas un nom concret. La place nous manque ici pour faire la démonstration, par la même méthode, qu'il appartient à la catégorie des noms locatifs.

C'est en utilisant de tels moyens qu'on a pu subdiviser l'ensemble des noms en dix catégories correspondant chacune à un trait syntactico-sémantique. Dans cette entreprise de

classement, la première tâche a été de déterminer, à partir de critères syntaxiques, quels noms sont de nature argumentale et quels noms sont de nature prédicative. Il a fallu ensuite, également à partir de critère syntaxiques, dresser la liste des traits permettant de subdiviser ces deux grandes classes. Dans l'état actuel de la recherche, cette liste comprend six traits pour les noms-arguments :

- humain non prédicatif
- animal
- végétal
- inanimé concret
- lieu
- temps

Quant aux prédicats nominaux, ils sont subdivisés au moyen des quatre traits suivants :

- humain prédicatif (e.g. *avare, vendeur de ...*)
- action
- état
- événement

3.2. Les opérateurs généraux : prédicats généraux et verbes supports généraux

La liste des traits correspond à un ensemble de catégories sémantiques et syntaxiques. Ce sont les catégories les plus générales. Elles sont, en vertu du principe qui a été énoncé au paragraphe 2.1., définies à partir de propriétés distributionnelles, syntaxiques et sémantiques repérables au niveau de la phrase. En particulier, les noms de chacune des différentes catégories ont en commun d'être en relation avec le même ensemble de verbes. Ces verbes sont appelés *opérateurs généraux*. Il y a deux sortes d'opérateurs généraux. Ceux qui définissent les traits syntactico-sémantiques des noms-arguments s'appellent *prédicats généraux* ; ceux qui définissent les traits des prédicats nominaux s'appellent *verbes supports généraux*.

Les *prédicats généraux* sont les prédicats qui sélectionnent la totalité des noms affectés d'un des six traits propres aux noms-arguments. Par exemple, les verbes *croître*, *pousser,*

s'étioler, mourir font partie des prédicats généraux des *végétaux*. Il est à noter que les prédicats généraux, parce qu'ils sélectionnent un très grand nombre de noms, ne peuvent être déterminés avec certitude qu'après un examen approfondi des différentes grandes classes de noms. Ainsi, il faudra encore beaucoup de recherches pour savoir avoir certitude quels sont les prédicats généraux des *inanimés concrets* et des *lieux*.

Passons aux *verbes supports généraux*, qui servent à la définition des traits *état, action* et *événement*. Il apparaît que *faire* est le verbe support général pour les actions, et qu'on a *avoir* ou *être* pour les états et *il y a* pour les événements (il s'agit ici d'un des emplois de *il y a*, non commutable avec *se trouver*). Mais l'excès de généralité de ces distinctions rend nécessaire l'élaboration de catégories plus étroites.

3.3. Limites de la notion de trait syntactico-sémantique

Rappelons qu'un outil théorique comme la notion de trait vise essentiellement à permettre la séparation des emplois des unités lexicales. Il y parvient dans un très grand nombre de cas. Par exemple, l'utilisation des traits *humain* et *inanimé concret* permet de séparer deux emplois du verbe *boire*:

N0 <hum> boire N1<inc> (*inc* = inanimé concret)

N0 <inc> boire N1<inc> (e.g. *L'éponge boit l'eau*)

On notera que la nécessité de séparer ces deux emplois de *boire* est confirmée par une propriété syntaxique affectant la forme du déterminant du nom complément :

Je bois (l'eau, de l'eau)

L'éponge boit l'eau

* *L'éponge boit de l'eau*

D'une manière générale, les traits sont très utiles pour séparer les emplois standard des unités lexicales de leurs emplois métaphoriques figés (catachrèses). Ils peuvent par ailleurs

aider à séparer nettement des emplois qui sont sémantiquement tout à fait distincts comme dans *filer un textile* et *filer un suspect*.

