

Mots. Les langages du politique

74 (2004)

Langue(s) et nationalisme(s)

Pierre Zweigenbaum et Benoît Habert

Accès mesurés aux sens

Avertissement

Le contenu de ce site relève de la législation française sur la propriété intellectuelle et est la propriété exclusive de l'éditeur.

Les œuvres figurant sur ce site peuvent être consultées et reproduites sur un support papier ou numérique sous réserve qu'elles soient strictement réservées à un usage soit personnel, soit scientifique ou pédagogique excluant toute exploitation commerciale. La reproduction devra obligatoirement mentionner l'éditeur, le nom de la revue, l'auteur et la référence du document.

Toute autre reproduction est interdite sauf accord préalable de l'éditeur, en dehors des cas prévus par la législation en vigueur en France.

revues.org

Revues.org est un portail de revues en sciences humaines et sociales développé par le Cléo, Centre pour l'édition électronique ouverte (CNRS, EHESS, UP, UAPV).

Référence électronique

Pierre Zweigenbaum et Benoît Habert, « Accès mesurés aux sens », *Mots. Les langages du politique* [En ligne], 74 | 2004, mis en ligne le 24 avril 2008. URL : <http://mots.revues.org/4673>

DOI : en cours d'attribution

Éditeur : ENS Éditions

<http://mots.revues.org>

<http://www.revues.org>

Document accessible en ligne sur : <http://mots.revues.org/4673>

Ce document est le fac-similé de l'édition papier.

© ENS Éditions

Pierre ZWEIGENBAUM¹, Benoît HABERT²

Accès mesurés aux sens

Les travailleurs actuels du texte numérisé sont confrontés à la multiplicité des documents électroniques (archives de presse, Web... et des logiciels). Le problème est alors de s'orienter dans le foisonnement des données textuelles et des outils. C'est aussi le lot commun de ceux qui naviguent sur le Web : le débousolement, par exemple face aux réponses d'un moteur de recherche. *Bruit*³ et *silence*⁴ brouillent notre accès aux données (hyper)textuelles. C'est pour cette raison que le consortium qui gère la Toile (<http://www.w3.org>) a lancé en 2001 le projet d'un Web sémantique : l'accès aux textes par le sens.

Naviguer dans les documents : à la recherche de « boussoles sémantiques »

Cette sémantique pour données textuelles volumineuses et hétérogènes⁵ vise la robustesse : la possibilité de traiter du texte « tout venant », « révisé » ou non, en quantité quelconque. La quantification – la mesure – y est donc centrale. Le grain visé reste grossier, bien loin des distinctions fines de la sémantique

1. Pierre Zweigenbaum, Mission de recherche en sciences et technologies de l'information médicale, DSI, Assistance publique — Hôpitaux de Paris et ERM 202, INSERM et CRIM, INaLCO STIM, 91, boulevard de l'Hôpital, 75634 Paris cedex 13 (pz@biomath.jussieu.fr).

2. Benoît Habert, LIMSI UPR CNRS 3251 et Université Paris X LIMSI — Groupe LIR (Langues, information et représentations), BP 133, 91403 Orsay cedex.

3. Nous mettons en italiques la terminologie du domaine, ou celle qu'il nous semble judicieux d'employer, et entre guillemets les notions courantes auxquelles il faut donner un sens plus précis, opératoire.

4. Le *bruit* est la proportion de documents non pertinents sur l'ensemble des documents rapportés par le moteur de recherche et le *silence* la proportion de documents pertinents non rapportés. Le complémentaire du bruit est le *rappel* et celui du silence la *précision*.

5. Le nombre de pages Web accessibles via la Toile avoisine le milliard. Ces pages vont de quelques mots à des dizaines de milliers, mélangent les langues et, inégalement relues, fourmillent de scories. Nous n'aborderons pas ici les méthodes de détermination de la langue d'un énoncé ou d'un document ; voir par exemple Grefenstette, Nioche (2000).

linguistique. Mais il varie, de grandes directions thématiques à des relations plus restreintes (hyponymie, synonymie, antonymie, relation partie-tout)⁶. Les méthodes et les outils privilégient en effet deux niveaux langagiers : celui de l'énoncé « long » et cohérent sur le plan thématique (le document, le fragment, le paragraphe, la phrase... d'une part, et le « mot » d'autre part.

Dans cette recherche de « poignées robustes » pour saisir le sens, on peut distinguer trois grands types de tâches, qui peuvent se combiner : découper, partitionner, répartir. Découper, c'est segmenter le flux textuel en « tronçons » plus ou moins longs ou en mots simples et « mots en plusieurs mots » (comme *carte bleue*, *carte de crédit*). On souhaite *constituer* les unités jugées particulièrement porteuses de sens (section 2). Partitionner, c'est rapprocher les tronçons ou les mots et faire des « tas » pour obtenir des catégories thématiques ou sémantiques (section 3). Répartir, c'est faire rentrer des tronçons ou des mots dans des classes prédéfinies (section 4). Nous traiterons à chaque fois du niveau du « mot » et de celui du tronçon. La section 5 fournira des éléments d'évaluation des performances sur ces trois axes et des indications sur les présupposés de cette sémantique « machinale »⁷.

