Revue québécoise de linguistique

Procédures de désambiguïsation pour les systèmes de recherche d'information

Pierre-André Buvet, Fabienne Moreau et Max Silberztein

Volume 32, numéro 1, 2003

TALN, Web et corpus

URI : https://id.erudit.org/iderudit/012249ar
DOI : https://doi.org/10.7202/012249ar

Aller au sommaire du numéro

Éditeur(s)

Université du Québec à Montréal

ISSN

0710-0167 (imprimé)
1705-4591 (numérique)

Découvrir la revue

Citer cet article

Buvet, P.-A., Moreau, F. & Silberztein, M. (2003). Procédures de désambiguïsation pour les systèmes de recherche d'information. *Revue québécoise de linguistique*, *32*(1), 177–197. https://doi.org/10.7202/012249ar

Résumé de l'article

Nous discutons de la nécessité de tenir compte de la polysémie nominale pour les systèmes de recherche d'information qui tiennent compte du contenu des textes numérisés. Nous présentons un prototype qui fonctionne en identifiant les substantifs d'un texte donné et en stipulant les domaines qui leur sont rattachés afin de faire ressortir une dominante, et ainsi de procéder au typage du texte en termes de domaine. Ce prototype a pour principale particularité d'utiliser le système INTEX et de faire appel aux descriptions formalisées du français effectuées au Laboratoire de Linguistique Informatique implémentées sous forme de dictionnaires électroniques et de grammaires locales. Nous montrons comment INTEX, en s'appuyant sur ces dictionnaires et ces grammaires, peut lever des ambiguïtés relatives à des substantifs.

Revue québécoise de linguistique, vol. 32, nº 1, 2003, © RQL (UQAM), Montréal

PROCÉDURES DE DÉSAMBIGUÏSATION POUR LES SYSTÈMES DE RECHERCHE D'INFORMATION

Pierre-André Buvet
Université de Paris XIII
Fabienne Moreau
Max Silberztein
Université de Franche-Comté

1. Introduction

L'un des objectifs des spécialistes du Web est la mise en œuvre de systèmes de recherche d'informations qui permettent de traiter efficacement l'hétérogénéité des innombrables documents présents sur le réseau. Les systèmes développés jusqu'à présent connaissent des limites liées à la notion de pertinence, c'est-à-dire qu'ils ont tendance à proposer aux utilisateurs n'importe quelle information au lieu de leur procurer celle qui répond précisément à leur attente. C'est pourquoi le traitement de l'information numérisée est en pleine évolution : il est de moins en moins question d'assimiler les langues naturelles à de simples assemblages de chaînes de caractères qui donnent lieu à des opérations de nature essentiellement statistique. Le fait d'admettre l'inanité des traitements en surface des documents a pour corollaire la prise en compte de leur contenu linguistique. De ce point de vue, les chercheurs en TALN ont un rôle majeur à jouer.

Le traitement du langage dans les systèmes de Recherche d'Informations peut s'effectuer de deux façons. L'une, d'inspiration cognitiviste, fait appel principalement à des ontologies où les mots sont structurés en fonction de relations sémantiques (Poibeau 2003)[1]. Dans une langue à caractère fortement flexionnel comme le français, cette manière de procéder suppose préalablement

1 Ces relations sémantiques sont principalement la synonymie contre l'antonymie, l'hyperonymie contre l'hyponymie ou la méronymie contre l'holonymie Ainsi, dans le domaine de la chimie, le mot *corrosion* est associé à *réaction chimique* par hyperonymie, à *corrosion électrochimique* par hyponymie et à *oxydation* par synonymie.

un traitement morphosyntaxique des unités lexicales afin de faire apparaître des formes lemmatisées dans le document. L'avantage de cette approche est de correspondre à la pratique des thesaurus en documentation. Son inconvénient est de ne pouvoir fonctionner que sur des univers restreints et donc de ne pas être adapté au Web, qui se caractérise, entre autres, par son extrême diversité thématique. L'autre façon est indépendante du thème des documents. Quels qu'ils soient, il s'agit de les analyser linguistiquement de telle sorte qu'ils soient sémantiquement caractérisés. Son intérêt est donc de pouvoir traiter la variété documentaire et de contribuer efficacement à son indexation. L'efficacité du traitement dépend de sa robustesse, qui est elle-même tributaire de dictionnaires à large couverture.

