

Quels dictionnaires pour l'étiquetage sémantique ?

Jean VERONIS

1. Introduction

L'étiquetage sémantique consiste à accoler à chaque mot d'un texte une étiquette correspondant au « sens » (je m'expliquerai plus loin sur la présence de ces guillemets) qu'a ce mot dans le contexte particulier où il apparaît. La tâche paraît conceptuellement simple : il semble qu'il suffise de se fixer un dictionnaire de référence, et pour chacune des occurrences de tel ou tel mot, de lui affecter le numéro du « sens » concerné dans le dictionnaire. Robert Martin (2001 ; voir aussi l'article dans ce numéro) décrit bien, à l'aide d'une simulation manuelle sur des exemples, le type de connaissances et de raisonnements qui sont mobilisés pour une telle opération. Automatiser la tâche réellement sur du texte courant n'est cependant pas une affaire triviale, et les informaticiens s'y essaient depuis plusieurs décennies avec des succès mitigés (on pourra se faire une idée de l'état de l'art dans Ide & Véronis, 1998 ; voir aussi Stevenson & Wilks, 2001 pour des résultats récents). Je ne m'intéresserai pas ici aux aspects algorithmiques de la question, car les problèmes linguistiques qui se posent en amont me semblent être suffisamment délicats pour reléguer les problèmes d'automatisation au second plan.

Je relaterai en effet tout d'abord deux expériences qui montrent que le jugement d'informateurs sur le caractère polysémique des mots est extrêmement peu fiable et que leur capacité à effectuer une telle tâche d'étiquetage lexical est tout à fait médiocre. Il semble donc peu raisonnable, dans l'état actuel des choses, de demander à une machine de faire mieux que des informateurs humains, et il me paraît préférable d'essayer de comprendre les causes d'une telle contre-performance.

Dans une deuxième moitié de cet article, j'essaierai de montrer que c'est le dictionnaire lui-même qui est la cause principale de la difficulté, et j'essaierai de tracer quelques pistes pour une nouvelle approche du travail lexicographique, qui peut à la fois systématiser la représentation de l'information lexicale dans le dictionnaire, et contribuer à la constitution de ressources utilisables pour la désambiguïsation lexicale automatique, et donc à l'étiquetage de textes.

2. Deux expériences sur la fiabilité du jugement humain

2.1. Jugements de polysémie

Dans une première expérience, j'ai extrait les occurrences de 600 mots (200 adjectifs, 200 noms et 200 verbes) d'un corpus d'un million de mots, constitué de questions écrites des parlementaires à la Commission Européenne. Ce corpus regroupe des textes brefs sur les thèmes les plus variés : environnement, santé, éducation, économie, etc. Les mots ont été choisis sur un simple critère de fréquence, de façon que chacun des mots retenus ait environ une soixantaine d'occurrences dans le corpus, sans faire d'hypothèse préalable sur leur caractère polysémique ou non. Ce sont des mots familiers : *abattage, abord, accompagnement, acier, aéroport, alimentation, appréciation, approbation*, etc. Au total, les mots concernés représentaient environ 36 000 occurrences.

Pour chaque mot, la soixantaine de contextes dans lesquels il figurait a été imprimée sur une page séparée, sous forme de lignes de concordances, le mot-cible apparaissant en gras dans une colonne centrale. Le matériel de l'expérience consistait donc en 600 pages au total, reliées en trois volumes : adjectifs, noms et verbes.

Un exemplaire de ce matériel a été distribué à six informateurs, choisis parmi les étudiants de maîtrise de sciences du langage à l'Université de Provence, qui avaient pour consigne de répondre pour chaque mot à la question suivante : « Ce mot a-t-il un ou plusieurs sens dans les contextes ci-dessous ? ». Les informateurs devaient cocher une des trois cases « oui », « non » ou « je ne sais pas ». Ils avaient une semaine pour s'acquitter de la tâche au rythme qu'ils choisissaient et percevaient une petite rémunération. Sans avoir de formation spécifique en lexicographie, tous avaient eu des cours de sémantique, avec une attention particulière aux notions de polysémie et d'homographie.

De façon peut-être un peu inattendue, tous les informateurs ont trouvé la tâche facile, bien qu'évidemment fastidieuse. Cette facilité est confirmée par le taux très faible de réponses du type « je ne sais pas » (4%), dont il avait

été bien souligné qu'il pouvait être fait usage à volonté. La plupart des mots (73%) ont été jugés comme n'ayant qu'un seul sens, avec des différences statistiquement significatives entre les catégories: les noms ont été jugés plus polysémiques que les verbes, eux-mêmes jugés plus polysémiques que les adjectifs.

