Publié dans Bulletin CILA (Commission interuniversitaire suisse de linguistique appliquée)
(« Bulletin VALS-ASLA » depuis 1994) 37, 46-62, 1983, qui doit être utilisé pour toute référence à ce travail

Documentation scientifique et technique: problèmes de multilinguisme – notion de réseau sémantique

1. *Objectifs et fonctions du service de documentation*

En schématisant, on peut classifier l'ensemble des opérations sur une chaîne documentaire dans une entreprise industrielle en trois groupes d'objectifs et de fonctions.

Le premier groupe, justiciable des techniques de gestion des stocks, porte sur les opérations d'acquisition, de prêt, de conservation des objets documentaires (livres, revues, rapports, microfiches, etc.). C'est la fonction ***bibliothèque*** qui s'intéresse à l'aspect extérieur des objets documentaires en tant que support, à leur localisation (emprunteur ou position sur un rayonnage) et, éventuellement, à l'édition d'un catalogue de ces objets.

Les deux autres groupes ont un aspect symétrique. Les objectifs essentiels de la fonction ***analyse*** peuvent se résumer ainsi: diffuser dès leur publication les analyses des documents à tous les intéressés d'une communauté scientifique et technique et, par symétrie, préparer la troisième fonction permettant par un procédé de ***recherche rétrospective*** *(retrieval)* de retrouver à tout moment les documents correspondant à un thème de recherche.

Pour atteindre ces deux derniers objectifs – diffusion et recherche rétrospective – les documents doivent être soumis à une opération d'***analyse du contenu*** dont résultera:
- la rédaction d'un résumé,
- l'indexation à l'aide de mots-clés ou «descripteurs», cette indexation créant en particulier les conditions de la recherche rétrospective.

En l'état actuel de nos connaissances sur l'analyse automatique du discours (dans une seule langue et *a fortiori* dans plusieurs langues impliquant l'analyse automatique du discours dans la langue d'origine et sa transposition dans la langue-objet, sous une forme canonique ou non, qui supposerait en outre résolu le problème de la lecture automatique), ***la rédaction d'un résumé ne peut être réalisée convenablement – et économiquement – que par l'homme.***

2. *Analyse de contenu*

De l'analyse de contenu résultent la rédaction d'un résumé et l'indexation à l'aide de mots-clés ou descripteurs, éventuellement prélevés dans un lexique contrôlé ou dans un thésaurus. D'autres méthodes d'analyse de contenu s'appuient sur des calculs de fréquences (fréquences de mots, de tournures de phrases), ainsi que sur le calcul de co-occurrences entre mots ou descripteurs dans un corpus donné.

Il est convenu d'appeler langage documentaire le moyen d'expression réduite d'un texte intégral rédigé en langage naturel, après le filtrage résultant de l'analyse de contenu d'un document. Cette expression réduite se présente généralement sous la forme suivante:
- ***indexation*** à l'aide de mots de caractérisation (prélevés ou non dans un lexique contrôlé, ou dans un thésaurus de descripteurs;
- ***résumé*** en langage naturel ou rédigé sous forme de phrases canoniques (dans le résumé, la syntaxe entre descripteurs peut éventuellement être exprimée par des indicateurs de rôle).

2.1 *Etablissement d'un résumé*

La signification d'un document écrit est détectée par une série d'intégrations successives portant sur la signification d'unités linguistiques et sémantiques de dimensions croissantes, et sur la signification des associations de ces unités à des niveaux de plus en plus complexes. Nous aurons, dans l'ordre:
- les lettres,
- les mots,
- les phrases,
- les ensembles de phrases.

La signification globale du document dépend de la signification des éléments considérés isolément à un certain niveau, et de la signification qu'ils prennent lorsqu'ils sont associés à des éléments de même niveau. La détection de la signification d'un texte passe, entre autres, par le repérage des pôles sémantiques. Mais la signification d'un texte ne peut être détectée par la seule utilisation de ses constituants linguistiques. En effet, pour «comprendre» un texte, le lecteur utilise un certain nombre de sous-ensembles hors texte faisant partie de son espace cognitif.

L'espace cognitif est constitué par l'ensemble des informations dans le texte et hors-texte dont dispose le lecteur pour appréhender la signification d'un texte. Voici quelques exemples de sous-ensembles de l'espace cognitif:

- filières de citations bibliographiques,
- substrat socio-culturel du lecteur,
- connaissance de l'auteur (biographie) et de son œuvre (bibliographie),
- images induites chez le lecteur-récepteur: à une description de l'auteur, le lecteur associe des images mentales. Cette re-création peut avoir lieu entièrement à partir des situations évoquées par le texte (c'est en particulier ce que fait le cinéaste en transposant une œuvre littéraire).