Mais les traits syntactico-sémantiques comportent aussi de graves inconvénients. Tout d'abord, la richesse de leur extension va évidemment de pair avec la pauvreté de l'information qu'ils fournissent. Cela les rend non pertinents dans les niveaux inférieurs de la classification. Inutile d'insister sur le fait que des descriptions comme « *N0 <hum> boire N1<inc>* » ou « *N0 <inc> boire N1<inc>* », si on les prenait pour des prescriptions strictes sur l'emploi des noms, autoriseraient une multitude de phrases inacceptables comme **Paul boit du caillou, * la table boit le fer, * le fer boit la table*.

Mais leur trop grande richesse entraîne également une certaine imprécision qui fait que, dans un certain nombre de cas, l'emploi des opérateurs généraux est déviant à des degrés divers. Cela fait que les opérateurs généraux doivent être considérés comme des approximations. Par exemple, nous l'avons vu, le verbe support général définissant les *actions* est *faire*. C'est incontestable du point de vue statistique. Mais, comme l'a mis en évidence Gaston Gross [Gross, 1996], « la notion d'action est trop générale pour être en mesure de prédire la forme requise du support ». Par exemple, le verbe support *faire* est maladroit avec les prédicats d'opérations industrielles et les prédicats de crimes (? *faire le démoulage d'une statue,* ? *faire un assassinat*). D'en d'autres cas, nous sommes en présence d'emplois-limites de *faire*, voire d'impropriétés, comme dans *?? faire une arrestation, * faire une gifle*. Il est clair que les prédicats *démoulage, assassinat* et *gifle* ont des verbes supports *appropriés* qui sont, respectivement, *procéder à, commettre* et *donner*.

Une sous-catégorisation des traits s'avère donc nécessaire. Elle est rendue possible par la notion de *classe d'objets*.

4. Les classes d'objets

Les classes d'objets sont un outil théorique nouveau qui a été élaboré par Gaston Gross. Elles peuvent être définies sommairement comme les sous-catégories des classes générales définies par les traits syntactico-sémantiques. Elles sont destinées à remédier à l'insuffisance des traits, qui fournissent des informations trop pauvres pour qu'on puisse avec eux séparer convenablement les emplois des unités lexicales. Cette notion est donc appropriée au traitement de la polysémie.

Un verbe hautement polysémique comme *prendre* constitue une bonne illustration de l'intérêt de la notion de classe d'objets (cf. [Gross, 1994]. En même temps, cet exemple montrera que les classes d'objets font à leur tour l'objet de subdivisions. Voici des phrases correspondant à quatre des emplois de *prendre* autres que celui de verbe support :

(1) *Nous avons pris l'autobus*

(2) *Nous avons pris la voiture*

(3) *Nous avons pris un steak*

(4) *Nous avons pris une bière*

Dans tous ces exemples, *prendre* sélectionne un nom affecté du trait *inanimé concret* en position de complément d'objet. Il est clair cependant qu'il illustrent plusieurs emplois différents.

Le verbe *prendre* est commutable, dans les phrases (1) et (2), avec des verbes comme *voyager en, aller quelque part en*. Cela suggère que *autobus* et *voiture* appartiennent à la classe d'objets des <moyens de transport>. Pourquoi dès lors séparer les emplois (1) et (2) ? Parce qu'*autobus* et *voiture* ne sont pas sélectionnés par les mêmes prédicats. Par exemple, on ne dira pas d'une voiture qu'on l'*emprunte* ou qu'elle *dessert* une localité. De telles restrictions distributionnelles, ainsi que des propriété syntaxiques sur lesquelles nous allons revenir, imposent un dégroupement : on subdivisera la classe d'objets des <moyens de

transport> en au moins deux sous-classes : les <moyens de transport en commun> et les <moyens de transport individuels>. D'autres prédicats permettront de subdiviser la classe des <moyens de transport en commun> en <moyens de transport en commun routiers>, qui, tel *autobus*, sont sélectionnés par des prédicats comme *rouler, déraper* ; en <moyens de transport en commun aériens> (*atterrir, décoller, ...*) ; en <moyens de transport en commun ferroviaires> (*être à quai, être en gare de*), etc.