Malgré la simplicité perçue de la tâche et la bonne confiance de chaque informateur dans son jugement, l'accord moyen entre paires d'annotateurs s'est avéré faible : on ne relevait que 49% d'accord une fois soustrait l'effet du hasard¹, toutes catégories confondues. L'accord était un peu meilleur pour les adjectifs, jugés moins polysémiques (67%), mais moins bon pour les noms (36%) et les verbes (37%). De telles valeurs sont généralement considérées dans la littérature comme révélatrices d'un désaccord important (Krippendorff, 1980).

2.2. Etiquetage lexical

Les 60 mots (20 adjectifs, 20 noms, 20 verbes) qui ont été globalement jugés comme les plus polysémiques par l'ensemble des six informateurs ont été sélectionnés pour servir de base à une deuxième expérience. Il a été demandé à six autres informateurs, également étudiants en maîtrise de sciences du langage, d'étiqueter chacune des 3 724 occurrences de ces mots dans le corpus à l'aide des numéros de sens du *Petit Larousse*. Les occurrences ont été intégrées dans le tableur Excel avec leurs contextes gauches et droits dans la limite du paragraphe, ainsi qu'une colonne supplémentaire permettant à l'annotateur d'entrer le sens voulu. Si l'annotateur jugeait que plusieurs sens pouvaient correspondre à un contexte particulier, il pouvait les entrer tous ; si par contre il jugeait qu'aucun sens ne convenait, il pouvait le signifier à l'aide d'un point d'interrogation. Comme précédemment, la tâche pouvait être réalisée en temps libre dans la limite d'une semaine, et les annotateurs percevaient une petite rémunération.

A nouveau, la tâche n'a pas été perçue comme difficile, et les annotateurs avaient une bonne confiance dans leurs réponses, ce qui est confirmé par le très faible taux (2%) de réponses multiples qui auraient pu révéler des hésitations. Pourtant, l'accord moyen entre paires d'annotateurs était très médiocre : 41% pour les verbes et les adjectifs, 46% pour les noms (une fois soustrait l'effet du hasard). Pour certains mots (par exemple *correct*, *historique*, *utile*, *communication*, *degré*, *lancement*, *station*), le résultat était même virtuellement indiscernable de réponses au hasard.

On pourrait penser que les divergences n'affectaient que les subdivisions mineures entre sens. J'ai donc refait les calculs en réduisant les étiquettes fournies par les annotateurs aux seules divisions de plus haut niveau dans la hiérarchie des entrées. L'amélioration observée était insignifiante pour les adjectifs et les verbes (46%) et un peu plus élevée pour les noms, qui n'atteignent malgré tout que 60% d'accord non imputable au hasard. Le désaccord observé entre annotateurs affecte donc bel et bien les grandes divisions de sens fournies par le dictionnaire, et pas seulement des différences de détail.

3. Discussion

La première expérience a montré que le jugement des informateurs diffère de façon importante quant au statut, polysémique ou non, d'un mot dans un corpus donné. La deuxième indique qu'ils sont aussi très largement en désaccord lorsqu'il s'agit d'étiqueter les occurrences d'un mot en contexte à l'aide d'un dictionnaire standard — au point que pour certains mots, il n'y avait pas plus d'accord que si les informateurs avaient répondu totalement au hasard.

Ces résultats me semblent jeter un nouveau jour sur l'étiquetage lexical des « sens » dans les corpus. On ne peut, à mon sens, incriminer le dictionnaire utilisé (le *Petit Larousse*). Ce dictionnaire a été choisi car il constitue en quelque sorte l'archétype du dictionnaire : produit de grande consommation, il est censé pouvoir être utilisé de façon correcte par tout locuteur raisonnablement cultivé, et particulièrement par des étudiants. D'autres choix auraient été possibles, comme celui du *Trésor de la Langue Française*, par exemple, mais les entrées longues et complexes auraient pu être source de difficulté (plusieurs pages pour certains des mots du corpus), et l'on aurait pu attribuer la variabilité des réponses à la difficulté d'analyse des entrées. Au demeurant, tous les dictionnaires que j'ai pu consulter présentent des caractéristiques analogues, qui conduiraient très probablement au même type de difficultés.

Le mot *degré*, qui a été très mal étiqueté par les annotateurs, illustre très clairement le type de problème auquel ils ont été confrontés. Au sommet de la hiérarchie, le *Petit Larousse* présente les divisions suivantes :

¹ Deux personnes qui répondent au hasard à un questionnaire ont en effet des chances non négligeables de tomber d'accord par accident. Le coefficient κ que nous utilisons permet de mesurer la part d'accord non imputable au seul hasard (Cohen, 1960).