L'établissement du résumé résulte d'un ensemble d'opérations intellectuelles conduites par l'analyste (dont l'espace cognitif est constitué en grande partie par des informations hors texte):

- lecture du texte intégral,
- filtrage du texte intégral par repérage des pôles sémantiques,
- intégration successive des pôles sémantiques,
- appréhension de la signification globale,
- compte-rendu de cette signification
  - soit en langage naturel,
  - soit en langage documentaire à l'aide de phrases canoniques,
  - soit en langage naturel à syntaxe réduite.

Un bon résumé doit permettre de savoir de quoi on parle dans le document à l'aide de mots de caractérisation, et ce qui en est dit à l'aide de mots opératoires. La rédaction doit être concise et pour cela éviter en particulier la répétition du titre (qui se trouve déjà dans un autre champ de l'enregistrement), et les phrases insignifiantes (e.g. «dans cet article, l'auteur décrit . . .»). Il existe deux obstacles majeurs à la réalisation automatique de résumés:

- d'ordre pratique: nécessité de saisie du texte intégral, qui n'apporte d'ailleurs pas toute l'information;
- d'ordre théorique: passage de la détection de la signification d'un texte à la production d'un résumé significatif et lisible du contenu d'un document.

2.2 *Indexation d'un document*

L'indexation est l'une des opérations, avec la rédaction d'un résumé, qui concourt à l'analyse de contenu d'un document. Elle peut se faire soit à l'aide d'indices de classification (CDU, classification internationale des brevets, par exemple), soit à l'aide de mots du langage naturel. Dans ce cas, les termes d'indexation peuvent être attribués selon des niveaux de contrôle et de structuration différents:

- indexation en langage libre (pas d'utilisation de lexique contrôlé ou de thésaurus);
- consultation ou comparaison avec un lexique contrôlé;
- recherche de descripteurs dans un thésaurus.

Les mêmes démarches peuvent se concevoir pour l'indexation d'une question lors d'une recherche rétrospective ou pour l'indexation d'un profil documentaire correspondant aux besoins de l'utilisateur dans le cas d'une diffusion sélective de l'information.

La diffusion sélective de l'information, ou DSI, est un processus de diffusion de notices bibliographiques à une communauté d'utilisateurs qui expriment leurs besoins à l'aide de descripteurs ou mots-clés représentatifs de leur profil documentaire. Certains systèmes de recherche documentaire proposent des profils standards, d'autres acceptent l'expression de profils personnalisés. L'indexation a pour but de faciliter et d'améliorer la recherche rétrospective dans une référothèque ou base de données bibliographiques. Dans une référothèque, en réponse à une interrogation, l'utilisateur obtient les références de documents (avec éventuellement le titre, les termes d'indexation et un résumé). Les référothèques sont des sources secondaires faisant référence à des documents primaires. Les bases de données factuelles donnent un accès au renseignement cherché.

La recherche rétrospective (transposition du vocable anglo-saxon *retrieval*) désigne l'opération de recherche dans un fichier de références bibliographiques ou référothèque. Dans les anciens fichiers «manuels», cette recherche se faisait soit à l'aide d'indices de classification (CDU, par exemple), soit à l'aide de mots-matière ou mots-vedette. Dans les fichiers informatisés, cette opération est réalisée soit par la recherche de chaînes de caractères (mots) dans les parties de texte libre en langage naturel (titre et résumé), soit par la recherche de termes d'indexation (mots-clés ou descripteurs) dans les champs correspondants du fichier. Sur les sous-ensembles de documents détectés à l'aide de ces mots libres et/ou des termes d'indexation, on peut effectuer les opérations booléennes (ET, OU, SAUF) et demander ensuite l'affichage *(display)* des résultats en divers formats (auteurs, titre, références du document, résumé, etc.).

La diffusion sélective sur profil est une opération analogue: elle s'effectue en indexant la question de recherche à l'aide des descripteurs correspondant aux besoins de l'utilisateur, mais en limitant la recherche rétrospective au chargement précédent du fichier.