Quant aux exemples (3) et (4), ils illustrent deux autres emplois de *prendre*. Le fait que le premier de ces emplois soit sur le même paradigme que *manger* alors que le deuxième est sur le paradigme de *boire,* impose leur séparation, leur dégroupage.

4.1. Les opérateurs appropriés : prédicats appropriés et verbes supports appropriés

Les exemples qui viennent d'être donnés montrent qu'une classe d'objets de noms-arguments est définie (non exclusivement) par un ensemble de prédicats qui sélectionnent les noms qui la composent. Ces prédicats sont appelés *prédicats appropriés*. Voici en exemple une liste d'opérateurs appropriés à la classe des moyens de transport collectifs (codée « *mt-c* »):

emprunter/N1:hum/N2:mt-c

manquer (le, Poss0)/N1:hum/N2:mt-c

partir par/N1:hum/N2:mt-c

prendre le dernier/N1:hum/N2:mt-c

prendre le premier/N1:hum/N2:mt-c

rater (le, Poss0)/N1:hum/N2:mt-c

revenir par/N1:hum/N2:mt-c

voyager par/N1:hum/N2:mt-c

Mais il est également possible de construire des classes d'objets de prédicats nominaux en prenant en considération la forme de leurs verbes supports. Sur la notion nouvelle de *verbe support approprié*, on se reportera à [Gross, 1996]. Comme on l'a vu dans le paragraphe 3.2., tous les noms pourvus du trait *action* n'admettent pas facilement le verbe support *faire*, et lui préfèrent des verbes supports particuliers. C'est à partir de ces verbes supports qu'on pourra subdiviser les *actions* en diverses classes d'objets : <opérations chirurgicales> (*pratiquer*), <crimes et délits> (*commettre*), <bruits vocaux> (*émettre, pousser*), <coups> (*donner*), etc.

Prenons un autre exemple (sommairement évoqué) : celui des prédicats nominaux pourvus du trait *événement* (voir à ce sujet [Gross et Kiefer, 1996]). On les subdivisera en différentes classes d'objets, suivant la forme de leurs verbes supports appropriés : les <événements fortuits> (*se produire*) et les <événements créés> (qui excluent le verbe support *se produire*), subdivisés eux-mêmes en <événements politiques>, <événements sociaux>, <spectacles> ; les <événements cycliques> (*tomber le Ndate*), etc.

Bien entendu, d'autres verbes que les verbes supports appropriés seront pris en compte pour construire les classes de prédicats nominaux. Par exemple les <événements créés> se caractérisent notamment par le fait qu'ils sont sélectionnés par des prédicats causatifs comme *provoquer, occasionner, être la cause de, être responsable de*. Quant aux <événements fortuits>, ils sont sélectionnés par des prédicats mettant en évidence l'importance du témoin, tels *être le témoin direct de, voir de ses propres yeux, être sur les lieux de, avoir entendu dire que*.

4.2. Rôle des propriétés sémantiques et syntaxiques dans la définition des classes d'objets

Nous venons de souligner l'importance de la prise en compte des prédicats appropriés et des verbes supports appropriés dans la définition des classes d'objets. Mais les classes d'objets sont aussi définies par des propriétés sémantiques et syntaxiques.

Par exemple, les <événements fortuits> dont il vient d'être question sont subdivisables, suivant des critères aspectuels, en <événements fortuits ponctuels> (*advenir, survenir, se faire, ...*) et en <événements fortuits duratifs> (*continuer, faire rage, ...*).