Découper en unités textuelles ou lexicales

Le plus souvent de manière implicite, lorsqu'ils découpent une suite brute de caractères, les logiciels d'analyse textuelle incorporent et mettent en œuvre des théories ou des postulats préthéoriques sur les unités lexicales et/ou thématiques à manipuler.

Identifier les mots

Le découpage en mots le plus aisé à réaliser informatiquement consiste à (permettre d') attribuer à chaque caractère un statut de séparateur ou de non-séparateur pour l'ensemble de l'énoncé à atomiser⁸. Les inconvénients sont clairs : certains mots n'ont pas d'existence isolée (*parce*, *fur*... ; certains caractères tantôt séparent tantôt conjoignent (l'espace dans *eau froide* et dans *eaux usées*). L'optique diamétralement opposée revient à lister par avance le

6. Nous n'aborderons pas d'autres apports à l'accès automatique au sens : l'écrouissage dit résumé automatique, mise en évidence des relations de coréférence, recherche dans les documents de fragments qui apportent une réponse précise à une question factuelle, etc.

7. Pour une analyse plus fine (inspirée par les travaux de Rastier), on se reportera à Pincemin (1999).

8. Lexico3 (Lamalle et autres, 2003) permet par exemple de choisir ce statut au début du traitement d'un corpus.

maximum de « mots en plusieurs mots » et à les « souder » dans les textes⁹. Les obstacles sont connus pour le français : l'inventaire, quelle que soit sa taille, tient du tonneau des Danaïdes (les « mots en plusieurs mots » étant un des moyens privilégiés de la néologie lexicale) ; une même séquence peut constituer une dénomination dans un domaine et une suite libre dans un autre (*plante succulente* en botanique *vs* en cuisine) ; la flexibilité de certaines suites de mots rend délicats tant le tracé d'une frontière entre mots simples et mots complexes que l'identification automatique (par exemple celle des mots discontinus du type : *ne... pas, à peine... que*). Dans la pratique, les logiciels d'analyse de données textuelles s'en tiennent à une position de compromis¹⁰ : soude d'un certain nombre de mots en plusieurs mots, en particulier d'une partie des locutions grammaticales ; possibilité de paramétrer le découpage en ajoutant/supprimant des agglutinations¹¹.

Les outils d'aide à l'acquisition automatique d'unités terminologiques peuvent concourir à un tel paramétrage. Ils repèrent les suites de mots susceptibles de constituer des dénominations en s'appuyant sur des indices syntaxiques (patrons du type : Nom Nom ; Nom Adjectif), sur des indices statistiques (mesures d'« attirance » entre mots d'une séquence) (Smadja, 1993 ; Daille, 1995) ou sur des contextes particuliers (par exemple définitoires (Rebeyrolle, 2000), comme dans *On appelle X le fait de...* Ils combinent éventuellement ces voies. La question de la légitimité de leur utilisation pour des corpus relevant de la langue générale et non de domaines scientifiques ou techniques reste entière (Lespinasse, 2002). Celle du statut à donner à leurs propositions aussi : pour celles qui sont validées comme des dénominations effectives, sont-elles valables en langue ou dans le cadre plus étroit d'un domaine voire d'un corpus ?

Souder des « mots en plusieurs mots » augmente la liste des formes, lorsque les mots simples constituants d'un « mot en plusieurs mots » apparaissent également seuls : *carte de crédit* peut s'ajouter à *carte, de et crédit*. Il en va de même pour l'étiquetage (*la_{pronom}* est distingué de *la_{article}*). La lemmatisation opère en sens inverse. Les variantes possibles sur le découpage lexical remodelent sur certains points les comportements fréquentiels des « mots », sans d'ailleurs forcément conduire à de grands déplacements dans les regroupements ou partitionnements d'items.

9. Ce que font les dictionnaires de *mots composés* du LADL et l'outil de projection et de découverte qu'est Intex (Silberstein, 1993).

10. La cohérence ne règne pas forcément. Ainsi l'étiqueteur Cordial (<http://www.synapse-fr.com>) considère-t-il comme des mots en plusieurs mots à la fois *droits de l'homme, personnes âgées* et... *situation économique, mode d'organisation*, ce qui renvoie à la couverture et à l'incohérence des dictionnaires papier ou électroniques qui ont été intégrés.

