Détection de liens de synonymie : complémentarité des ressources générales et spécialisées

Dans le cadre d'une aide à la structuration de terminologies, l'utilisation de données sémantiques générales nous a conduits à proposer des règles d'inférence de liens de synonymie entre des candidats termes complexes. Au regard de l'évaluation faite par un expert du domaine en contexte applicatif, il s'avère que nous obtenons des résultats intéressants. Ces premiers résultats montrent que, contrairement à une opinion assez largement admise, l'utilisation de ressources générales comme un dictionnaire de langue pour le traitement de documents techniques se justifie. Nous cherchons ici à caractériser plus précisément l'apport de ces ressources. Nous avons confronté les résultats obtenus aux liens inférés à partir de ressources lexicales plus spécialisées. Il s'avère que peu de liens inférés sont communs d'une ressource à l'autre. Ceci souligne la complémentarité des différentes sources et l'intérêt spécifique des informations de la langue générale pour la structuration de terminologies.

Termes-clés : structuration de terminologie ; variation sémantique ; synonymie ; ressources lexicales ; langue spécialisée *vs.* langue générale.

1 Introduction

Le travail présenté ici s'inscrit dans un projet de développement d'outils d'aide à la structuration et à la mise à jour de terminologies. Il résulte d'une collaboration entre le *LIPN* et la Direction des études et recherche d'électricité de France (DER-EDF). La terminologie constituée est ensuite utilisée dans un système de consultation de documents techniques (*SCDT*). Notre objectif est de fournir des liens sémantiques, de synonymie notamment, entre termes extraits d'un corpus technique. Ces liens doivent faciliter la navigation dans les documents techniques. Nous présentons et confrontons les résultats de l'utilisation de plusieurs types de ressources sémantiques pour inférer des liens de synonymie : un dictionnaire de langue, des classes de synonymes construites manuellement et le thesaurus EDF.

Ce travail s'inscrit dans le débat concernant le statut des langues de spécialité et l'apport des ressources lexicales comme un dictionnaire de langue pour le traitement des documents techniques. On constate un fort contraste dans la description des unités lexicales figurant dans les dictionnaires d'usage et leurs emplois dans des langues de spécialité : les mots des documents techniques ne sont pas toujours répertoriés dans les dictionnaires et quand ils le sont, c'est souvent avec des sens plus variés et différents. L'hypothèse couramment admise est donc que ce type de ressources n'est pas exploitable pour le traitement de documents spécialisés pour lesquels on s'attache à construire des ressources spécifiques.

Les premiers résultats que nous avons obtenus soulignent cependant l'apport d'un dictionnaire de langue pour l'aide à la structuration de terminologie. Dans cette expérience préliminaire (Hamon *et al.* 1998), le dictionnaire *Le Robert* a été utilisé pour inférer des liens de synonymie entre termes. La validation des résultats par un expert du domaine concerné montre que 37% des liens inférés expriment effectivement une relation de synonymie (*action de protection* / *action de sauvegarde*) et plus largement que la moitié des liens est utile pour la structuration de terminologie (*rapport de sûreté* / *analyse de sûreté*, que l'expert analyse comme un lien de méronymie).

Sur la base de ces premiers résultats, nous cherchons ici à caractériser plus précisément l'apport de ce type de ressources lexicales en les confrontant à des données plus spécialisées dans la perspective du développement d'outils d'aide à la navigation dans les documents techniques. Dans ce qui suit, nous opposons donc des ressources « générales » comme un dictionnaire d'usage à des ressources spécialisées. En pratique ces ressources diffèrent surtout par l'usage plus ou moins spécialisé qui en est fait.

1.1 Utilisation de liens sémantiques dans un SCDT

De nombreuses applications dans les domaines de spécialité nécessitent l'utilisation de terminologies: indexation contrôlée, aide à la rédaction, consultation de documents, etc. Nous nous intéressons ici à ce dernier type d'applications. La taille croissante de la documentation technique conduit en effet les entreprises à développer des outils de navigation dans leurs documents.

Le système de consultation de documentation technique développé par EDF (Gros *et al.* 1996), (Gros *et al.* 1997) fournit un accès hypertexte au document suivant différents modes:

• Une table des matières;
• Un accès plein texte par mots-clés, étendu par la consultation d'une terminologie;
• Un index du domaine intégrant des liens de synonymie et d'hyponymie entre une entrée et une sous-entrée;
• Un index de l'activité modélisant la tâche de l'utilisateur.

Le processus d'aide à la construction et à la structuration de terminologies doit faciliter l'intégration de nouveaux documents dans un système de consultation de documents techniques en fournissant des liens sémantiques entre les termes extraits du document. Les liens de synonymie sont introduits dans le système pour enrichir l'index et la terminologie proposés aux utilisateurs.