IRIS (INTEX pour la Recherche d'Informations Spécialisées), le prototype de système de recherche d'informations dont il est question ici, ressortit à la deuxième approche. Son mode de fonctionnement consiste notamment à identifier les substantifs d'un texte donné et à stipuler les domaines qui leur sont rattachés afin de faire ressortir une dominante, et ainsi de procéder au typage du texte en termes de domaine (cf. infra). Du point de vue de son architecture, IRIS a pour principale particularité d'utiliser le système INTEX et de faire appel aux descriptions formalisées du français effectuées au LLI (Laboratoire de Linguistique Informatique) implémentées sous forme de dictionnaires électroniques et de grammaires locales (cf. infra).

D'une façon générale, le figement, la polymorphie et la polysémie sont des caractéristiques des langues naturelles qui constituent autant de difficultés majeures pour les systèmes opérant sur des données linguistiques. Pour les résoudre, les systèmes doivent comporter des descriptions formalisées s'appuyant sur des théories linguistiques qui tiennent compte de ces phénomènes. L'une de ces théories est le modèle des classes d'objets (cf. infra). Pour l'élaboration du système IRIS, c'est ce modèle qui a été choisi, car il a permis de réaliser des descriptions formalisées à large couverture. Si la résolution des difficultés afférentes au figement dépend en grande partie de la réalisation des bases de données lexicales, celle des difficultés imputables à la polymorphie ou à la polysémie implique aussi la prise en compte obligatoire de l'environnement des formes problématiques. Il s'ensuit que les systèmes doivent comporter des modules dédiés à l'analyse syntacticosémantique des différentes phrases qui composent un texte numérisé donné. Nous montrons ici comment le prototype IRIS envisage de traiter les phénomènes de polysémie nominale dans la mesure où un tel traitement est indispensable à son bon fonctionnement.

Après avoir discuté de la nécessité de traiter la polysémie du point de vue de la recherche d'informations, nous indiquons quelles composantes du prototype IRIS prennent en charge les procédures de désambiguïsation. Nous évoquons ensuite les descriptions formalisées du français réalisées au LLI puis

nous montrons comment le système INTEX a recours à ces descriptions pour lever des ambiguïtés relatives à des noms.

2. Problématique

Tout système opérant sur des données linguistiques considérables doit tenir compte de la polysémie du fait qu'il s'agit d'une caractéristique majeure des langues naturelles. Dans le domaine de la recherche d'informations, c'est essentiellement la polysémie nominale qui doit être traitée dans la mesure où, indépendamment de l'approche adoptée, la catégorisation des textes numérisés résulte de l'interprétation sémantique des différents substantifs qu'ils comportent. Les propriétés sémantiques des noms sont de deux ordres; elles correspondent soit à leur sens, soit à leur domaine. Si la polysémie, par définition, a trait au premier type de propriété, par contre, elle concerne indirectement le second type de propriété, le domaine d'un substantif donné pouvant varier selon son emploi. Du point de vue des systèmes du type IRIS, les spécifications sémantiques majeures sont les domaines des substantifs puisqu'ils contribuent directement au typage des documents (cf. infra)[2]. Pour cela, il est nécessaire que le sens des noms soit totalement désambiguïsé. C'est pourquoi il est prévu que le prototype IRIS recoure à des procédures de désambiguïsation.

Les classes d'objets étant des descriptions formalisées dont le rôle est de traiter la polysémie, il sera fait appel à elles. Il n'est pas question ici de montrer comment désambiguïser les noms à l'aide des classes, dans la mesure où cela a été déjà fait (Gross 1994), mais comment traiter la polysémie à l'aide d'INTEX en prenant appui sur ces descriptions formalisées. Auparavant, nous précisons quels sont les rapports entre les sens et les domaines des noms dans la mesure où ils justifient la mise en place du dispositif que nous présentons.