Considérons le résumé ci-dessous (rédigé en syntaxe réduite):

Chapitre 6 – Analyse morpho-syntaxique appliquée à la réalisation et à la consultation de banques de données textuelles

- Définition de la ***donnée textuelle*** par rapport à la ***donnée factuelle.***
- Champ d'application de l'***analyse automatique*** du ***discours:***
  - limites pratiques imposées par les problèmes de ***saisie*** ou de ***lecture automatique*** du ***texte intégral;***
  - défi majeur du *point de vue théorique,* l'exploitation des ***pôles sémantiques*** pour la *production* de ***résumés.***
- *Utilisation* de ***séparateurs*** pour la *production* d'un ***résumé à syntaxe réduite:***
  - ***segmentation*** en ***paragraphes, phrases, propositions;***
  - *mise en évidence* des ***mots*** de ***caractérisation;***
  - *facteurs d'amélioration* de la ***lisibilité*** et de la ***consultabilité.***
- *Production* d'***énoncés élémentaires*** à partir de ***résumés*** en ***syntaxe réduite:***
  - ***étiquetage sémantique*** de l'***énoncé*** à l'*aide* de ***configurations sémiques;***
  - pour la *production* d'un ***énoncé-pivot métalinguistique.***
- Conversion de ***résumés*** en ***langage naturel*** en ***résumés*** à ***syntaxe réduite.***
- ***Traduction automatique*** de ***résumés*** par l'*intermédiaire* des ***énoncés-pivot.***
- *Utilisation* des ***résumés*** en ***syntaxe réduite*** pour la *réduction* des ***paraphrases*** et la *comparaison* de ***documents*** par *confrontation* d'***énoncés.***
- *Production* de ***résumés*** en ***saisie interactive:***
  - ***indexation automatique;***
  - *mise à jour* de ***thésaurus;***
  - ***recherche*** de ***configurations sémiques*** pour l'***étiquetage lexical*** du ***thésaurus.***

---

dans lequel nous avons repéré par trois types différents de caractères:
- ***italique gras*** (mots de caractérisation),
- *italique maigre* (mots opératoires),
- romain (mots vides),

les divers mots ou syntagmes selon leur niveau de signification pour l'indexation d'un document ou d'une question lors d'une recherche rétrospective.

Dans le «texte» (titre et résumé d'un document) d'une référothèque, on trouve, en effet, plusieurs catégories de mots selon l'intérêt qu'ils présentent pour l'indexation de ce document:
- mots vides (articles et prépositions),
- mots-outil, locutions-outil (par rapport à, à partir de, etc.),
- mots athématiques (indiquant comment le sujet est exposé, pas ce qui en est dit), ce sont souvent des mots indicateurs des conditions opératoires, ils n'indiquent pas le thème traité dans le document, mais précisent ce qui est dit à propos des mots de caractérisation qui définissent ce document;
- mots de caractérisation (ce sont des mots thématiques ou des mots-clés spécifiques qui définissent le sujet ou thème traité dans le document).

Mots de caractérisation: Lors d'une recherche rétrospective, l'utilisation des mots de caractérisation est primordiale. Elle permet, en particulier, d'isoler un sous-ensemble de documents dans lequel on peut ensuite rechercher des conditions opératoires particulières.

Mots athématiques: Les mots athématiques présentent peu d'intérêt pour l'indexation et la recherche rétrospective, car ce sont des mots que l'on peut trouver régulièrement dans n'importe quel texte d'une étendue suffisante, indépendamment de son contenu. Ils servent à s'exprimer au sujet des choses plutôt qu'à exprimer les choses elles-mêmes. Ce sont des termes plus ou moins communs à tous les sujets, à toutes les situations.

Mots vides: Les mots vides sont des mots qui ne sont pas utilisés pour constituer le fichier inverse d'une référothèque. Ce sont essentiellement les articles et prépositions, les mots-outil et locutions-outil (en particulier les locutions comprenant des articles, des prépositions, des conjonctions, des pronoms: «par rapport à», «de l'», «à partir de», etc.). L'ensemble des mots vides constitue l'anti-dictionnaire.

De tels «textes» peuvent être soumis à une analyse morpho-syntaxique prenant en compte l'étude des formes de mots (bases + désinences) selon le genre, le nombre, etc. et les règles de combinaison régissant la formation des syntagmes et des phrases. La morphologie étudie la nature et la formation des mots, en dehors de leurs rapports et de leurs fonctions dans la phrase, la syntaxe s'intéresse à l'arrangement des mots dans la phrase. Si l'on dispose d'un analyseur morpho-syntaxique et des lexiques

convenables (lexique de mots vides, lexique de mots opératoires, lexique de mots ou syntagmes de caractérisation) on peut réaliser l'indexation automatique.

4. *Indexation et langage naturel*

De tels lexiques, constitués à partir des signifiants du langage naturel, présentent toutefois quelques inconvénients lorsqu'ils sont utilisés seuls lors d'une recherche rétrospective.