11. C'est le cas d'Alceste (Reinert, 1996).

Découper en unités thématiques

Le thème du discours évolue au cours d'un texte. La segmentation thématique est l'opération qui consiste à délimiter des unités textuelles possédant une bonne homogénéité thématique, ou encore, à déterminer les points de transition d'un thème à un autre dans un texte (Manning, Schütze, 1999, p. 566-570). Les logiciels de segmentation thématique prennent généralement le paragraphe comme approximation d'unité thématique (Folch, 2002; Sébillot, 2002); la phrase est également employée. Il est aussi possible de détecter des ruptures thématiques sans préjuger de la taille des unités. Ainsi, Ferret (1998) étudie l'homogénéité thématique au travers des liens sémantiques qui existent entre les mots du passage concerné. Il examine les variations de cohésion apportées par l'ajout de chaque mot d'un texte, et place des frontières de segments aux minima de cohésion. Rien n'empêche de se recalculer a posteriori, si on le juge utile, sur les frontières de phrases ou de paragraphes les plus proches; mais les hésitations que l'on pourra rencontrer dans ce choix rappellent l'arbitraire du positionnement précis d'une frontière thématique, alors que la réalité est souvent faite de chevauchements. On remarquera que le marquage thématique par des humains ne constitue pas, lui non plus, une tâche aisée: les hésitations d'un annotateur donné sont nombreuses, l'accord entre annotateurs limité. Par ailleurs, cette notion de découpage en thèmes a été mise au point pour des textes journalistiques. Elle ne s'applique pas de la même façon à tous les types de documents (par exemple, articles médicaux, romans ou dépêches de presse).

Un intérêt de la segmentation thématique est de découper dans un corpus des segments textuels qui possèdent un certain degré d'homogénéité interne. En regroupant les segments par classe, on obtient des sous-corpus qui ont chacun une homogénéité plus grande que le corpus pris dans son entier. L'hypothèse est qu'un tel sous-corpus constitue une unité textuelle plus pertinente pour l'étude du sens des mots: classification des mots, désambiguïsation sémantique, etc. Folch (2002) montre par exemple les différences de résultats obtenus par une analyse factorielle sur le corpus initial et sur des segments thématiquement sélectionnés.

Proposer des catégories thématiques ou sémantiques

L'objectif est de regrouper les items (mots ou tronçons) «qui se ressemblent». On parle alors de *classification (clustering)*. Le résultat peut constituer une partition: tout item est dans une classe et les classes sont disjointes deux à deux. Les classes n'étant pas données au départ, il s'agit d'un apprentissage *non supervisé* (Cornuéjols, Miclet, 2002).

Obtenir des classes/rerelations sémantiques pour les mots

Deux grandes approches sont mises à contribution. La première voie s'appuie sur des « patrons » lexico-syntaxiques caractéristiques d'une relation sémantique donnée, comme ceux indices de l'hyponymie en français mis en évidence par Borillo (1996). Un schéma du type *SN tel que SN+*, c'est-à-dire un syntagme nominal suivi de *tel* (ou *telle*, *tels*, *telles*), de *que* et d'un ou de plusieurs groupes nominaux correspond souvent à une relation d'hyperonymie entre le premier SN et ceux qui suivent *tel que*, comme dans la phrase *Des cations tels que le sodium, le potassium, le calcium et le magnésium peuvent être dosés par une méthode de routine*. Déclenché par ce patron morpho-syntaxique, l'outil propose une relation d'hyperonymie entre *cations* et *sodium*, *potassium*, *calcium*, *magnésium*. Cette approche, introduite dans Hearst (1992), fait désormais place à des techniques d'apprentissage automatique de règles, qui tablent sur les distributions observées des mots d'une même classe sémantique (Morin, 1999). Deux limites sont apparues à l'usage. Certains patrons sont plus fiables que d'autres pour repérer une relation donnée. D'autre part, les patrons pertinents et leur fiabilité varient avec le domaine et le genre textuel : satisfaisant sur des énoncés didactiques, le patron donné en exemple l'est nettement moins sur du texte journalistique tout venant (Hearst, 1998).

Dans la deuxième approche, l'objectif global est de trouver des « airs de famille » entre les mots (Manning, Schütze, 1999, p. 294-308). Pour rendre le texte traitable, une première étape est celle de la réduction, de la simplification des traits associés aux mots. C'est une optique distributionnelle qui prévaut : un mot est caractérisé par les contextes dans lesquels il figure. La définition de ces contextes varie (section 5.1). Une deuxième étape est celle de l'obtention d'un indice synthétique de la proximité relative entre deux mots. Elle se concrétise par le calcul d'un indice de *similarité* entre les mots deux à deux en fonction des contextes qu'ils partagent, de ceux qui sont propres au premier, de ceux qui sont propres au second et de ceux qui ne sont employés ni par l'un ni par l'autre. De multiples indices existent (Losee, 1998, p. 51-55). La similarité de Jaccard se calcule par exemple ainsi¹² :