1.2 Structuration d'une terminologie

Le processus de constitution d'une terminologie se divise en deux grandes phases (Dagan *et al.* 1994). Dans un premier temps, les termes candidats sont extraits d'un corpus pertinent pour le domaine étudié. L'ajout de liens sémantiques permet ensuite d'obtenir un réseau terminologique candidat, plus complexe.

Le système d'EDF repose sur cette démarche. L'extraction des candidats termes est assurée par le logiciel d'extraction de terminologie *Lexter* (Bourigault 1994). Les candidats termes sont organisés en un réseau syntaxique. Les liens sémantiques sont ajoutés dans le réseau syntaxique sous la forme de classes conceptuelles (Assadi 1997) ou de liens de causalité (Garcia 1998). Notre travail porte sur cette deuxième étape: il vise à enrichir le réseau initial de liens de synonymie entre candidats termes.

Notre méthode repose sur l'utilisation de ressources lexicales telles qu'un dictionnaire de langue. Les bases de connaissances lexicales en langue de spécialité étant rarement disponibles, nous avons évalué la pertinence et l'utilité des informations sémantiques générales dans les documents techniques (Hamon *et al.* 1998). À partir du lien de synonymie (*commande / ordre*) un lien de synonymie est inféré entre les candidats termes *commande manuelle* et *ordre manuel*.

Les résultats obtenus montrent la pertinence des informations sémantiques générales et apportent une réponse expérimentale au débat sur le rôle de ces connaissances dans le traitement des textes de spécialité. Cependant, bien que les résultats soient intéressants du point de vue de l'expert, il est nécessaire de les combiner à des données spécialisées. Nous avons donc cherché à caractériser la contribution respective des différents types de données lexicales. Nous avons confronté le dictionnaire de langue avec deux ressources spécialisées disponibles au sein de la DER-EDF et *a priori* pertinentes pour l'un de nos corpus d'expérimentation.

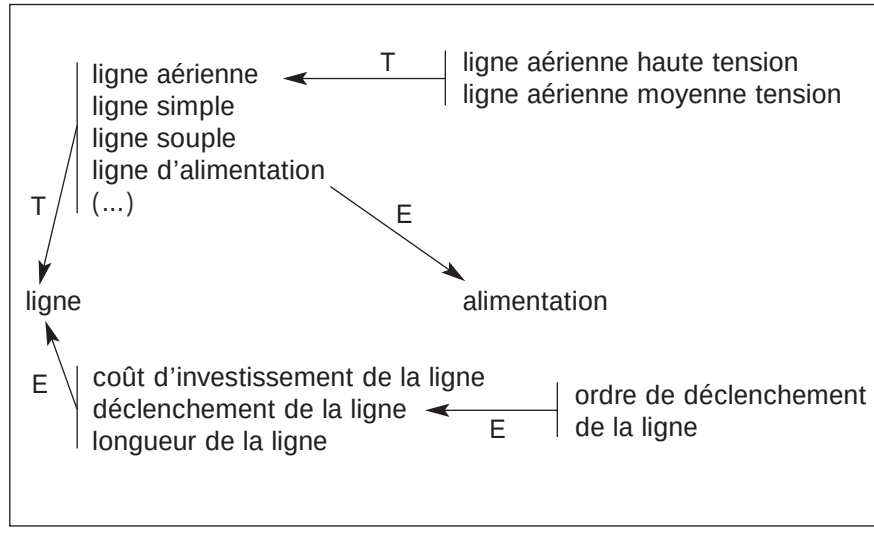

Figure 1:
Fragment du réseau syntaxique Lexter (T = tête, E = expansion).

1.3 Présentation du corpus de travail

Le corpus des dossiers de système élémentaire (*DSE*) comporte environ 160 000 mots et décrit en partie le fonctionnement des centrales nucléaires.

Le corpus est analysé par *Lexter* (Bourigault 1994) qui en extrait 17 675 candidats termes (2 865 noms, 1 306 adjectifs et 13 504 groupes nominaux), structurés en réseau syntaxique (*cf.* figure 1). Chaque candidat terme complexe (par ex. *ligne d'alimentation*) est décomposé en une tête (*ligne*) et une expansion (*alimentation*).

2 Inférence de liens de synonymie

2.1 Principe général

Notre définition de la synonymie est proche de celle proposée dans *WordNet* (Miller *et al.* 1993). Alors que la synonymie peut être vue comme une relation graduée, nous la considérons comme une relation d'équivalence contextuelle. Ainsi, à l'instar de la synonymie cognitive de (Cruse 1986), nous définissons une relation de synonymie cognitive contextuelle entre deux termes X et Y dans un contexte C si les deux termes sont syntaxiquement identiques et substituables – *salve veritate* – dans le contexte C.