Ces inconvénients sont liés:

- d'une part, au «bruit» (dépistage de documents non pertinents) engendré par des phénomènes linguistiques tels la polysémie et la quasi-synonymie (flou contextuel du langage naturel);
- d'autre part, au «silence» (on ne retrouve pas tous les documents pertinents correspondant à une notion) engendré par l'absence de relation entre les notions désignées par des synonymes et des hyperonymes.

Quelques définitions et exemples vont nous permettre de cerner ce problème et d'examiner en quoi l'utilisation d'un thésaurus peut contribuer à le résoudre.

5. *Relations entre signifiants du langage naturel et objets ou notions auxquels ils font référence*

5.1 *Synonymie*

Peuvent être considérés comme rigoureusement synonymes deux signifiants en relation avec la même configuration sémique*. Mais les cas de synonymie vraie sont extrêmement rares. La plupart des «équivalents» donnés dans les dictionnaires de synonymes sont des pseudo-synonymes ou paronymes:

- cas de synonymie vraie:
  - variantes orthographiques (clé = clef),
  - sigles de substitution (dans un contexte donné),
  - équivalents d'étymologie savante ou populaire;
- cas de pseudo-synonymie résultant de:
  - mots polysémiques dont la synonymie varie selon le contexte,
  - mots dont le sens renforce (ou atténue) celui du synonyme proposé,
  - analogie du mot avec le synonyme proposé,
  - inclusion d'une notion spécifique dans une notion générique;

- relation floue entre pseudo-synonymes:
  - relation de ressemblance entre A et B d'une part, et B et C d'autre part,
  - la relation de ressemblance n'est pas transitive;
- mesure de la pseudo-synonymie:
  - il existe au moins un sème* distinctif entre configurations sémiques;
- utilisation de l'antonymie pour vérifier la pseudo-synonymie:
  - si A est antonyme vrai de B, d'une part, et B est antonyme vrai de C, d'autre part, on doit avoir A synonyme de C. Mais l'antonymie est souvent, comme la pseudo-synonymie, une relation floue. De ce fait, entre A et C on obtient au mieux une pseudo-synonymie ou paronymie.

La notion de «synonymie documentaire» répond encore plus rarement à la définition de synonymie stricte du point de vue sémantique. La relation de substitution n'est pas, du point de vue sémantique, une relation d'équivalence. Exemple: si dans un corpus donné, «abri vitré» peut être substitué à «châssis» d'une part, et à «serre» d'autre part, cela impliquerait que «châssis» est équivalent à «serre». La même remarque s'applique à l'antonymie. Dans la plupart des cas, les synonymes présentent au mieux une relation d'équivalence floue ou relation de similitude.

5.2 *Polysémie*

Au sens classique du terme, il y a polysémie lorsqu'un seul et même signifiant peut s'associer à plusieurs signifiés, c'est-à-dire lorsqu'un mot a plusieurs sens ou plusieurs emplois (en réalité, à tout mot correspond une configuration sémique comprenant plusieurs sèmes, à l'exception de ceux qui désignent les primitives). Nous distinguons ici la polysémie superficielle et la polysémie profonde. La polysémie superficielle est celle d'homographes dont les différences d'origine des divers sens sont très marqués (e.g. l'origine des trois homographes «son»), et dont l'ambiguïté peut le plus souvent être levée en surface par une analyse morpho-syntaxique. D'autres homographes présentent une polysémie profonde par extension ou restriction de sens dans le champ sémantique du mot original (e.g. le «bureau» en tant que meuble, local ou fonction; le

* Nous appelons «sème» un élément de signification, nécessaire et suffisant, pour distinguer une notion d'une autre. La configuration sémique est un ensemble de sèmes représentant les propriétés ou les caractéristiques élémentaires d'une notion ou d'un objet.

«délit» qui, dans le langage courant, est un hyperonyme de «délit» au sens pénal). La polysémie s'oppose à la monosémie (une seule acception par signifiant).