Formule pour l'article Zweigenbaum-Habert

$$\frac{|\text{contextes partagés}|}{|\text{contextes partagés}| + |\text{contextes propres à mot}_1| + |\text{contextes propre à mot}_2|}$$

12. Les barres verticales notent la cardinalité d'un ensemble.

Lorsque deux mots partagent tous leurs contextes, la similarité est de 1, lorsqu'ils n'en partagent aucun, elle est de 0. La troisième étape consiste à regrouper les mots en sous-ensembles en fonction des distances découlant de l'étape précédente¹³. Le regroupement peut reposer sur la classification hiérarchique (Lebart et autres, 1997, p. 155-176) : on obtient un arbre (*dendrogramme*) de mots. Cet arbre peut être élagué à un niveau donné pour obtenir un ensemble de « classes » du niveau de finesse souhaité. On peut également obtenir directement un regroupement en ensembles disjoints (techniques d'agrégation autour des centres mobiles ou « nuées dynamiques » ; ouvrage cité, p. 148-154). Dans les deux cas, le nombre de classes que l'on retient est un paramètre important. Au-delà d'une demi-douzaine, les résultats sont peu fiables. C'est en outre par commodité et avec optimisme qu'on dénomme *classes* les regroupements résultants. Il reste en effet toujours à les trier (éliminer les intrus au sein d'un regroupement et enlever les groupes « poubelles ») et ensuite à les interpréter. Les mots retenus au sein d'un groupe obéissent enfin à des relations sémantiques variables : synonymie certes, mais aussi antonymie, voire association thématique plus lâche.

Regrouper des fragments textuels en « thèmes »

Une fois obtenus des fragments textuels¹⁴, on fait émerger des groupes de fragments possédant des thèmes proches. La classification se fonde sur une mesure de distance entre les fragments textuels pris deux à deux, un fragment étant caractérisé par les mots qui le composent. Cette caractérisation est classiquement représentée par un vecteur dont chaque dimension correspond à un mot, la valeur associée à ce mot mesurant son « importance », ou force discriminante. Cette importance dépend de facteurs multiples, les plus couramment employés étant son nombre d'occurrences dans le fragment (*tf*) et le nombre de fragments dans lesquels le mot apparaît (*df*). Par exemple, la mesure *tf.idf* est un rapport de ces deux quantités (Salton, 1987). Deux fragments sont considérés comme d'autant plus proches qu'ils partagent un grand nombre de mots communs importants ; cette similarité est instrumentée par une comparaison de leurs vecteurs de mots. Outre l'indice de Jaccard¹⁵, le cosinus des deux vecteurs est couramment employé (Salton, 1987).

Les mêmes limites que celles observées pour les mots s'appliquent aux fragments textuels. Les arbres résultants ont souvent des branches malades, il

13. La distance est l'inverse de la similarité : un mot est d'autant plus proche d'un autre que la similarité entre eux est grande.

14. Leur taille peut varier de la phrase au document en passant par le paragraphe.

15. Voir Losee (1998, p. 43-62), pour une comparaison raisonnée d'indices courants.

reste un problème d'élagage des regroupements proposés. On peut également se retrouver avec un espace de fragments qui sont difficiles à séparer, qui n'ont pas d'oppositions très discriminantes. Enfin, la partition résultante fait ressortir des contrastes entre des groupes de documents possédant des traits similaires. Les traits sélectionnés dépendent cependant de la collection de documents de départ. Ils peuvent refléter en effet des différences non seulement de thème, mais aussi de domaine de discours, d'origine, de registre.

Trouver le sens ou le thème d'une unité textuelle

On dispose de catégories prédéterminées, d'un *répartitoire*, pour reprendre la terminologie de Damourette et Pichon : il s'agit de placer chaque item dans la catégorie qui lui convient. On parle alors de *catégorisation*. Les catégories étant fixées, il s'agit d'un apprentissage *supervisé*.

Donner aux mots leur sens

L'optique est différente selon qu'on s'attache à attribuer à un mot donné, sur la base de l'ensemble de ses contextes, une classe sémantique (par exemple, animé, évènement, mouvement) ou selon qu'on souhaite, en fonction d'un inventaire préalable des sens d'un mot, attribuer à chaque occurrence en contexte l'un de ces sens (désambiguïsation sémantique – *Word Sense Disambiguation* ; par exemple, *artère* – « vaisseau sanguin » vs *artère* – « avenue »).