DEGRE n.m. **I.** Litt. Marche d'un escalier. [...] **II.** Chacun des états intermédiaires pouvant conduire d'un état à un autre. [...] **III.** Intensité relative (d'un état affectif, moral ou pathologique). [...] **IV.** Chacune des divisions, correspondant à l'unité, d'une échelle de mesure. [...]

Les divisions I et IV ne posent pas de grandes difficultés, mais la distinction entre II et III est source de confusion. Prenons par exemple les deux phrases suivantes :

(1) ...les trois principaux **degrés** de cette élimination étatique: le génocide, la déportation en masse et l'assimilation forcée...

(2) Ils s'inquiètent de ce qu'ils perçoivent comme un **degré** croissant d'anarchie...

Il est extrêmement difficile de déterminer dans ces deux exemples si le mot *degré* réfère à un « état intermédiaire pouvant conduire d'un état à un autre » ou bien à « l'intensité relative d'un état affectif, moral ou pathologique ». Le lexicographe avait peut-être une logique derrière cette subdivision, mais l'entrée ne livre aucun indice permettant au lecteur de la retrouver.

Il serait pourtant simple, dans ce cas, de classer les usages sur la base de critères syntaxiques. Un premier ensemble d'usages du mot degré accepte la détermination par des cardinaux (*un, deux, trois*, etc.) ou des ordinaux (*premier, second, dernier*, etc.), tandis qu'un autre ensemble, disjoint, accepte les intensifieurs du type *fort/faible* (les autres adjectifs du paradigme sont *alarmant, élevé, minimal, différent, croissant*, etc.). En d'autres termes, le premier ensemble d'usages est comptable, le second intensifiable. On voit immédiatement que la phrase (1) se rattache au premier ensemble, tandis que la phrase (2) relève du second. J'ai testé ce critère sur plusieurs centaines d'exemples extraits de divers corpus, et j'ai rencontré très peu de difficultés. Pourtant, aucun des dictionnaires que j'ai pu consulter ne donne cette propriété simple. On pourrait penser qu'elle était la clé cachée des divisions II et III faites par le lexicographe dans le *Petit Larousse* (« chacun » renvoie au comptable, « intensité » renvoie à l'intensifiable), mais le contenu des divisions contredit cette interprétation (on trouve ainsi sous III les exemples *brûlure au premier, deuxième, troisième degré*).

Il est toujours facile de pointer du doigt les faiblesses ou les erreurs de tel ou tel dictionnaire. C'est d'ailleurs un jeu dont les lexicographes eux-mêmes sont friands. Cependant, ma critique est ici d'une autre nature : je ne suis pas en train de dénoncer des erreurs occasionnelles ; je remets en cause le style et l'organisation même des entrées. Dans la quasi totalité des 60 mots utilisés dans la deuxième expérience, les entrées ne contiennent pas suffisamment d'indices de surface pour permettre aux annotateurs de mettre en correspondance tous les contextes avec un sens particulier de façon fiable. Pire, la division même des entrées ne prend que rarement en compte les contraintes distributionnelles du type de celles que j'ai mentionnées ci-dessus — et est en fait très souvent en contradiction avec ces contraintes. Il résulte de ce manque d'ancrage sur les indices et propriétés distributionnelles des mots un caractère vague de nombreuses définitions, qui est particulièrement apparent dans des mots abstraits et hautement polysémiques tels que *degré, économie, communication, formation*, etc., qui constituent une part importante de nombreux textes.

La tradition lexicographique prend ses racines dans une approche du sens et de la définition qui remonte à Aristote. Pendant plusieurs siècles, les dictionnaires ont essayé de décrire le « sens » des mots plutôt que leurs *usages* (à part quelques indications occasionnelles sur le registre ou le domaine). Ce n'est que très récemment, à la suite du travail pionnier de Hornby (1942, 1954) que certains dictionnaires (par exemple le *Oxford Advanced Learner's Dictionary*², le *Cobuild* ou le *Longman Dictionary of Contemporary English*) ont commencé à incorporer de façon systématique des informations syntaxiques, collocationnelles ou paradigmatiques, en se basant sur l'examen systématique de corpus plutôt que sur l'introspection du lexicographe. En ce qui concerne le français, le *Dictionnaire du Français Contemporain (DFC)* de Jean Dubois a commencé à aller dans cette direction dans les années 1960, en essayant d'intégrer les schémas de valence (nombre et nature des compléments) dans la division ou le « dégroupement » (c'est-à-dire l'éclatement en homographes) des entrées. Toutefois, le projet n'a pas bénéficié de l'étude de corpus informatisés qui auraient pu rendre l'entreprise plus systématique. Le *Trésor de la Langue Française (TLF)* a quant à lui été basé sur l'étude d'un grand corpus (essentiellement littéraire), et il fournit également quelques indications sur la valence. On est toutefois très loin d'un relevé systématique des propriétés distributionnelles des lexies (collocations, etc.). Il est vrai que jusqu'à une date assez récente on ne disposait guère de grands corpus, ni des outils informatiques permettant recherches et traitements complexes. Curieusement, alors que ces ressources commencent à apparaître, les projets lexicographiques semblent au point mort : le *DFC* et le *TLF* sont restés sans suite. Le *Dictionnaire Explicatif et Combinatoire* de Mel'čuk et son équipe (voir Mel'čuk, Clas &