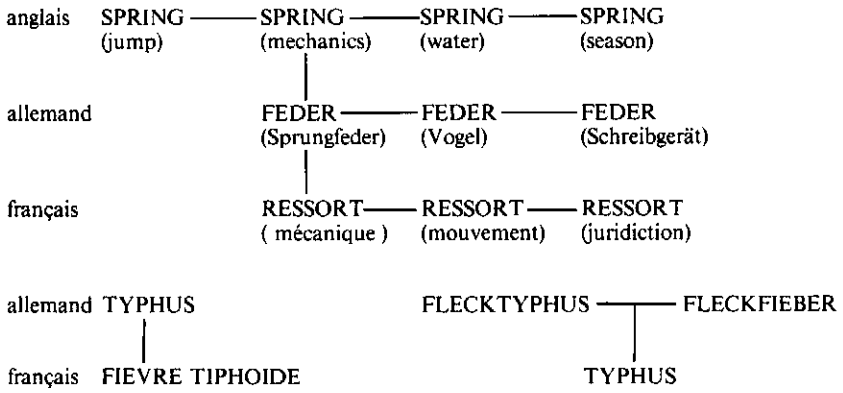
5.3 *Hyperonymie et trou lexical*

L'hyperonyme est un signifiant qui se réfère à une notion générique (e.g. le mot «parents» est un hyperonyme de «père» et «mère»). Dans certaines langues il peut ne pas exister de terme spécifique en relation avec un hyperonyme. Ainsi, par exemple, en français il n'existe que deux hyperonymes «beau-frère» et «belle-sœur» pour désigner les huit relations possibles entre deux époux, d'une part, leurs frères et sœurs, d'autre part. En russe, il existe six mots différents. Le français présente donc huit trous lexicaux, le russe deux. D'autres fois, on ne peut désigner une notion que par une périphrase, faute d'un mot précis qui s'y réfère, c'est également un trou lexical.

5.4 *Polysémie et multilinguisme*

Dans les systèmes multilingues, la transitivité est d'autant moins assurée que les mots utilisés dans chacune des langues sont polysémiques et ne font pas référence aux mêmes notions.

Problèmes de transitivité multilingue



Dans les exemples ci-dessus, on note:
- une polysémie présentée par chacun des termes dans les trois langues mais avec des variations de signification présentant des orientations différentes dans chacune des langues. La signification de chaque terme doit alors être précisée entre parenthèses. La multiplicité des polysémies dans les trois langues conduit à des relations non transitives entre les termes de chacune d'elles;
- l'homographie (chaînes de caractères identiques) qui introduit un croisement sémantique (quelquefois dangereux!) entre deux notions dans deux langues différentes («typhus» et «fièvre typhoïde»). On notera d'ailleurs que le mot «Fieber» (fièvre) présent dans l'un des deux équivalents de «typhus» donnés par le Brockhaus, introduit un nouveau risque de confusion avec «fièvre typhoïde».

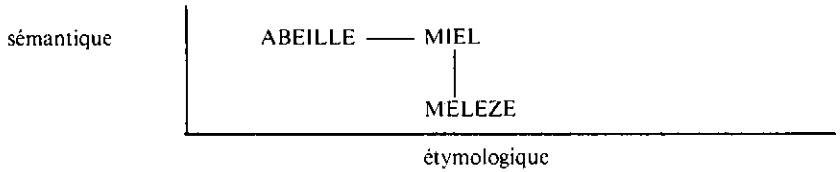
5.5 *Etymologie et dérivations lexicales*

L'étude des racines des mots et de leur dérivation lexicale par préfixation et suffixation n'apporte que peu de solutions quant aux relations entre notions auxquelles ces mots font référence.
La dérivation est parfois distinguée de la composition: dans le cas de la dérivation il y a utilisation de suffixes (petit → petitesse, charger → chargement); dans le cas de composition, le mot nouveau est «fabriqué» à partir d'autres mots (garde-barrière, tourne-disque, etc.). Le problème des préfixes est plus délicat: certains linguistes les placent dans la composition ce qui semble difficilement acceptable. En effet, comme les suffixes, les préfixes sont des éléments qui ne peuvent se rencontrer sans radical (utile → inutile), alors que dans le cas de la composition, les deux mots peuvent être séparés et s'employer isolément.

Mais l'utilisation des suffixes et des préfixes ne conduit pas à des modèles uniques de fonctionnement des variations de sens:
- la dérivation «coiffeur-coiffeuse» n'est pas transportable dans «mitrailleur-mitrailleuse»;
- les «hors-d'œuvre variés» dont il ne reste plus qu'une sorte ne sont pas nécessairement des «hors-d'œuvre avariés»;
- les mots «libraire» (en français) et «librarian» (en anglais) ont rigoureusement la même étymologie et désignent cependant des fonctions converses: le bon libraire se débarrasse au plus vite de ses livres, le bon bibliothécaire (librarian) se fait un point d'honneur de les conserver.