L'inférence d'un lien de synonymie entre termes complexes repose sur l'hypothèse que la compositionnalité des termes complexes préserve la synonymie. Cette hypothèse est évidemment simplificatrice : ce n'est pas parce que *Le Robert* donne *arrêt* et *interruption* comme synonymes dans certaines acceptions que le terme complexe *arrêt du réacteur* doit s'entendre au sens de *interruption du réacteur*. En pratique, l'inférence d'un lien de synonymie suppose que deux termes complexes comportant des éléments synonymes et construits selon le même schéma syntaxique soient attestés en corpus. Nous considérons que deux termes sont synonymes si leurs composants sont identiques ou synonymes. Dans l'exemple ci-dessus, *interruption du réacteur* n'étant pas attesté dans le corpus, aucun lien de synonymie n'est inféré. En revanche dès lors que *arrêt de l'appoint* (ellipse pour *arrêt de l'appoint en acide borique*) et *interruption de l'appoint* sont tous deux des termes attestés, nous faisons l'hypothèse qu'ils sont synonymes.

Cette démarche se rapproche en fait de celle de Basili *et al.* (1997) dans le sens où elle exploite des données de la langue générale pour des corpus spécialisés lorsque ces données sont corroborées en corpus par l'existence de constructions parallèles (dans notre cas) ou par la similarité des contextes d'apparition (pour R. Basili et ses collègues).

2.2 Détection des candidats termes synonymes

La méthode générale d'inférence des liens de synonymie est présentée dans Hamon *et al.* (1998). Elle se décompose en deux étapes.

La première est une étape de filtrage qui réduit la taille des données utilisées lors de l'application des règles d'inférence. Un lien entre deux termes est conservé si ceux-ci sont tous les deux présents dans le document étudié. Par exemple, le lien (*portion / tronçon*) est retenu si ces lemmes figurent sous une forme quelconque dans le corpus.

La deuxième étape est le processus inférentiel à proprement parler. Nous avons conçu trois règles pour inférer des liens de synonymie entre candidats termes complexes. Un lien de synonymie est ajouté entre deux candidats termes du réseau syntaxique si l'une des trois conditions suivantes est vérifiée :

- Règle 1 : les têtes sont identiques et les expansions sont synonymes (*action de protection / action de sauvegarde*) ;
- Règle 2 : les têtes sont synonymes et les expansions sont identiques (*capacité faible / puissance faible*) ;
- Règle 3 : les têtes sont synonymes et les expansions sont synonymes (*classement d'équipement / classification de matériel*).

Nous contraignons les composants des termes à posséder la même catégorie syntaxique. Par ailleurs, nous avons choisi de ne pas tenir compte des prépositions et des formes fléchies des termes. Ce parti pris réduit le coût du calcul des liens sémantiques et a peu d'incidence sur les résultats.

Les liens initiaux sont d'abord utilisés pour amorcer la détection des candidats termes complexes. Puis, tant que de nouveaux liens sont trouvés, nous réitérons le processus en prenant en compte les liens précédemment détectés.

La méthode a été testée sur des corpus de taille différente : *Menelas* (85 000 mots), *DSE* (160 000 mots), *Crater* (750 000 mots). Un algorithme efficace permet l'application à des corpus techniques importants. Une interface de validation est en cours de réalisation.

2.3 Protocole de validation

Un expert du domaine a validé les résultats. Les liens inférés ont été acceptés ou rejetés en tenant compte du contexte d'application de ces résultats : un système de consultation de documents techniques. De plus, le statut terminologique des candidats termes liés a été pris en compte contrairement à la première expérience (Hamon *et al.* 1998). Ainsi, lors de la validation des résultats, les liens dont l'un des

candidats termes liés ne pouvait avoir le statut de terme, ont été rejetés.

Nous avons choisi de présenter les candidats termes liés sous la forme fléchie de leur première occurrence rencontrée dans le corpus. Lors de la validation, l'expert a pu accéder aux groupes nominaux maximaux ainsi qu'aux phrases dans lesquels se trouvent les candidats termes liés.

Toutefois, bien que notre objectif soit la détection de liens de synonymie entre des candidats termes complexes, nous avons constaté que les liens inférés peuvent être typés différemment (Hamon *et al.* 1998). Nous laissons donc à l'expert la possibilité de modifier le type du lien. En effet, bien qu'il ne s'agisse pas de liens de synonymie, il est intéressant, dans le cadre d'une aide à la structuration de terminologie, de conserver tous les liens sémantiquement pertinents. L'évaluation des résultats tient compte de cette caractéristique.

Afin d'assurer une certaine cohérence, les liens inférés ont été structurés et présentés suivant deux modes de regroupement :

• Structuration par famille : les liens sont regroupés en fonction du lien initial utilisé par les règles d'inférence. Par exemple, les liens (*débit maximum / volume limite*, *débit nécessaire / volume requis*, *débit d'acide borique / volume d'acide borique*, *débit d'eau / volume d'eau*, *débit total / volume total*) sont présentés ensemble puisqu'ils sont inférés à partir du même lien initial (*débit / volume*) ;

• Structuration par classe : nous avons regroupé les liens qui constituent les chemins de longueur *n* entre deux termes pour les proposer ensemble à la validation. Ainsi, par exemple, les liens *gamme logarithmique / échelle logarithmique*, *échelle logarithmique / mesure logarithmique* sont regroupés dans la même classe.