Lorsque des classes sémantiques sont déjà fixées et que l'on connaît un certain nombre de mots de ces classes, on peut répartir de nouveaux mots dans ces classes en se fondant sur les exemples dont on dispose. On cherche les exemples d'emploi (des mots) d'une classe qui sont le plus similaires des exemples d'emploi du mot-cible, la mesure de similarité étant ici encore à choisir parmi un large éventail. L'une des méthodes souvent employées est celle des plus proches voisins. Étant donné un ensemble de mots assortis d'une étiquette sémantique, on peut chercher quels sont les k mots étiquetés les plus proches d'un mot non étiqueté donné. L'étiquette inconnue est alors déterminée en fonction de celles de ses k plus proches voisins (par exemple, par vote chez Nazarenko et autres, 2001). Le nombre de mots déjà étiquetés peut être proche de la taille d'un dictionnaire de langue (une centaine de milliers de vedettes). Mais il peut s'agir aussi d'une poignée de mots pour chacune des classes du répartitoire. On parle dans ce cas-là de mots-amorces (*seed words*) (Riloff, Shepherd, 1997).

La désambiguïsation sémantique (Ide, Véronis, 1998 ; Manning, Schütze, 1999, chap. 7) pour un mot donné suppose de disposer d'une part d'un inven-

taire de ses sens possibles (par exemple, ceux d'un dictionnaire de langue), d'autre part d'un *ensemble d'apprentissage*, c'est-à-dire de contextes dans lesquels le mot a été manuellement étiqueté (on a indiqué *le* sens qu'il a dans *ce* contexte). La taille de ces contextes peut varier, de quelques mots à gauche et à droite à la phrase voire au paragraphe. La nature des traits extraits du contexte également (Leacock, Chodorow, 1998): mots, lemmes, étiquettes morpho-syntaxiques... La tâche du *classifieur* automatique est double. Elle consiste dans un premier temps à extraire des contextes correspondant à chaque sens la configuration de traits (présence/absence, « poids » relatif, associations) qui les caractérisent. Dans un deuxième temps, face à un emploi où le mot-cible n'est pas étiqueté, le classifieur cherche la configuration de traits qui est la plus proche des traits employés dans le contexte examiné. Le sens correspondant peut alors être attribué.

Les problèmes rencontrés en désambiguïsation automatique sont de trois types (outre les incertitudes qu'un locuteur ne saurait lever). En premier lieu, le bien-fondé des étiquettes choisies. C'est en effet sur les distinctions de sens que les ouvrages de référence, les dictionnaires, divergent le plus. En second lieu, la création de contextes annotés destinés à l'entraînement du classifieur est coûteuse. Enfin, les sens d'un mot sont souvent de probabilité très inégale: il est alors difficile de rassembler des contextes permettant d'entraîner le classifieur de manière fiable pour les sens rares. Désambiguïsation sémantique et rattachement à une classe sémantique partagent certaines contraintes. D'abord, il faut disposer de suffisamment de contextes pour chaque sens ou pour chaque catégorie sémantique. En second lieu, plus le nombre de classes du répartitoire est élevé, plus difficile est la tâche de catégorisation.

Filtrer ou router un document

De nombreuses tâches pratiques reposent sur la détermination de la catégorie d'un (fragment de) document. L'objectif est généralement de répondre aux besoins d'informations personnalisées de chacun. Détecter des articles de presse qui traitent d'un thème donné permet de les *router* vers les personnes ayant affiché ce thème parmi leurs centres d'intérêt. À l'inverse, la caractérisation de thèmes indésirables est le ressort des logiciels anti-spam chargés de filtrer les messages importuns (le *pourriel*).

Une fois fixée la représentation des documents, typiquement vectorielle, comme on l'a vu les méthodes sont similaires à celles employées en catégorisation de mots (section 4.1). On dispose d'un ensemble de documents déjà catégorisés (ensemble d'apprentissage: par exemple, dépêches Reuters déjà étiquetées, ou résumés scientifiques déjà catalogués dans le corpus OHSU-MED; voir Hersh et autres, 1994). On entraîne sur cet ensemble un classifieur

automatique. On applique ce classifieur aux documents à catégoriser. Limites et contraintes sont également similaires : il est plus facile de décider entre deux catégories (SPAM vs non-SPAM) qu'entre 135 types d'articles (Reuters) ou encore que d'affecter un ou plusieurs des 20 000 mots clés du thesaurus MeSH (OHSUMED). Comme le montrent ces exemples, les catégories employées sont généralement liées aux besoins des utilisateurs. Même pour une tâche donnée, il peut y avoir extrêmement peu de recouvrement dans les catégories utilisées¹⁶.