² qui a fait suite à *The Idiomatic and Syntactic English Dictionary* publié par Hornby dès 1942.

Polguère, 1995) semble être à l'heure actuelle le seul projet de description approfondie des propriétés distributionnelles (« combinatoires » pour employer la terminologie des auteurs) des lexies. Malheureusement, les auteurs ont opté pour un parti pris d'introspection et rejettent l'approche empirique sur corpus (le corpus ne sert qu'à posteriori, à titre de vérification), ce qui semble très dommageable pour l'exhaustivité et la fiabilité du relevé.

Je suis convaincu qu'une rupture radicale doit être opérée avec la lexicographique traditionnelle si l'on souhaite progresser dans la question de l'étiquetage lexical, ainsi que dans toutes les branches du traitement automatique des langues qui font intervenir le « sens des mots ». Il nous faut accepter de changer de paradigme, et de passer de la description des « sens » à celles des *usages*. Les dictionnaires mentionnés plus haut amorcent un pas timide dans cette direction, mais l'information distributionnelle qu'ils fournissent reste pour la plus grande partie un ajout à une base traditionnelle. J'essaierai de montrer à travers un exemple dans la section suivante que l'information distributionnelle peut constituer le fondement même de l'organisation du dictionnaire : les entrées peuvent être divisées en *classes d'usage* cohérentes sur la seule base de cette information, sans recours à l'analyse du « sens », et aux considérations subjectives qu'une telle analyse requiert.

Certes, un certain courant structuraliste a pu donner un temps l'illusion que le « sens » pouvait faire l'objet d'une étude scientifique. On pourrait ainsi décomposer le « sens » des mots en « primitives » ou en « sèmes », qui s'opposeraient sur le modèle des paires minimales qui a fait recette en phonologie. Des voix n'ont pas manqué de souligner le caractère totalement ad hoc des « sèmes » : outre qu'on n'a jamais pu formuler le moindre critère formel pour leur élaboration (Todorov, 1966), on s'aperçoit très vite qu'il est nécessaire de postuler plus de sèmes que de mots, ce qui sape le projet componentiel à la base (Eco, 1976). On pourrait sans doute prolonger la discussion, mais l'histoire a déjà tranché : en une quarantaine d'années, l'analyse en « sèmes » n'a guère dépassé la description d'ensembles d'objets spécifiques tels que les sièges ou les habitations, ou l'illustration d'exemples soigneusement choisis dans les manuels de linguistique.

Le point de vue selon lequel l'étude du « sens » des mots devrait être remplacée par celle de leurs usages (c'est-à-dire de leurs propriétés distributionnelles) n'est bien sûr pas totalement nouveau. On en trouve la trace chez Meillet (1926) : « Le sens d'un mot ne se laisse définir que par une moyenne entre [ses] emplois linguistiques ». Wittgenstein (1953) a défendu une position analogue dans les *Philosophische Untersuchungen*, et Harris (1954 : 155-158) l'a adoptée dans son programme linguistique en définissant le sens comme une fonction de la distribution (« meaning as a function of distribution »). Pour l'instant, elle n'a pas fait toutefois l'objet d'une incorporation systématique dans des projets lexicographiques de grande envergure.