L'expert a essentiellement utilisé cette deuxième présentation des résultats pour valider les liens, les candidats termes liés apparaissant dans le graphe avec leur voisinage. Ce type de présentation propose une vue globale sur un ensemble de liens et favorise la cohérence de la validation.

3 Utilisation de différentes ressources lexicales

Cette partie présente les résultats de l'application de l'inférence de liens de synonymie sur le corpus de travail à partir de trois sources lexicales différentes. La confrontation de ces différents ensembles de résultats permet de mieux évaluer l'apport respectif de chaque source lexicale.

3.1 Exploitation d'un dictionnaire de langue

Nous avons utilisé pour cette étude les informations sémantiques du dictionnaire *Le Robert* fournis par l'Inalf. Même s'il comporte des indications de synonymie, ce n'est pas à proprement parler un dictionnaire de synonymes. Cependant, ce dictionnaire, largement disponible, est reconnu comme un standard. Il permet d'effectuer des expériences dans des conditions réelles. Des listes de liens de synonymie ont été extraites. Une entrée peut comporter différentes listes de synonymes correspondant à chacun de ses sens mais les synonymes eux-mêmes ne sont pas désambiguïsés. Les indications de sens n'étant ni explicites ni homogènes, nous les avons négligées (Ploux *et al.* 1998). De plus, les liens de synonymie extraits du dictionnaire pouvant être très contextuels ou exprimer, par exemple, des relations d'analogie, l'application de la propriété de transitivité provoque de nombreuses erreurs : nous n'exploitons pas cette propriété de la synonymie.

Le dictionnaire couvre 40% des candidats termes extraits du corpus des DSE. À partir des données du dictionnaire, nous avons cherché à inférer des liens de synonymie sur ce corpus. L'étape de filtrage conserve 3 129 liens exprimant principalement des relations de synonymie entre mots simples. Ces liens permettent ensuite d'inférer 590 liens sur les candidats termes complexes. L'expert a jugé que 199 liens inférés (33,7 %) sont pertinents : *air de l'enceinte / atmosphère de l'enceinte*, *changement de gamme / modification d'échelle*, *fiche de correction / fiche de modification*. Parmi ces liens, 101 liens ont été retenus comme exprimant des relations de synonymie (*débit nul de refroidissement / débit nul de réfrigération*, *fluide actif / liquide radioactif*, *prescription de sûreté / règle de sûreté*) et 84 comme des liens de type Voir-Aussi (*liaison d'alimentation / ligne d'alimentation*). Les autres liens sont typés comme des relations d'hyperonymie ou de méronymie (*phénomènes naturels / phénomènes physiques*). Une partie des erreurs est due au fait que des candidats termes liés ne peuvent avoir le statut de termes. Nous estimons que le taux de précision est proche de celui de la première expérience (37 % de liens de synonymie valides, 50 % de liens sémantiquement valides) si le statut terminologique de candidats termes n'est pas pris en compte.

3.2 Amorçage à l'aide de classes de synonymes construites manuellement

Des classes de synonymes ont été constituées manuellement par un expert pour un corpus du même domaine que celui de notre corpus de travail. Nous avons à notre disposition un ensemble de 500 classes constituées de 1 335 termes. Considérées comme des classes d'équivalence, ces classes fournissent

3 456 liens de synonymie. Ces liens reposent sur des relations morpho-syntaxiques (*appoint en acide borique / appoint en bore*) ou des relations sémantiques (*eau de refroidissement / fluide réfrigérant*). Il s'agit de liens entre des termes complexes mais aussi entre termes simples et termes complexes (*appareil de mesure / capteur*). De telles ressources spécialisées sont précieuses mais rares. Elles sont coûteuses à construire et demandent à être mises à jour régulièrement.