L'évolution lexicale, fondée très souvent sur des dérivations par analogie, conduit à utiliser des mots ayant même racine pour désigner des

objets sans relation sémantique. Tandis que, inversement, des objets ayant une grande affinité sémantique n'ont aucun rapport étymologique:



5.6 *Lexique et relations sémantiques*

Dans le domaine de la lexicographie, le terme de lexique désigne soit un recueil comprenant la liste des unités lexicales utilisées par un auteur, par une technique, par une science; soit une sorte de dictionnaire bilingue ou plurilingue comprenant une liste d'unités avec leurs équivalents dans une ou plusieurs langues. La notion de lexique peut être étendue à tout répertoire ordonné alphabétiquement ou non. Les éléments de ces répertoires présentent une relation d'ordre total. Pour l'indexation des documents, on prélève quelquefois les mots dans un lexique contrôlé qui est, par rapport au lexique général de la langue, un lexique clos concernant un vocabulaire de spécialité.

La relation d'ordre alphabétique qui préside à l'organisation d'un lexique est telle qu'elle ne permet pas de rapprochement sémantique, par exemple:

- entre les parties et un tout
  AFGHANISTAN
  AFRIQUE
  ALBANIE
  ALGERIE
  ALLEMAGNE
  AMERIQUE
  ARABIE
  ARGENTINE
  ASIE
  AUTRICHE
  BIRMANIE

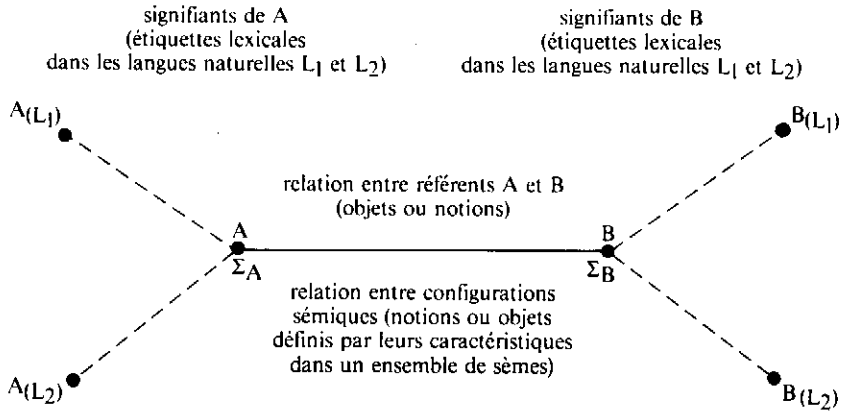
  BOLIVIE
  BRESIL
  . . .
  . . .
  EUROPE
  . . .

- ou entre des hyponymes (mots se référant à une notion spécifique: chat, chien, vache) et des hyperonymes (mots se référant à une notion générique: animal domestique, mammifère), ou réciproquement.

  De telles relations peuvent être organisées en réseau sémantique, dont une application particulière est le thésaurus utilisé pour améliorer les performances des systèmes de recherche documentaire.

6. *Réseau sémantique extra-lexical*

Nous définissons un réseau sémantique comme étant un système de relations entre objets, ou notions, auxquels les signifiants du langage naturel font référence. Dans le réseau sémantique, ces notions ou objets peuvent être représentés par des configurations sémiques (ensembles de sèmes élémentaires). Un tel réseau peut être représenté sous la forme d'un graphe (un hypergraphe en réalité) dans lequel, à chaque sommet, correspond une configuration sémique, et réciproquement:



Dans le graphe des relations entre configurations sémiques, chaque sommet correspond à une notion ou à un objet. L'étiquette lexicale est la dénomination de sommet à l'aide de signifiants (mot ou groupe de mots) dans une langue naturelle.

Chaque sommet peut recevoir autant d'étiquettes lexicales que de langues naturelles prises en compte dans le système. Si, dans une langue donnée, un sommet peut recevoir plusieurs étiquettes, celles-ci correspondent à un cas de synonymie. Si un seul mot dans une langue sert à étiqueter plusieurs sommets, il y a polysémie. Le trou lexical correspond à un sommet pour l'étiquetage duquel il n'existe pas d'unité lexicale dans une langue donnée.

L'étiquette lexicale peut être un mot simple (chaîne de caractères précédée et suivie d'un espace: chat), un mot composé (chou-fleur) ou un syntagme (chemin de fer).

7. *Système de recherche documentaire (SRD)*

7.1 *Fonction du SRD*

C'est un système informatique permettant d'effectuer des recherches rétrospectives dans une référothèque. Le SRD comprend en général, outre la référothèque elle-même, un logiciel d'interrogation muni d'un langage de commande (sélection d'un fichier, recherche d'éléments dans certains champs du fichier, opérations booléennes entre sous-ensembles de références sélectionnées, affichage des résultats, etc.).