Des critères purement syntaxiques, voire morpho-syntaxiques, pourront permettre de séparer des emplois. Par exemple, la présence et l'absence respectives d'une forme nominalisée impose la séparation de deux emplois du verbe *prendre* qui ne sont pas sans rapport puisqu'ils sont hyponymiques de *ingérer, avaler* :

J'ai (avalé, pris) un médicament *la prise d'un médicament*

J'ai (avalé, pris) un steak ** la prise d'un steak*

On exprimera une partie de cette situation en termes de classes d'objets : un seul des deux emplois du verbe *prendre* hyponymiques de *ingérer* a une forme nominalisée ; c'est quand il sélectionne en position de complément un nom appartenant à la classe d'objets des <médicaments>.

Un autre exemple de définition d'une classe d'objets à partir de critères majoritairement syntaxiques sera pris au sein des noms-arguments possédant le trait *humain*. Ceux-ci sont subdivisés en 36 classes d'objets. Parmi celles-ci figure la classe d'objets des <défauts>. Elle se caractérise par sa syntaxe particulière, qui a été étudiée par Nicolas Ruwet et Jean-Claude Milner [Milner 1978] : les noms de qualités qui la composent figurent en position de

deuxième complément du verbe *traiter de* (e.g. *traiter quelqu'un de traître*), et ils entrent dans les structures d'insultes (*Traître ! Espèce de traître ! Traître que tu es ! Ce traître de Luc !*)

Le dernier exemple concerne les noms appartenant à la classe des <moyens de transport en commun> [Gross, 1996]. Ils ont une syntaxe qui les distingue des noms de la classe des <moyens de transport individuels>, comme le montrent les deux exemples suivants:

l'autobus de 8h47 ** la voiture de 8h47*

l'autobus de Marseille ** la voiture de Marseille*

4.3. Définition et intérêt des classes d'objets

Les exemples qui viennent d'être donnés montrent qu'une classe d'objets est un ensemble de noms qui partagent un ensemble de propriétés distributionnelles (ils peuvent figurer en position d'argument des mêmes prédicats), syntaxiques et, bien entendu, sémantiques. On aura noté que la théorie sémantique qui sous-tend cette notion est qu'il est impossible de déterminer le sens des mots sans prendre en compte leurs propriétés syntaxiques.

Voici, très sommairement, quelques applications de la notion de classe d'objets :

a) Les classes d'objets servent avant tout à séparer les différents emplois des prédicats. Nous reprenons l'exemple du verbe *prendre*, en nous limitant à cinq emplois :

prendre/N0:hum/N1:inc<nourriture>

prendre/N0:hum/N1:inc<boisson>

prendre/N0:hum/N1:inc<moyen de transport>

prendre/N0:hum/N1:inc<médicament>

prendre/N0:hum/N1:inc<voie>

b) Les classes d'objets rendent compte avec élégance des phénomènes de synonymie. Grâce aux classes d'objet, par exemple, on peut décrire les différents emplois d'adjectifs comme *piquant, rigoureux, vif*.

c) Les classes d'objets présentent de grands avantages dans le cadre de la construction d'un dictionnaire bilingue. Elles permettent notamment de rendre compte des idiotismes intraduisibles littéralement (e.g. *passer un contrat avec quelqu'un, négocier un virage, en appeler à quelqu'un*) et de proposer des traductions adaptées.

5. Un aperçu de la forme actuelle de la base de données lexicales du LLI

Nous entendons, dans cette section, donner au lecteur une représentation aussi concrète que possible de la forme de la base de données lexicales du LLI.

5.1. Aperçu de la description des classes d'objets

Rappelons la liste actuelle des traits syntactico-sémantiques:

Noms non prédicatifs	- *humain non prédicatif* - *animal* - *végétal* - *inanimé concret* - *lieu* - *temps*
Noms prédicatifs	- *humain prédicatif* (noms de qualités, déverbaux) - *action* - *état* - *événement*

Chacune des classes générales correspondant à ces traits sont subdivisées en classes d'objets. Voici un extrait de la subdivision des noms humains en classes d'objets :

adepte : protestant, taoïste, ...
âge : enfant, vieillard, ...
âge (collectif) : jeunesse, vieillesse, ...
appellatif : monsieur, sire, ...
collectif : foule, troupe, ...
défaut moral : menteur, voleur, ...
défaut physique : difforme, boiteux, ...
défaut psychologique : imbécile, ...
...
soldat : zouave, biffin, ...
spécialiste : juriste, spéléologue, ...
sportif : footballeur, escrimeur, ...
titre : duc, prince, ...