Lors de l'étape de filtrage 281 liens sont retenus. Les règles permettent d'inférer 167 liens sémantiques entre des termes complexes. Lors de la validation, 143 liens sont jugés pertinents (85,6 %): *concentration en bore du circuit primaire / teneur en bore du circuit primaire*, *Procédure contrôle / procédure d'essais*. Parmi ces liens, 106 liens sont des relations de synonymie (*circuit d'alimentation / réseau d'alimentation*, *bilan de fuite global du circuit primaire / bilan de fuite global du primaire*) et 23 liens de type Voir-Aussi (*état normal d'exploitation / conditions normales d'exploitation*)

Nous avons également appliqué les règles d'inférence en utilisant les liens extraits des classes de synonymes à la suite de liens du dictionnaire. On constate que 41 liens peuvent être inférés par l'une ou l'autre des deux ressources indifféremment. L'utilisation conjointe de ces deux ressources lexicales permet d'inférer 40 nouveaux liens supplémentaires. Il s'agit de liens obtenus par la règle 3. Si un lien (ex. *tronçon du circuit de réfrigération intermédiaire / portion du circuit RRI*) ne peut être inféré qu'à partir de deux liens initiaux issus de deux sources différentes (*tronçon / portion*, fourni par le dictionnaire et *circuit de réfrigération intermédiaire / circuit RRI*, fourni par les classes de synonymes), seule l'utilisation conjointe de ces deux ressources permet de l'inférer. L'expert a validé 9 des 40 liens supplémentaires, 3 liens exprimant des relations de synonymie (*liaison d'alimentation / ligne de distribution*) et 2 des relations Voir-Aussi (*vidange du réservoir / purge des bâches*). Le faible nombre de liens pertinents est dû à l'application de la règle 3. En effet, comme nous l'avons déjà constaté (Hamon *et al.* 1998), cette règle ne permet pas de proposer beaucoup de liens pertinents. Cette règle est cependant précieuse puisqu'elle permet d'inférer des liens difficiles à trouver manuellement par les terminologues.

3.3 Inférence de liens à partir d'un thesaurus

Le thesaurus EDF (EDFDOC) contient 20 000 termes simples ou complexes organisés entre 330 champs sémantiques (Circuit électrique, Technologie des câbles, Organisation administrative), eux-mêmes regroupés en 45 points de vue (classes très générales). Trois types de liens sont proposés: les liens associatifs (Voir-Aussi: *source autonome / alimentation de secours*), les liens hiérarchiques (Hyponymie: *sécurité des personnes / protection de l'opérateur*) et les liens de synonymie (Est employé pour: *panier filtrant / tamis*). Nous avons utilisé tous les liens sémantiques du thesaurus, soit 25 000 liens.

Bien que la plupart des liens du thesaurus expriment des relations sémantiques ou morpho-syntaxiques entre termes complexes, nous avons cherché à saturer le réseau terminologique candidat de la même manière que pour les classes de synonymes. Ainsi, 389 liens sont conservés lors du filtrage et 55 liens sont inférés. Ce faible résultat tient à la complexité des termes présents dans le thesaurus. Les liens inférés portant sur des termes plus complexes, peu de liens sont détectés. Lors de la validation, 36 liens sur 55 (65,4 %) sont retenus par l'expert (*capteurs de pression / mesures de pression*, *signalisations lumineuses locales / voyants locaux*, *arrêt de l'appoint / interruption de l'appoint*). Les liens de synonymie ne représentent qu'une faible partie des liens validés (4 / 36) alors que 15 liens sont typés comme des liens Voir-Aussi (*indicateurs locaux / mesure locale*, *capteurs de niveau / régulation de niveau*).

4 Caractérisation de l'apport spécifique de chaque source lexicale

La table 1 présente les résultats obtenus à partir des trois sources. La proportion de liens inférés par rapport au nombre de liens retenus lors de l'étape de filtrage varie suivant le type de source utilisé. Cette proportion est de 3/5 pour les classes de synonymes et seulement de 1/5 pour le dictionnaire dont la productivité est donc faible. Néanmoins, du fait de la taille de ce dernier, les liens inférés à partir du dictionnaire représentent 78% du total des liens inférés tandis que seulement 22% des liens sont inférés à partir des classes de synonymes. Les informations sémantiques extraites du dictionnaire permettent d'inférer beaucoup de liens à un faible coût alors que la constitution des classes de synonymes est très coûteuse.

Les résultats présentés ci-dessus font apparaître des différences numériques significatives dans l'apport des différentes sources lexicales utilisées. Les sections qui suivent comparent les résultats obtenus par inférence avec les données présentes dans les classes de synonymes et le thesaurus qui contiennent eux-mêmes des termes complexes. Nous cherchons ainsi à mieux caractériser l'apport spécifique

	Filtrage des liens		Inférence des liens sur le corpus	Validation
	Nombre de liens conservés	Nombre de termes	Nombre de liens inférés	Nombre de liens
Dictionnaire	3 129	1 299	590	199 (33,7 %)
Classes de synonymes	281	344	167	143 (85,6 %)
Thesaurus EDF	389	478	55	36 (65,4 %)
Dictionnaire puis classes	3 376	1 547	756	315 (41,6 %)

Tableau 1 :
Résultats de la méthode d'inférence

de chaque source lexicale et leur complémentarité.

4.1 Complémentarité des classes de synonymes construites manuellement et du dictionnaire

La confrontation des liens inférés à partir du dictionnaire de langue avec les liens fournis par les classes de synonymes ou inférés à partir de ceux-ci est riche en enseignements. Le tableau 2 synthétise ces résultats.