Certains SRD permettent, à l'aide d'un thésaurus, la recherche de descripteurs correspondant à des notions plus génériques ou plus spécifiques. L'extension (ou la restriction) de la question peut être automatique dans certains SRD. Dans un SRD, la meilleure adéquation entre la réponse et la question (pertinence) a lieu lorsque l'indexation de la question avec des descripteurs ou mots-clés coïncide avec celle des documents traitant le problème.

7.2 *Rôle du thésaurus*

Un thésaurus est un instrument de contrôle de la terminologie utilisée pour l'indexation des documents (à l'entrée dans le SRD, ou lors d'une recherche rétrospective). Du point de vue de sa structure, un thésaurus est généralement constitué par:

- la liste alphabétique des termes ou descripteurs, avec indication des descripteurs en relation;
- un système de relations entre notions désignées par les descripteurs.

Dans un thésaurus, on peut exprimer les relations suivantes:
- relation générique-spécifique,
- relation tout-partie,
- relation associative,
- relation introduite par les facettes,
- appartenance à un champ sémantique.

Les deux premières relations sont exprimables par des arborescences. Dans la plupart des thésaurus, ces arborescences sont représentées par un système de retrait (indentation) typographique:

F	arbre	étiquette lexicale en français
E	tree	étiquette lexicale en anglais
D	Baum	étiquette lexicale en allemand
TG	végétal	terme générique *(broader term)*
TS	chêne	
	noyer	termes spécifiques *(narrower terms)*
	sapin	
TA	industrie papetière	

Dans quelques thésaurus, les relations entre descripteurs dans un champ sémantique sont représentées sous forme de schéma fléché.

7.3 *Relations dans un thésaurus*

Relation générique-spécifique

C'est une relation d'inclusion du type «poule ⊂ oiseau ⊂ animal». La représentation classique de cette relation est l'arborescence (ou son équivalent sous forme d'inclusions de sous-ensembles les uns dans les autres). Les règles d'établissement d'une configuration sémique spécifique sont les suivantes:
- pour apparaître sur un sommet de niveau n+1, une configuration sémique doit mettre en évidence au moins un sème distinctif de plus que la configuration située sur le sommet de niveau n;
- le sème distinctif ne peut être le même sur les collatéraux issus d'un même sommet.

Le sème supplémentaire, pour passer du niveau n au niveau n+1, doit donc être «distinctif» de ceux du niveau n, et «discriminant» entre sommets collatéraux de niveau n+1.

Relation tout-partie

Soit l'élément $\Sigma_j = \Sigma_i + s_j$ (pédale) qui est spécifique de l'ensemble Σ_i (organe de transmission). Considérons maintenant l'élément «pédale» en tant que partie d'un certain nombre de machines (embrayage d'automobile, bicyclette, machine à coudre):

«pédale» n'est pas un spécifique de «bicyclette»
«bicyclette» n'est pas un générique de «pédale»

Dans cette relation, le même type d'éléments appartient généralement à une très grande diversité d'ensembles non définis à l'avance. Chaque élément est susceptible d'entrer dans une relation polyhiérarchique (e.g. les «habitants» dans divers découpages administratifs, les «organes» d'un grand nombre de machines, etc.) sans autre caractéristique que l'appartenance à cet ensemble. Mais il faut distinguer entre relation tout-partie, d'une part:

«pédale» $\in \{M_1\}$ (ensemble des pièces d'une automobile)
«carburateur» $\in \{M_1\}$ (ensemble des pièces d'une automobile)
«volant» $\in \{M_1\}$ (ensemble des pièces d'une automobile)

et relation générique-spécifique, d'autre part:

{chien, vache, cheval} $\subset$ {mammifère} $\subset$ {animal}

Relation de substitution

La relation de substitution est utilisée dans un certain nombre de thésaurus pour indiquer l'emploi préférentiel (TP = terme préféré) d'un descripteur, par rapport à un autre terme (NP = non préféré) qui ne doit pas être utilisé pour l'indexation. Il s'agit de «synonymes documentaires» qui ne sont pas nécessairement des synonymes au sens où on l'entend généralement, surtout lorsque certains des termes préférés sont polysémiques. Par exemple, la substitution «fréquence TP période» indiquant une relation réciproque entre deux notions, est admissible dans le cas d'un phénomène vibratoire; elle ne l'est plus lorsqu'on parle de période géologique. La relation de substitution n'est donc sémantiquement admissible que dans un corpus défini.

Relation associative

Relation entre deux configurations sémiques telles que:
- les deux termes de la relation ne soient pas déjà connus dans une autre relation (générique-spécifique, tout-partie, substitution);
- les deux configurations des sèmes communs permettent de calculer une distance sémantique entre elles.