Voici encore un exemple ; il s'agit d'un extrait de la liste des classes d'objets de noms prédicatifs affectés du trait *état* :

défauts : bêtise, paresse, ...
états physiques : faiblesse, fatigue, ...
états psychologiques : excitation, sérénité, ...
maladies somatiques : tuberculose, grippe, ...
maladies psychiques : névrose, paranoïa, ...
sentiments : gaieté, tristesse, ...
...

5.2. Présentation schématique d'une classe

Chaque classe d'objets fait l'objet de six fichiers. Un premier fichier dresse la liste des noms communs, munis de leur trait syntactico-sémantique et de l'indication de la classe d'objets (ainsi que du domaine de connaissance). Le second fichier comprend les noms propres. Les trois fichiers suivants décrivent les prédicats appropriés : verbes, adjectifs et noms prédicatifs. Chacun de ces prédicats porte l'indication de la nature des arguments. Enfin, le sixième fichier recense les suites figées. Voici un extrait des fichiers se rapportant à la classe des <voies>:

Dictionnaire des noms communs

autoroute/G:nf/T:loc/C:voie/D:transp
avenue/G:nf/T:loc/C:voie/D:transp

route/G:nf/T:loc/C:voie/D:transp
... (*nf* = nom féminin ; *T:loc* = Trait:locatif ; *C* = Classe d'objets ; *D* = Domaine)

Dictionnaire des prédicats appropriés

Sujet : <voie>

conduire à/N0:loc<voie>/N1:loc
desservir/N0:loc<voie>/N1:loc
mener à/N0:loc<voie>/N1:loc
...

Sujet <humain> objet <voie>

bitumer/N0:hum/N1:loc<voie>
emprunter/N0:hum/N1:loc<voie>
goudronner/N0:hum/N1:loc<voie>
...

Dictionnaire des adjectifs appropriés
boueux/N0:loc<voie>/D:transp
cahoteux/N0:loc<voie>/D:transp
congestionné/N0:loc<voie>/D:transp
...

Dictionnaire des prédicats nominaux appropriés
asphaltage/N0:hum/N1:loc<voie>/W:procéder à (*W* = verbe support)
goudronnage/N0:hum/N1:loc<voie>/W:procéder à
pavage/N0:hum/N1:loc<voie>/W:procéder à
...

Dictionnaire des suites figées
rebrousser chemin
se frayer un chemin
tous les chemins mènent à Rome
...

5.3. Les champs

Les exemples simplifiés ci-dessus n'enregistrent pas la totalité des champs de chaque entrée de la base de données lexicales. Nous nous en sommes tenu aux principaux : ceux qui sont relatifs aux traits (champ *T*), à la classe d'objet (champ *C*), à la position les arguments (*N0, N1*,...), au verbe support (champ *W*), à la catégorie grammaticale (champ *G*), et au

domaine de connaissance (champ *D*). A propos du champ *D* (domaine), on notera que le LLI a élaboré, par une comparaison systématique de tous les dictionnaires disponibles, une liste d'environ 500 domaines, avec le souci de montrer leur importance dans la levée de l'ambiguïté (cf. [Mathieu-Colas, 1994]).

Il existe en outre un champ *F*, qui décrit les variations morphologiques induites par la flexion; un champ *V*, qui rend compte des variantes graphiques ; un champ *R*, qui note les registres de langue ; un champ *M* qui, dans le cas des locutions verbales figées, présente la suite verbale en termes de catégories grammaticales (e.g. *V Prép N* pour « *tirer sur la ficelle* »).

D'autres champs (respectivement *S* et *A*) notent la synonymie et l'antonymie. On trouvera également un ou plusieurs champs réservés à la traduction dans une langue étrangère. Des projets de dictionnaires multilingues sont en effet à l'étude.