On remarque tout d'abord que très peu des liens établis manuellement par l'expert (19 liens) peuvent être trouvés par inférence. Dans la mesure où les classes de synonymes contiennent beaucoup de liens entre termes complexes, on pourrait s'attendre à ce que les règles d'inférence, appliquées sur des classes de synonymes, retrouvent des liens déjà présents dans ces mêmes classes. Ainsi le lien de synonymie (*condition de fonctionnement / régime de fonctionnement*) qui est donné dans les classes est également inféré à partir du lien (*condition / régime*) lui aussi établi par l'expert. Le nombre de ces liens est cependant faible (19 liens) par rapport au nombre de nouveaux liens inférés à partir des classes de synonymes (148 liens). Il s'avère que l'expert, lors de la construction des classes, n'a pas eu le souci d'en faire la clôture inférentielle : il privilégie la cohérence intrinsèque des classes. Ceci souligne la complémentarité des deux démarches, humaine et algorithmique, pour la détection des liens de synonymie. L'expert a validé 17 des 19 liens inférés également présents dans les classes sémantiques : *conditions de fonctionnement / régimes de fonctionnement*, *classe sismique / classification sismique*, *limitation de durée / limitation de temps*. Le typage de ces liens se répartit équitablement entre la relation de synonymie et la relation Voir-Aussi.

	Nombre de liens inférés également construits par l'expert	Nombre de liens inférés non construits par l'expert
Classes de synonymes	19	148
Dictionnaire	18	572
Dictionnaire puis classes	32	724

Tableau 2 :
Proportion de liens construits par l'expert parmi les liens inférés

On observe par ailleurs que peu de liens inférés à partir du dictionnaire (18 sur un total de 590) sont donnés par l'expert dans les classes de synonymes. Il y a donc beaucoup de liens qui sont inférés à partir du dictionnaire que l'expert valide effectivement comme liens de synonymie quand on les lui soumet mais qu'il n'a pas pensé à inclure dans ses classes de synonymes. Ceci s'explique par le fait que l'expert travaille de manière privilégiée sur la langue technique. De son propre aveu, rechercher les liens de langue générale est un surcroît de travail. Ceci souligne l'intérêt spécifique des ressources générales. Parmi les liens proposés par l'expert dans les classes sémantiques, tous les liens sont validés : *contrôles périodiques / inspections périodiques*, *baisse de pression / réduction de pression*. Cependant, peu de liens expriment des relations de synonymie (3 liens). Il s'agit essentiellement de liens de type Voir-Aussi (15 liens). La proportion de ces liens inférés à partir du dictionnaire et présents dans les classes est également faible au regard du nombre total de liens dans les classes.

Le recouvrement entre les deux sources lexicales est donc faible : les données spécialisées fournies par l'expert et les données issues du dictionnaire de langue apparaissent largement complémentaires.

4.2 Le thesaurus, un apport plus marginal

De la même manière, le recouvrement entre les liens inférés à partir du dictionnaire ou des classes et les liens donnés par le thesaurus est

	Nombre de liens inférés déjà présents dans le thesaurus	Nombre de liens inférés non présents dans le thesaurus
Dictionnaire	2	588
Classes de synonymes	1	166
Dictionnaire puis classes	2	754

Tableau 3:
Proportion de liens du thesaurus parmi les liens inférés

très faible. Ne figurent dans le thesaurus que deux liens inférés à partir des liens extraits du dictionnaire et un seul lien inféré à partir des liens construits par l'expert (voir le tableau 3): *circuit de réfrigération / circuit de refroidissement* et *appareil de mesure / dispositif de mesure*. Le premier lien est validé comme un lien de synonymie alors que le second est retenu comme un lien de type Voir-Aussi. Dans le cas du thesaurus, ce faible recouvrement indique en fait un intérêt très marginal pour la détection de liens de synonymie entre termes.

Nous pouvons avancer deux types d'explications qui demandent à être confirmées par un expert du domaine: la couverture et la normalisation du thesaurus. Le thesaurus semble ne couvrir que très partiellement le domaine du corpus. Nous ne retrouvons que 28 liens du thesaurus parmi les classes de synonymes, qui ont été construites à partir d'un corpus du domaine. La couverture est d'autant plus faible que les termes présents dans le thesaurus sont en grande partie complexes. De surcroît, la faible productivité du thesaurus pour la détection de liens de synonymie montre que les liens initiaux ne reflètent souvent pas des liens terminologiques du domaine. Les termes liés dans le thesaurus n'entrent pas dans les constructions parallèles recherchées par les règles d'inférence.