Éléments d'évaluation

Partitionner et répartir constituent deux activités distinctes mais souvent complémentaires. L'opposition et l'interaction entre les deux activités est analogue à celles de la botanique et de la zoologie entre *mise au point de taxonomies* (organiser les êtres vivants en espèces, genres... et *détermination* ou *identification* (trouver l'espèce, le genre, etc., dont relève une plante ou un animal). Le niveau du tronçon et celui du mot sont également liés. Ainsi, dans Sébillot (2002), les 9 500 paragraphes de 200 articles du *Monde diplomatique* entre 1987 et 1997 sont partitionnés par classification hiérarchique en 80 classes thématiques, dont 27 sont retenues, car jugées pertinentes par 4 évaluateurs sur un total de 5. Les noms de chaque sous-thème sont à leur tour partitionnés sur la base des adjectifs, noms et verbes avec lesquels ils cooccurrent dans un empan de plus ou moins 5 mots. L'enchaînement de partitions au niveau du tronçon (le paragraphe dans ce cas) et au niveau du mot a pour visée d'assurer une meilleure homogénéité des contextes des mots.

Distinguer les tâches – découper, partitionner, répartir –, leurs combinaisons éventuelles, les choix effectués pour la réalisation de chacune d'elles fournit une grille d'analyse pour les logiciels d'analyse textuelle disponibles. Par exemple, Alceste repose sur deux découpages. Les mots sont tronqués (forme de *racinisation-stemming*) pour faciliter les rapprochements (sans recourir à une lemmatisation reposant sur une analyse morphologique). Certaines suites de mots sont « soudées ». Des anti-dictionnaires (modifiables) éliminent des mots jugés vides. Les tronçons sont des « phrasettes » : des suites de mots ni trop longues ni trop petites qui s'affranchissent des limites initiales des phrases pour constituer des contextes de taille similaire donc plus facilement comparables. Une double classification hiérarchique des tronçons permet de constituer de grandes classes thématiques (on peut choisir le nombre de

16. Voir Beauvisage, Assadi (2002) sur les catégories des annuaires du Web.

classes désiré). L'analyse factorielle des correspondances permet de visualiser les oppositions et rapprochements entre ces classes, les phrasettes et les mots caractéristiques de chaque classe permettent de les interpréter. Cordial assure une lemmatisation morphologique, un étiquetage morpho-syntaxique (et même sémantique), regroupe des suites de mots et découpe en «phrases». À l'inverse, face à un besoin d'analyse sémantique «outillée», cette distinction découper/partitionner/répartir aide à formuler la nature fondamentale de la tâche ou des tâches à accomplir et à examiner les choix possibles pour chacune d'elles.

La place donnée aux «airs de famille»

La méthode distributionnelle de Harris (1991) repose sur l'étude des dépendances entre opérateurs et opérands. Les analyseurs syntaxiques capables de traiter de gros volumes de textes dans leur état brut, avec leurs possibles scorries et malformations, sont des analyseurs robustes qui produisent généralement une représentation de dépendances syntaxiques de surface, qui reste une approximation des dépendances harrissiennes. Les méthodes fondées sur la simple cooccurrence de mots dans une fenêtre donnée opèrent sur une approximation plus grossière encore. Que cette fenêtre soit définie par un nombre de mots à gauche et à droite (fenêtre «graphique») ou par une unité textuelle (phrase, paragraphe, document), elle mène à une modélisation extrêmement appauvrissante. Il en est de même pour les représentations vectorielles de (tronçons de) documents. Cependant, la redondance des textes compense la pauvreté de cette approximation. La fréquence d'occurrence d'un mot est évidemment un point clé dans les différentes méthodes de collecte d'informations distributionnelles¹⁷. Curran et Moens (2002) montrent que du moment que l'on est en mesure d'augmenter significativement la taille du corpus d'étude en lui adjoignant des données similaires, les méthodes les plus simples (cooccurrences) deviennent aussi performantes que des méthodes plus complexes (analyse syntaxique)¹⁸.

Sens mouvants

Par commodité sans doute, les accès mesurés au sens privilégient des sens fixes, «immuables», en nombre limité pour un mot donné, et discrets, c'est-à-

17. Voir l'expérience comparative de Grefenstette (1996) entre les contextes de mots pleins étiquetés dans une fenêtre étroite et les dépendances syntaxiques fines fournies par l'analyseur syntaxique robuste SEXTANT (Grefenstette, 1994).

18. Voire même plus rapides malgré l'augmentation de la taille des données, car de complexité informatique moindre.

dire sans recoupement ni continuité¹⁹. La tâche de répartition peut cependant se trouver contrariée par le «vieillessement» à la fois des catégories et des traits qui leur sont associés. En routage d'informations, par exemple, certaines catégories disparaissent, d'autres apparaissent, d'autres se transforment (les traits caractéristiques ont changé). Au rebours de l'hypothèse de sens fixes, les contextes caractéristiques d'un mot peuvent changer d'une partie d'un corpus à une autre. Si ces parties correspondent à de grands thèmes, ces changements peuvent donner accès aux sens propres à chaque domaine, comme dans Sébillot (2002). Si les parties correspondent à des acteurs sociaux distincts, ces changements peuvent conduire aux notions qui occasionnent des divergences. C'est par le biais des fluctuations significatives ou au contraire des stabilités, d'une partie à l'autre, des associations < *mot, contexte* > qu'on peut progresser vers le repérage automatique des mots qui font consensus relatif et de ceux qui au contraire témoignent de divergences. C'est la démarche suivie dans Habert et autres (1999).