4. Un exemple: le mot *barrage*

Dans cette section, je voudrais faire la démonstration que les entrées peuvent être subdivisées et organisées sans recours à l'analyse du « sens », mais grâce un examen systématique des propriétés distributionnelles relevées en corpus. En même temps, le relevé de ces propriétés distributionnelles est une ressource de première importance pour l'étiquetage lexical des corpus. J'utiliserai le vocable *barrage* comme exemple, qui a le mérite tout en restant assez simple pour nos contraintes d'espace, de présenter un éventail intéressant des types d'informations utilisables. Il va de soi que la plupart de ces informations ont été utilisées, de façon implicite ou explicite, par les lexicographes (voir par exemple Moon, 1987 ; Mel'čuk *et al.*, 1995 : 59 sqq.). C'est leur utilisation planifiée et systématique, sans tentative d'analyse du sens, qui fait l'originalité de la présente proposition. L'analyse ci-après se base sur l'examen d'environ 10 000 occurrences du mot *barrage* sur Internet et dans des corpus variés. Elle est volontairement simplifiée, par souci de concision, et de multiples usages (tels que *tir de barrage*, *match de barrage* ou *barrage de guitare* ont été sacrifiés).

4.1. Dérivation

Un premier critère de subdivision nous est fourni par la morphologie dérivationnelle : seuls certains usages de *barrage* sont des nominalisations *actives* du verbe *barrer*. Par *active*, je veux signifier que ces nominalisations ne sont pas seulement la trace étymologique d'une création lexicale, mais bien d'une transformation productive. Le nom est alors strictement interchangeable avec le verbe, à une modification de la structure syntaxique près. C'est le cas dans la phrase suivante :

(3) Après le **barrage** d'une route par les habitants, les 8 et 9 juin, l'ex-gouverneur de Chubul, Carlos Maestro, et le ministre de gouvernement de l'époque et aujourd'hui gouverneur, José Luis Lizurume, avaient signé un accord dans lequel ils accédaient à une partie des réclamations des manifestants. (*Web*)

L'énoncé peut faire l'objet d'une *reformulation* telle que :

(4) Après que les habitants **aient barré** la route, les 8 et 9 juin, l'ex-gouverneur ...

ou bien, sous forme de deux phrases successives :

(5) Les habitants **avaient barré** la route les 8 et 9 juin. L'ex-gouverneur ... et le ministre ... avaient alors signé un accord ...

La valence du verbe est exactement conservée, avec la réorganisation syntaxique classique :

le barrage de X par Y ↔ Y barre X

Ce cas est à distinguer des nominalisations *figées*, qui n'ont plus de valeur qu'étymologique, comme dans l'exemple suivant :

(6) Un groupe terroriste a dressé un faux **barrage** sur la route reliant Mascara à Mohammedia, interceptant une Renault Express et un fourgon Peugeot J-5 transportant des voyageurs. (*Web*)

Il est intéressant de remarquer que l'emploi en nominalisation active, qui est celui fourni en premier par tous les dictionnaires (« action de barrer »), est extrêmement rare. Sur quelque 10 000 occurrences du vocable *barrage*, je n'en ai trouvé qu'une seule instance.

4.2. Propriétés syntaxiques

Les usages qui ne constituent pas une nominalisation active peuvent à nouveau se subdiviser en fonction de leur valence. L'information sur la valence est évidemment bien plus productive pour les verbes, mais elle nous permet ici de discriminer une première sous-classe qui sélectionne un complément introduit par *sur* :

(7) ... les deux pays comprennent leur intérêt à régler l'affaire du **barrage sur le Danube** (*Chirac*)

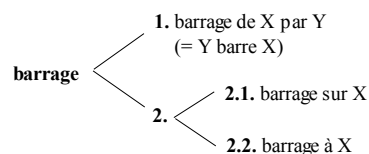
(8) Selon le commandant de la police Shahar Ayalon, la police a arrêté une voiture avant un **barrage sur l'autoroute** reliant Jérusalem à la vallée du Jourdain. (*Web*)

et une autre qui sélectionne un complément introduit par *à* (ou *contre*) :

(9) ... que M. Kim Jong-il impose son pouvoir (évolutif), seul **barrage à un chaos généralisé** (*Monde Diplo.*)

(10) ... où l'islamisme, hier considéré à Washington comme un **barrage aux révolutions**, conduit maintenant l'immense cortège des déceptions, des frustrations, des humiliations ... (*Monde Diplo.*)

Le schéma de l'entrée peut alors être révisé de la façon suivante :



Aucun dictionnaire ne mentionne de façon explicite cette propriété simple, qui constitue de plus un indice utilisable par des programmes d'étiquetage automatique.

La validité de la subdivision est confirmée par une autre propriété syntaxique, celle de la combinaison avec un verbe support (*faire*) avec chute de la détermination, qui ne s'applique qu'à la classe 2.2 :

(11) ... seul M. Kim Jong-Il peut **faire barrage** à un chaos généralisé

(12) ... on considèrerait que l'islamisme **ferait barrage** aux révolutions ...