Il existe enfin, dans la description de chaque prédicat, un champ *TR* qui précise si l'emploi verbal en question relève ou non des grandes transformations (nominalisation, adjectivation, passivation, interrogation totale, etc.).

Conclusion

Nous nous sommes borné dans cet article à présenter la forme générale de la banque de données lexicales du LLI, en insistant sur les choix théoriques et méthodologiques dont dépend sa mise en oeuvre. Nous avons passé sous silence un certain nombre d'autres applications de la notion de classe d'objets, dans le domaine de l'analyse des noms composés, de la syntaxe des prépositions, des connecteurs et des phénomènes anaphoriques, dans le domaine aussi de l'étude de la synonymie et de l'antonymie. La diversité des domaines mentionnés indique assez à quel point la notion de classes d'objets est féconde.

Soulignons pour terminer ce qui caractérise la base informatisées de données linguistiques du LLI. C'est l'ambition de rassembler de façon combinée un maximum d'informations linguistiques sur chacune des entrées. Ces informations sont, pour l'essentiel, de nature distributionnelle, syntaxique et sémantique. D'une part, le dictionnaire du LLI partage avec celui du LADL (voir par exemple [Gross (M.), 1975] la caractéristique d'être un « lexique-grammaire » soucieux d'énumérer les propriétés syntaxiques, notamment transformationnelles, des entrées dans le cadre de la phrase. D'autre part, c'est aussi un dictionnaire qui entend énumérer les propriétés sémantiques des entrées en les reliant aux autres propriétés, notamment distributionnelles et syntaxiques. On pourrait dire, en reprenant une expression de Michel Mathieu-Colas, que c'est un dictionnaire « sémo-syntaxique ». C'est cette volonté d'*intégration* de toutes les propriétés linguistiques des entrées qui légitime principalement la forme qui a été donnée à la base de données linguistiques du LLI. D'autres aspects de cette base de données, comme son caractère explicite et extensif, trouvent leur légitimité dans la finalité même du projet, qui est de servir de support à des dictionnaires de formats divers, et dans la nature informatique du support.

REFERENCES

Gross (G.)

1994, « Classes d'objets et description des verbes » in *Langages 115*, Paris, Larousse

1996, « Prédicats nominaux et compatibilité aspectuelle » in *Langages 121*, Paris, Larousse

Gross (G.) et Kiefer (F.)

1996, « La structure événementielle des événements », in *Folia Linguistica*

Gross (M.)

1975, *Méthodes en syntaxe*, Paris Hermann

Mathieu-Colas (M.)

1994, *Les mots à traits d'union,* Paris, Didier

Milner (J.-C.)

1978, *De la syntaxe à l'interprétation*, Paris, Editions du Seuil

UN NOUVEL OUTIL THEORIQUE EN LEXICOGRAPHIE : LES CLASSES D'OBJETS

Denis Le Pesant

RESUME

Cet article vise à décrire la forme de la banque informatisée de données lexicales qu'élabore l'équipe *Lexique et Grammaire* du LLI (Laboratoire de Linguistique Informatique : CNRS, UMR 195). Il s'agit d'un dictionnaire explicite et extensif. Chaque entrée fournit de façon combinée un maximum d'informations linguistiques, notamment syntaxiques, distributionnelles et sémantiques. Les unités lexicales sont regroupées dans des classes d'un type nouveau appelées *classes d'objets*.

THE *CLASSES OF OBJECTS* A NEW THEORETICAL TOOL IN LEXICOGRAPHY

Denis Le Pesant

ABSTRACT

We intend in this article to describe the form of the lexical data base developped by the LLI (Laboratoire de Linguistique Informatique : CNRS, UMR 195). Each entry of that explicit and extensive dictionary provides the utmost wide range of integrated linguistic data, especially syntactic, distributional and semantic data, and the lexical units are put together in a new type of syntactico-semantic classes called *classes of objects*.