Rester mesuré

Le partitionnement de mots en fonction des proximités distributionnelles comme celui de tronçons débouche sur des groupements qui nécessitent sans conteste possible un travail humain en aval : émondage et interprétation. Certaines «classes» constituent de simples artefacts à éliminer. Des intrus doivent être enlevés de certains groupes, par ailleurs cohérents. Il en va de même pour la répartition de mots ou de tronçons dans des catégories prédéfinies. L'examen des performances humaines sur ces deux tâches (Véronis, 2004) permet d'éviter l'illusion d'une sémantique entièrement automatisée. L'ajout d'étiquettes sémantiques au fil du texte en fonction d'un inventaire prédéfini manifeste les divergences entre annotateurs. L'accord sur la manière de regrouper les mots en fonction de leur emploi n'est guère meilleur. Au total, les accès mesurés au sens que nous avons présentés doivent être avant tout compris comme l'occasion d'une répartition du travail fructueuse entre dégrossissage machinal et affinage humain.

19. Voir a contrario Ploux, Victorri (1998)

Bibliographie

- BEAUVISAGE T., ASSADI H., 2002, «Les annuaires du Web : entre intermédiation neutre et choix éditorial marqué», *Réseaux*, n° 116, p. 141-170.
- BORILLO A., 1996, «Exploration automatisée de textes de spécialité: repérage et identification de la relation lexicale d'hyponymie», *Linx*, n° 34-35, p. 113-124.
- CORNUÉJOLS A., MICLET L., 2002, *Apprentissage artificiel. Concepts et algorithmes*, Paris, Eyrolles.
- CURRAN J. R., MOENS M., 2002, «Improvements in automatic thesaurus extraction», *Proceedings Workshop on Unsupervised Lexical Acquisition*, Philadelphie, ACL, p. 59-67.
- DAILLE B., 1995, «Repérage et extraction de terminologie par une approche mixte statistique et linguistique», *TAL*, n° 36, 1-2, p. 101-118, *Traitements probabilistes et corpus*, Habert B. dir.
- FERRET O., 1998, «Une segmentation thématique fondée sur la cohésion lexicale», dans Zweigenbaum P. (dir.), *Actes de TALN 1998, Traitement automatique des langues naturelles*, Paris, ATALA, p. 32-41.
- FOLCH H., 2002, *Articuler les classifications sémantiques induites d'un domaine*, thèse de doctorat, Villetaneuse, Laboratoire d'informatique de Paris Nord (LIPN), Université Paris 13.
- GREFENSTETTE G., NIOCHE J., 2000, «Estimation of English and non-English language use on the WWW», *Actes de RIAO 200, Content-Based Multimedia Information Access*, Paris, CID, p. 237-246.
- GREFENSTETTE, G., 1994, *Explorations in Automatic Thesaurus Discovery*, Dordrecht & Boston, Kluwer Academic Publishers.
- GREFENSTETTE G., 1996, «Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches», dans Boguraev B., Pustejovsky J. (dir.), *Corpus Processing for Lexical Acquisition*, Cambridge, Massachusetts, MIT Press, chap. 11, p. 205-216.
- HABERT B., FOLCH H., ILLOUZ G., 1999, «Sortir des sens uniques: repérer les mots "mouvants" dans le domaine social», *Sémiotiques*, n° 17, *Dépasser les sens uniques dans l'accès automatisé aux textes*, Habert B. dir., p. 121-151.
- HARRIS Z. S., 1991, *A theory of language and information. A mathematical approach*, Oxford, Oxford University Press.
- HEARST M. A., 1992, «Automatic acquisition of hyponyms from large text corpora», dans A. Zampolli (dir.), *Proceedings of the 14th COLING*, Nantes, p. 539-545.
- HEARST M. A., 1998, «Automated discovery of WordNet relations», dans Fellbaum C. (dir.), *WordNet: an electronic lexical database*, Cambridge, Massachusetts; The MIT Press, chap. 5, p. 131-151.
- HERSH W. R., BUCKLEY C., LEONE T. J., HICKAM, D. H., 1994, «OHSUMED: An interactive retrieval evaluation and new large test collection for research», *Actes 17th ACM SIGIR*, p. 192-201.