Aucun des dictionnaires que j'ai consultés, pas même le *TLF*, ne mentionne cette construction, pourtant très fréquente. Elle constitue elle aussi un indice de choix pour la désambiguïsation automatique.

4.3. Information paradigmatic

Un autre type d'information est de nature paradigmatic. Ainsi, une partie seulement des usages de *barrage* accepte un véritable hyperonyme, *ouvrage*. Cette affirmation peut sembler étrange dans la mesure où une tradition qui remonte à Aristote (et la mode récente des « ontologies ») peut laisser penser que tous les mots, et tous leurs « sens », possèdent un hyperonyme, et que l'espace lexical est organisé en une gigantesque taxinomie. Tout dépend, bien sûr, de ce que l'on souhaite appeler hyperonyme. Dans la perspective distributionnelle que je défends ici, j'adopterai une vue très stricte, qui restreint cette notion aux seuls cas corroborés par une réalisation syntagmatique observable, par exemple par des anaphores ou des énumérations :

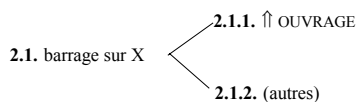
- (13) le **barrage** d'Assouan ... **cet ouvrage** géant, monstrueux (*Monde Diplo.*)
- (14) Ainsi la Tennessee Valley Authority... annonçait en mai sa décision de démolir le **barrage** Columbia pour protéger une moule menacée par **cet ouvrage** (*Web*)
- (15) la tutelle ... met en avant l'argument du mauvais suivi et de contrôle des projets et des études des **ouvrages comme les barrages** (*Web*)
- (16) les **ouvrages** "lourds" du GAP, **comme le barrage** Atatürk ou les tunnels jumeaux d'Urfa... (*Monde Diplo.*)
- (17) L'article 2 contient des dispositions relatives au marquage de la frontière sur les **barrages**, ponts **et autres ouvrages** fixes situés sur le parcours concerné du fleuve. (*Sénat*)

Ces attestations sont très fréquentes pour le couple *barrage* (hydraulique) / *ouvrage*. Malgré l'examen minutieux de mes 10 000 exemples, je n'ai pu trouver aucune réalisation équivalente pour les autres usages de *barrage*, tels que les barrages de police ou la classe 2.2 (*barrage à quelque chose*). Bien sûr, on peut tenter d'appliquer les tests classiques, et affirmer qu'un barrage de police est une sorte de barrière, d'obstacle, ou bien qu'un barrage à quelque chose est une sorte d'action, mais l'on est ici au niveau des inférences, et non des relations lexicales : aucune attestation en corpus ne corrobore ces relations. Les dictionnaires donnent le plus souvent un « hyperonyme » comme tête de la définition, mais rien ne permet de distinguer les hyperonymes vrais des autres, comme le montre cet exemple du *TLF* :

B. -- P. méton. **Barrière, obstacle** qui ferme un passage. *Barrage de police; forcer un barrage ...*

2. TRAV. PUBL. **Ouvrage** construit sur un cours d'eau ...

Le comportement différencié du mot *barrage* vis à vis de l'hyperonymie permet en tous cas de subdiviser la classe 2.1 (*barrage sur*):



D'autres types d'information paradigmatic peuvent être utilisés, comme par exemple la présence ou l'absence de synonymes. Je restreindrai cette notion aux synonymes stricts, qui permettent une reformulation exacte dans un contexte donné. La substituabilité peut être vérifiée en s'assurant que les candidats synonymes ont un comportement identique en termes de distribution (valence, verbes support, détermination, etc.). Dans le cas de *barrage*, seule la classe 2.2 de *barrage* (*barrage à*) accepte un synonyme strict, *obstacle* :

- (18) la volonté de faire **barrage** (=obstacle) à une probable expansion du communisme (*Monde Diplo.*)

Ceci ne nous permet pas de subdiviser les classes de façon plus fine, mais confirme que 2.2 doit être une classe séparée.

4.4. Co-occurrences

Les relations de co-occurrence entre mots peuvent se diviser en plusieurs catégories, et il n'est pas possible de les présenter de façon détaillée dans le cadre de cet article. L'un des types les plus productifs est constitué par les *restrictions de sélection*, qui se trouvent à la croisée des chemins entre information syntagmatique et paradigmatic. D'une part, elles ont une base syntaxique, puisqu'elles expriment les « préférences » qu'entretiennent des mots en