Les liens du thesaurus ont été construits afin d'organiser conceptuellement les termes en fonction de différents domaines d'activité. Il est possible que les objectifs de normalisation sous-jacents dans le thesaurus rendent difficile son utilisation pour le traitement de corpus. Si les candidats termes extraits automatiquement sont en fait des variantes des formes normalisées du thesaurus, il faut mettre en œuvre des techniques plus lourdes et plus complexes telles que la génération de variantes (Jacquemin 1997) pour inférer des liens de synonymie à partir du thesaurus.

À maints égards, le thesaurus semble avoir un statut intermédiaire entre le dictionnaire de langue et la source spécialisée que sont les classes de synonymes. Malheureusement, dans la visée qui est la nôtre, il combine les faiblesses de l'un et de l'autre. Le thesaurus, après l'étape du filtrage, ne fournit qu'un nombre réduit de liens initiaux (comme les classes de synonymes) mais, à la différence des liens initiaux des classes, ceux du thesaurus ne sont pas directement exploitables pour le corpus et ont une faible productivité (ce qui rapproche le thesaurus et le dictionnaire).

Cette confrontation des résultats obtenus en utilisant des sources lexicales de natures différentes confirme l'intérêt des dictionnaires de langue pour le traitement de corpus spécialisées. Quand elles existent, les ressources spécialisées construites manuellement sont souvent incomplètes du point de vue de la synonymie. Un dictionnaire de langue, à la différence d'une source lexicale comme le thesaurus EDF, fournit un apport numériquement conséquent et qualitativement complémentaire.

5 Le rôle des ressources lexicales dans l'acquisition de connaissances spécialisées

Ces dernières années ont vu se développer les travaux portant sur le traitement automatique des textes techniques en vue de l'acquisition de connaissances spécialisées (lexicales, terminologiques ou ontologiques). Les méthodes reposant sur les seules données du corpus ayant montré leurs limites, ces travaux reposent souvent sur une approche mixte combinant des sources de connaissances préexistantes et des corpus spécialisés. Ils diffèrent cependant par le type des sources utilisées.

Il s'agit le plus souvent de sources spécialisées. Naulleau *et al.* (1996) utilisent certaines informations sémantiques du thesaurus EDF pour construire des classes de termes dans une perspective de filtrage de documents. Morin (1998) exploite des liens d'hyponymie d'un thesaurus d'agronomie pour amorcer l'acquisition de nouveaux liens à partir de corpus. Habert *et al.* (1998) confrontent à un corpus portant sur la médecine coronarienne une nomenclature médicale assez générale pour l'étendre et l'adapter à ce domaine particulier. Maynard *et al.* (1998) désambiguïsent les termes d'un corpus médical en s'appuyant à

la fois sur leurs distributions en corpus et sur un thesaurus médical.

Les données de la langue générale sont plus rarement utilisées, l'idée qu'elles sont peu pertinentes pour les textes techniques étant assez largement admise. Basili *et al. (*1997) montrent cependant comment la confrontation de *WordNet* et d'un corpus technique permet de construire un *WordNet* spécialisé, adapté au domaine considéré. Nous avons tenté de montrer que la contribution d'un dictionnaire de langue à la structuration d'une terminologie spécialisée est réelle (Hamon *et al.* 1998). Au regard de ces expériences mais aussi de Habert *et al.* (1998), il ressort que les données lexicales générales sont utiles dès lors qu'elles sont contrôlées par des attestations en corpus.

En pratique, c'est souvent faute de ressources spécialisées adéquates que l'on en vient à exploiter des sources lexicales générales. Peu de travaux ont exploré les deux pans de l'alternative et cherché à caractériser la contribution respective des sources lexicales spécialisées et des sources lexicales décrivant la langue générale.

C'est ce que nous avons tenté ici dans le cadre d'une expérience particulière. Des données lexicales générales et spécialisées sont combinées et projetées sur un corpus pour élaborer de nouvelles connaissances spécialisées. Les résultats montrent la complémentarité de ces sources pour la structuration de la terminologie du corpus étudié.

6 Conclusion

L'utilisation de ressources lexicales de différents types dans une méthode d'inférence de liens sémantiques permet de détecter un nombre important de liens sémantiques entre des candidats termes extraits d'un corpus technique. L'étude et la confrontation des résultats suivant le type de ressources a montré la complémentarité de données de la langue générale, comme *Le Robert*, et de données très spécialisées construites manuellement par un expert du domaine.

De plus, le faible recouvrement entre les liens inférés à l'aide de ces deux ressources et les liens présents dans la source de statut intermédiaire, le thesaurus EDF, justifie l'utilisation d'informations de la langue générale pour la structuration d'une terminologie du domaine étudié. Les résultats de cette expérience laissent penser que dans le contexte d'un système d'accès à l'information, il est probablement préférable de constituer manuellement de petites ressources très spécialisées et adaptées au corpus plutôt que d'exploiter un thesaurus lui-même coûteux à maintenir. L'utilisation d'un thesaurus spécialisé n'apporte pas une aide significative à la structuration d'une terminologie alors que des informations issues d'un dictionnaire de langue largement disponible, combinées à des petites ressources construites à partir du corpus, permet d'obtenir de meilleurs résultats et une relativement bonne couverture du corpus étudié.