Lorsque le nombre de notions associées est supérieur à deux, la relation entre elles peut être transitive (relation de similitude) ou intransitive (relation de ressemblance).

Relation de similitude

Relation d'équivalence floue (transitive, réflexive et symétrique). La notion de similitude constitue le pont entre «équivalence» et «ressemblance», la relation ordinaire la plus proche d'une relation de similitude est une relation d'équivalence. La relation associative entre configurations sémiques peut être, en fonction de la transitivité, une relation de similitude; ou de la non-transitivité, une relation de ressemblance.

Relation de ressemblance

Contrairement à la relation de similitude, la relation de ressemblance n'est pas transitive:

A ressemble à B
B ressemble à C
A ressemble à C

8. Indexation automatique

Soit le résumé de la fiche 128-0521-0320:

L'arc électrique qui, autrefois, était un phénomène indésirable, est maintenant
*** ARC ELECTRIQUE ***
considéré comme un élément dynamique et vital du disjoncteur. Les études effectu
DYNAMIQUE vital *DISJONCTEUR*
ées en physique du plasma permettent d'arriver à une commande dynamique de l'arc
***PHYSIQUE DES PLASMAS* COMMANDE DYNAMIQUE ARC**
dans les disjoncteurs. Description à l'aide de schémas des phénomènes apparaiss
DISJONCTEUR* *SCHEMA
ant lors de la séparation des contacts d'un disjoncteur en prenant l'exemple d'u
SEPARATION CONTACT *DISJONCTEUR*
n disjoncteur à SF6 à autosoufflage. Stabilisation de l'arc. Calcul du diamètre
***DISJONCTEUR A SF6* AUTOSOUFFLAGE STABILISATION ARC CALCUL diamètre**
de la buse de soufflage. Tension transitoire de rétablissement. Evolution de l'a
BUSE DE SOUFFLAGE* *TENSION TRANSITOIRE DE RETABLISSEMENT
rc au cours de la coupure. Refroidissement de la colonne d'arc. Importance de la
COUPURE D'ARC* *REFROIDISSEMENT* *COLONNE D'ARC
tension de claquage du gaz.
TENSION DE CLAQUAGE* *CLAQUAGE DES GAZ

présenté ici avec les conventions suivantes:
- «L'arc électrique qui, autrefois, . . .» lignes impaires = texte frappé par l'opératrice.
- Lignes paires = informations affichées par le système après élimination des mots vides:

ARC ELECTRIQUE
– terme d'indexation (TG ou TS) existant déjà dans le thésaurus.
vital
– chaîne de caractères non reconnue ni dans l'antidictionnaire, ni après l'exploitation du dictionnaire des bases morphologiques.
COMMANDE DYNAMIQUE ARC
– trois descripteurs unitermes reconnus dans le fichier inverse.

L'indexation automatique consiste à indentifier dans le texte (titre et résumé) tout mot qui, après une analyse morphologique, peut être considéré comme descripteur entrant dans la constitution du fichier inverse d'interrogation, soit comme descripteur uniterme, soit comme terme d'indexation (TG ou TS) (en tant qu'uniterme ou élément d'un syntagme).

Pour cela, il faut pouvoir comparer les chaînes de caractères du texte avec:
- l'antidictionnaire ou lexique des mots vides;
- le lexique des bases morphologiques plus les désinences possibles (si la comparaison ne situe la chaîne de caractères dans aucun des deux lexiques, proposer son introduction dans le lexique des bases morphologiques et, par conséquent, dans la lexique des descripteurs) et, à la suite de cette comparaison:
  - afficher la forme lexicale trouvée dans le lexique «descripteurs»,
  - vérifier si les formes ainsi repérées ont une signification compatible avec le texte,
  - dans l'affirmative, proposer leur validation à l'analyste indexeur,
  - éventuellement, utiliser les unitermes ou syntagmes ainsi repérés comme termes d'indexations TS (ou TG),
  - mettre à jour le fichier inverse.

Le projet d'indexation automatique présenté ici doit être nettement distingué de la construction automatique de thésaurus. Cette construction automatique consiste à:
- proposer à l'indexeur des liens TS-TG en fonction de l'expérience du système,
- mettre à jour le thésaurus avec les liens ainsi déterminés.

La réalisation de l'étape «indexation automatique» (qui peut éventuellement être considérée comme une étape finale) est un préalable à la réalisation d'une construction automatique de thésaurus.

Merlin Gerin
Service de documentation
F-38050 Grenoble Cedex

André Deweze