relation de dépendance (verbe-objet, etc.) ; d'autre part, elles permettent le groupement des mots en paradigmes qui peuvent apparaître dans une position syntaxique donnée (par exemple *lire un(e)* <*livre, journal, revue, lettre, rapport...*>). Ces relations ne sont données qu'occasionnellement, à travers les exemples, par les dictionnaires. Elles peuvent pourtant être extraites de façon relativement aisée à partir de grands corpus, en utilisant des filtres grammaticaux et statistiques appropriés suivis d'une vérification manuelle. Dans le cas de *barrage*, cette information est extrêmement productive. Elle n'entraîne pas de subdivision supplémentaire par rapport aux classes établies jusqu'ici (ce pourrait être le cas pour d'autres vocables), mais elle permet de confirmer leur validité. Par exemple, les verbes fréquents qui prennent *barrage 2.1.1* comme objet sont *construire, édifier, démolir*, etc., tandis que ceux associés avec *barrage 2.1.2* constituent un sous-ensemble totalement disjoint : *dresser, lever, franchir, forcer*, etc. (Figure 1). Cette information est très utile pour l'étiquetage automatique : une recherche très simple des co-occurents dans notre corpus nous montre que la racine *constr-* apparaît 665 fois dans une fenêtre de 4 mots avant et après *barrage(s)*, et que dans la totalité des cas *barrage* est alors le barrage hydraulique (classe 2.1.1). De même, *dress-* apparaît 44 fois en co-occurrence avec *barrage(s)* et correspond aux barrages routiers (classe 2.1.2) dans la totalité des cas.

2.1.1	<p>▼ Adj grand, futur, gigantesque, coûteux barrage haut barrage: <i>Le haut barrage d'Assouan</i> barrage gigantesque, monumental, ultramoderne barrage hydraulique, hydroélectrique</p> <p>NP le barrage X : <i>le barrage Atatürk</i></p> <p>SPrep le barrage de X : <i>le barrage d'Assouan</i> barrage sur un fleuve, une rivière : <i>le barrage sur le Danube</i></p>	<p>▲ N le lac, les eaux, les turbines, les lâchages du barrage le chantier, la construction, l'achèvement, le financement, l'inauguration, la rupture du barrage un projet, un programme de barrage(s) un système, une série de barrages</p> <p>V(S) le barrage engloutit (des forêts, des bâtiments...), contrôle (le débit, les inondations), alimente (qqc en énergie, en électricité), fournit (de l'énergie, de l'électricité à qqc), se rompt, se brise</p> <p>V(O) construire, édifier, financer, détruire un barrage</p> <p>Prep en amont, en aval, en contrebas du barrage</p>
2.1.2	<p>▼ Adj faux barrage barrages routiers, policiers, militaires barrage filtrant, volant</p> <p>SPrep barrage de police, de l'armée, des douanes, de miliciens, de soldats, d'hommes en armes barrage sur une route, un chemin, une voie ferrée barrage de/en pierres, de/en béton, de/en terre</p>	<p>▲ V(S) une route hérissée de barrages des barrages se dressent, foisonnent sur la route</p> <p>V(O) dresser, établir, lever, démanteler un barrage rencontrer, éviter, forcer, franchir un barrage</p>

Figure 1. Restrictions de sélection pour *barrage 2.1.1* et *2.1.2*

Un autre type de co-occurrence extrêmement utile (en particulier pour les besoins du traitement automatique des langues) est constitué par les *corrélats*, c'est-à-dire les mots qui apparaissent dans le voisinage du mot-cible, sans être nécessairement reliés par un rapport de dépendance syntaxique (au contraire des restrictions de sélection). Cette liste inclut la précédente, mais contient de nombreux autres mots, qui sont dans un simple rapport thématique avec la classe d'usage considérée. La Figure 2 donne quelques-uns de corrélats pour les classes 2.1.1 et 2.1.2 de *barrage*. De telles listes sont faciles à obtenir par traitement informatique, et constituent des indices importants pour la désambiguïsation automatique.

- | | |
|-------|---|
| 2.1.1 | agricole, bassin, canal, centrale, construction, construire, cours, crue, débit, développement, digue, eau, électricité, électrique, en amont, en aval, énergie, environnement, fleuve, grand, hydraulique, hydroélectrique, inondation, irrigation, lac, lit, mètre cube, niveau, ouvrage, poisson, production, projet, réservoir, retenue, rivière, terre, vallée, etc. |
| 2.1.2 | arme, armé, armée, arrêter, bloquer, camion, chauffeur, circulation, contrôle, défense, démantèlement, démanteler, dresser, faux, forcer, franchir, frontière, gendarme, guerre, israélien, militaire, palestinien, passage, passer, police, policier, racket, région, route, routier, secteur, soldat, surveillance, terroriste, véhicule, ville, voiture, etc. |

Figure 2. Quelques corrélats pour *barrage* 2.1.1 et 2.1.2

On voit à travers cet exemple que les différents usages constituent des *classes cohérentes* à l'intérieur desquelles les différentes propriétés distributionnelles sont fortement corrélées.