Ce travail appelle un double prolongement. En montrant la nécessité de combiner différentes ressources pour la structuration de terminologies, cette étude fait apparaître deux aspects cruciaux des outils d'aide à la construction de terminologie. La réutilisation des données exploitées dans cette expérience a demandé la mise au point de quelques algorithmes. L'intégration de connaissances hétérogènes pour l'extraction d'information étant un problème en soi, il est nécessaire de proposer des outils et algorithmes permettant d'assurer la cohérence de ces informations lorsqu'elles sont utilisées conjointement. De plus, afin d'aider le terminologue au moment de la validation mais aussi lors de l'évaluation, il est important de structurer les résultats en fonction de mesures de pertinence. Cette classification des liens inférés doit tenir compte des validations et permettre également d'éliminer rapidement un certain nombre de liens erronés.

Remerciements

Ce travail est le fruit d'une collaboration avec la DER-EDF. H. Boccon-Gibod, Y. Abbas, M.-L. Picard (DER-EDF) et D. Bourigault (CNRS) ont mis leurs données et outils à notre disposition. Ce travail a bénéficié des discussions que nous avons avec eux ainsi qu'avec C. Jacquemin (Limsi) et B. Habert (UMR 8503).

Thierry Hamon,
Laboratoire d'informatique de Paris-Nord,
Université Paris-Nord,
Villetaneuse,
France.

Daniela Garcia,
Direction des études et recherches d'électricité de France,
Clamart,
France.

Adeline Nazarenko,
Laboratoire d'informatique de Paris-Nord,
Université Paris-Nord,
Villetaneuse,
France.

Bibliographie

Assadi (H.), 1997 : « Knowledge acquisition from texts : Using an automatic clustering method based on noun-modifier relationship », dans *Proceedings of the 35th Annual Meeting of the ACL - Student Session, Madrid, Spain.*

Basili (R.), Pazienza (T.) et Velardi, (P.), 1997: «Acquisition of selectional patterns in sublanguages», dans *Machine Translation*, n°8, p. 175-201.

Bourigault (D.), 1994: *Lexter, un logiciel d'extraction de terminologie. Application à l'extraction des connaissances à partir de textes.* Thèse de mathématiques, informatique appliquée aux sciences de l'homme, EHESS, Paris.

Cruse (D. A.), 1986: *Lexical semantics*, Cambridge University Press.

Dagan (I.) et Church (K.), 1994: *Termight*: «Identifying and translating technical terminology», dans *Proceedings of ANLP'94, Stuttgart, Germany*, p. 34-40.

Garcia (D.), 1998: *Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système informatique Coatis*, Thèse de doctorat nouveau régime en informatique, Université de Paris-Sorbonne, Paris.

Gros (C.) Assadi (H.), Aussenac-Gilles (N.) et Courcelle (A.): «Task models for technical documentation accessing», dans *Proceedings of the 9th European Workshop on Knowledge Acquisition (EKAW'96), Nottingham*.

Gros (C.) et Assadi (H.), 1997: «Intégration de connaissances dans un système de consultation de documentation technique», dans *Actes de ISKO'97*, Presses universitaire du Septentrion.

Habert (B.), Nazarenko (A.), Zweigenbaum (P.) et Bouaud (J.), 1998: «Extending an Existing Specialized Semantic Lexicon», dans *Proceedings of LREC'98, Granada*, p. 663-668.

Hamon (T.), Nazarenko (A.) et Gros (C.), 1998: «A step towards the detection of semantic variants of terms in technical documents», dans *Proceedings of Coling-ACL'98, Montreal, août 1998*, p. 498-504.

Jacquemin (C.), 1997: *Variation terminologique: Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*, Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes, Nantes.

Maynard (D.) et Ananiadou (S.), 1998: «Acquiring Contextual Information for Term Disambiguation», dans *Proceedings of the First Workshop on Computational Terminology, Montreal, August*, p. 86-90.

Miller (G. A.), Beckwith (R.), Fellbaum (C.), Gross (D.) et Miller (K.), 1993: *Introduction to WordNet: An on-line lexical database*, Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton.

Morin (E.), 1998: «Prométhée, un outil d'aide à l'acquisition de relations sémantiques entre termes», dans *Actes de la Conférence TALN 1998, Paris*.

Naulleau (E.), Monteil (M.-G.) et Habert (H.), 1996: «Recycling an existing thesaurus to characterize and process terms», dans *Proceeding of Euralex'96, Göteborg*.

Ploux (S.) et Victorri (B.), 1998: «Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes», dans *Revue Tal*, vol. 39 n°1, p. 161-182.