- IDE N., VÉRONIS J., 1998, «Introduction to the special issue on word sense disambiguation: the state of the art», *Computational Linguistics*, n° 24, 1, p. 1-40.
- LAMALLE C., MARTINEZ W., FLEURY S., SALEM A., FRACCHIOLLA B., KUNCOVA A., MAISONDIEU A., 2003, *Lexico3, outils de statistique textuelle, manuel d'utilisation*, Syled – Cla2T, version 3.41., Université de la Sorbonne nouvelle-Paris 3.
- LEACOCK C., CHODOROW M., 1998, «Combining local context and WordNet similarity for word sense identification», dans Fellbaum C. (dir.), *WordNet: an electronic lexical database*, Cambridge, Massachusetts, The MIT Press, p. 265-283.
- LEBART L., MORINEAU A., PIRON M., 1997, *Statistique exploratoire multidimensionnelle*, 2^e édition, 2^e cycle, Paris, Dunod.
- LESPINASSE K., 2002, *Acquisition sémantique en langue générale: la paradocummentation textuelle pour l'indexation de documents audiovisuels sur la politique*, thèse en sciences du langage, Université de la Sorbonne nouvelle-Paris 3.
- LOSEE R. M., 1998, *Text Retrieval and Filtering: Analytic Models of Performance*, Information Retrieval, Dordrecht: Kluwer Academic Publishers.
- MANNING C. D., SCHÜTZE H., 1999, *Foundations of Statistical Natural Language Processing*, Cambridge, Massachusetts, The MIT Press.
- MORIN E., 1999, «Des patrons lexico-syntaxiques pour aider au dépouillement terminologique», *Traitement Automatique des Langues*, n° 40: 1, p. 143-166.
- NAZARENKO A., ZWEIGENBAUM P., HABERT B., BOUAUD J., 2001, «Corpus-based extension of a terminological semantic lexicon», Bourigault D., Jacquemin C., L'Homme M. C. (dir.), *Recent Advances in Computational Terminology*, Amsterdam, John Benjamins, chap. 16, p. 327-351.
- PINCEMIN B., 1999, «Sémantique interprétative et analyse automatique de textes: que deviennent les sèmes?», *Sémiotiques*, n° 17, *Dépasser les sens iniques dans l'accès automatisé aux textes*, Habert B. dir., p. 71-120
- PILOUX S., VICTORRI B., 1998, «Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes», *Traitement Automatique des Langues*, n° 39: 1, p. 161-182.
- REBEYROLLE J., 2000, *Forme et fonction de la définition en discours*, doctorat en linguistique, Université de Toulouse-le-Mirail.
- REINERT M., 1996, «Un logiciel d'analyse lexical: ALCESTE», *Les cahiers de l'analyse des données*, n° 4, p. 471-484.
- RILOFF E., SHEPHERD J., 1997, «A corpus-based approach for building semantic lexicons», *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, p. 117-124.
- SALTON G., 1987, *Introduction to Modern Information Retrieval*, Singapore, McGraw-Hill.
- SÉBILLOT P., 2002, *Apprentissage sur corpus de relations lexicales sémantiques. La linguistique et l'apprentissage au service d'applications du traitement automatique des langues*, Habilitation à diriger des recherches, Université de Rennes I, IRISA, Documents d'habilitation n° 41.
- SILBERZTEIN M., 1993, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson (Informatique linguistique).

- SMADJA F., 1993, «Retrieving collocations from text: Xtract», *Computational Linguistics*, n° 19, 1, p. 143-177, Special Issue on Using Large Corpora, I.
- VÉRONIS J., 2004, «Quels dictionnaires pour l'étiquetage sémantique?», *Le français moderne*, n° 1, *Traitement automatique des langues et linguistique*, Fuchs C., Habert B. dir.

Résumé / Abstract / Compendio

Accès mesurés aux sens

On rencontre un besoin croissant d'accès sémantique robuste à des données textuelles volumineuses et hétérogènes. Nous présentons ici en trois grands types les méthodes qui aident à obtenir cet accès, et qui s'appliquent aux mots comme aux textes : *découper* en unités porteuses de sens, *partitionner* pour obtenir des catégories thématiques ou sémantiques, et *répartir* dans des classes prédéfinies.

Mots clés : analyse sémantique automatique, sémantique quantitative.

Measured accesses to the meanings

There is a growing need for robust semantic access to large, heterogeneous textual data. We present here under three categories the methods which help to achieve such an access, and which apply both to words and to texts : segmenting into meaning-bearing units, partitioning to obtain thematic or semantic categories, and distributing into pre-defined classes.

Key words : quantitative automatic semantics, semantic analysis.

Accesos medidos a los sentidos

Se necesita cada vez más un acceso semántico a datos textuales voluminosos y heterogéneos que sea robusto. Presentamos aquí tres grandes tipos de métodos que favorecen la obtención a este acceso y que se aplican tanto a los textos como a las palabras : recortar en unidades que transportan el sentido, particionar para obtener categorías temáticas o semánticas, y distribuir por clases predefinidas.

Palabras clave : semántica automática cuantitativa, análisis semántico.