5. Conclusion

Dans cet article, j'ai essayé de montrer que la tâche d'étiquetage du « sens » des mots en corpus à l'aide d'un dictionnaire était une tâche difficile, dont les annotateurs humains s'acquittaient assez mal. La raison me semble provenir principalement du dictionnaire lui-même, qui, d'une part, ne contient pas assez d'indices de surface, distributionnels, pour permettre de relier de façon sûre un « sens » et un contexte précis, et dont, d'autre part, les critères de division des entrées ne sont pas suffisamment rigoureux pour une telle tâche. Je partage complètement l'opinion de Robert Martin sur la nécessité d'un dictionnaire le plus riche possible pour la désambiguïsation ; je suis beaucoup plus sceptique sur l'adéquation des dictionnaires actuels, pour des raisons tenant à leur principes même. Bien que quelques dictionnaires aient fait un effort louable vers une plus grande systématisation, en particulier à l'aide de corpus informatisés, je soutiens la thèse que la lexicographie doit changer radicalement de paradigme, passant de la description du « sens » à celle des *usages* — un programme somme toute déjà présent dans la pensée de philosophes ou de linguistes comme Wittgenstein et Harris, mais jamais mis en oeuvre de façon extensive dans la réalisation de dictionnaires. En l'absence de ressources lexicales adéquates, l'étiquetage sémantique de corpus doit sans doute être entrepris avec prudence, car les énormes efforts qu'il requiert risquent d'être gaspillés dans la production de données peu fiables et peu réutilisables. Un tel étiquetage ne peut, à mon sens, que se faire de façon incrémentale en conjonction avec la construction d'un dictionnaire du type nouveau : le corpus sert à l'élaboration des entrées, les entrées à leur tour permettent d'affiner l'étiquetage du corpus, de façon itérative. Pour le français, ni les maisons d'édition, ni les organismes de recherche publics ne semblent pour l'instant intéressés par des projets de lexicographie audacieux, tels qu'ont pu l'être en leur temps le *Dictionnaire du français contemporain* ou le *Trésor de la langue française*. Un dictionnaire nouveau, utilisant de grands corpus et les outils informatiques qui commencent à être disponibles, serait pourtant une ressource inestimable pour les technologies du langage, et pour l'enseignement du français, particulièrement comme langue seconde. Il contribuerait très certainement à la défense du français, dont on nous dit que la place est menacée sur la nouvelle scène mondiale, organisée autour de la globalisation des marchés et de l'information.

Jean VERONIS
 Equipe DELIC
 Université de Provence
 29, Avenue Robert Schuman, 13100 Aix-en-Provence (France)
 Jean.Veronis@up.univ-mrs.fr

Références

- COHEN, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- ECO, Umberto (1976). Peirce and contemporary linguistics. *Versus*, 15, 49-65.
- HARRIS, Zellig S. (1954). Distributional Structure. *Word*, 10, 146-162.
- HORNBY, Albert S. (1942). *The Idiomatic and Syntactic English Dictionary*, Tokyo: Kaitakusha.
- HORNBY, Albert S. (1954). *A guide to patterns and usage in English*. London : Oxford University Press.
- IDE, Nancy M., VERONIS, Jean (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1), 1-40.
- KRIPPENDORFF, Klaus (1980). *Content Analysis: An Introduction to its Methodology*. Sage Publications.
- MARTIN, Robert (2001). *Sémantique et automate*. Paris : P.U.F.

- MEILLET, Antoine (1926). *Linguistique historique et linguistique générale*. Vol. 1. Paris : Champion. (2ème édition).
- MEL'ČUK, Igor, CLAS, André, POLGUERE, Alain (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain : Éditions Duculot.
- MOON, Rosamund (1987). The analysis of meaning. In Sinclair, John M. *Looking up. An account of the COBUILD project in lexical computing* (pp. 86-103). London: HarperCollins Publishers.
- STEVENSON, Mark, WILKS, Yorick (2001). The Interaction of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics*, 27(3): 321-349.
- TODOROV, Tzvetan (1966). Recherches sémantiques. *Langages*, I(1), 5-43.
- WITTGENSTEIN (1953). *Philosophische Untersuchungen*. [trad. angl. G.E.M. Anscombe: *Philosophical Investigations*, Oxford, 1953; trad. franç. P. Klossowski: *Investigations Philosophiques*, Paris, 1961].