La notion de polysémie varie selon les théories linguistiques. Il s'ensuit des différences de traitement en fonction des modèles. Admettons que la polysémie est la possibilité d'attribuer différentes valeurs à une même forme. Deux types de polysémies sont alors concevables (Le Pesant et Mathieu-Colas 1998: 20-22) :

2 Il va de soi que le sens des noms contribue également à caractériser les documents. Ainsi, des textes peuvent être rassemblés sur la base non plus d'un domaine commun mais d'un sujet commun, par exemple les félins. L'identification de tous les noms d'animaux relatifs à cette catégorie dans des textes est alors suffisante pour procéder au typage des documents. Il s'agit d'une stratégie de recherche d'informations différente de celle qui est évoquée ici. Elles sont évidemment complémentaires et l'implémentation de la stratégie basée sur le sens des noms pourrait, à terme, être un prolongement de celle fondée sur le domaine des noms dont il est question ici.

a) la polysémie lexicale, p. ex. le nom *caporal* :

(1) *Luc s'est adressé au caporal* (*caporal* est un nom de <grade >)

(2) *Luc fume du caporal* (*caporal* est un nom de <tabac>)

b) la polysémie régulière, p. ex. le nom *porte* :

(3) *La porte est fracturée* (*porte* est un nom <<inanimé concret>>)

(4) *Luc a franchi la porte* (*porte* est un nom <<locatif>>)

Dans les deux cas, le modèle des classes d'objets considère qu'il est possible de lever les ambiguïtés rattachées aux substantifs, entre autres, si l'on tient compte de leur combinatoire dans la mesure où il est postulé que la sémantique procède de la syntaxe, c'est-à-dire que le sens d'une unité lexicale s'explique par un environnement qui lui est propre.

La notion de domaine, telle que nous l'envisageons (cf. Buvet et Mathieu-Colas 1999), est à priori indépendante de celle de polysémie puisqu'il s'agit du «champ d'expérience dont relève le mot» (Quemada et coll. 1984)[3]. Pour autant, le traitement de la polysémie lexicale est absolument nécessaire pour caractériser un texte à partir des domaines de ces substantifs, car ces derniers sont le plus souvent ambigus[4]. Il s'ensuit qu'ils peuvent ne pas avoir le même domaine selon leur emploi[5].

Le tableau ci-dessous récapitule les différentes situations possibles pour les substantifs du double point de vue la polysémie et des domaines.

3 Dans les dictionnaires du LLI, les spécifications de domaines sont complétées par celles de registres de telle sorte qu'on puisse préciser si un mot relatif à un domaine donné relève, ou non, d'une langue spécialisée. De ce fait, des mots comme *ictère* et *jaunisse* sont l'un et l'autre décrits comme des noms du domaine de la médecine, mais distingués du fait que le premier est présenté comme un terme spécialisé.

4 Les textes dont les noms présentent un faible taux de polysémie sont des textes extrêmement spécialisés (Lerat 1995). Ce sont de tels documents que peuvent traiter les systèmes de recherche fondés sur une approche cognitiviste de la langue, car ils ne tiennent pas compte de la polysémie. Or, ces documents sont loin d'être représentatifs de ceux qui figurent sur le Web.

5 De ce point de vue, le traitement de la polysémie régulière devrait également être pris en compte, car un substantif n'a pas toujours le même domaine selon ses variations de sens. Ainsi, le nom de pays *France* en tant que locatif relève plutôt du domaine de la géographie ou du tourisme (*J'ai visité la France*); par contre, lorsque ce substantif correspond à un humain collectif, son domaine est plutôt l'histoire ou la politique (*La France n'a jamais failli à ses engagements*).

Tableau 1

POLYSÉMIE	DOMAINE	EXEMPLE
–	0	*certitude* (langue générale)
–	1	*écosystème* (écologie)
–	>1	*drogue dure* (médecine et société)
+	0	*amour*1 (langue générale : *Luc éprouve de l'**amour** pour Léa*) *amour*2 (langue générale : *Luc a l'**amour** des belles choses*) *amour*3 (langue générale : *Luc est un **amour***)
+	1	*bain-marie*1 (cuisine : *Luc l'a cuit dans le **bain-marie***) *bain-marie*2 (cuisine : *Luc l'a cuit au **bain-marie***)
+	>1	*arbre*1 (agriculture) *arbre*2 (technique3) *arbre*3 (linguistique) …

L'objectif du prototype IRIS est de typer un texte numérisé en comptabilisant les domaines de ses noms mis en évidence à l'aide des dictionnaires du LLI. De ce point de vue, les trois premiers cas de figure ne sont pas équivoques, il suffit que les substantifs soient identifiés pour leur attribuer un ou plusieurs domaines[6]. Les deux cas suivants peuvent également accepter ce type d'opération, car les différentes acceptions correspondent à un même domaine. En revanche, le dernier cas pose des problèmes d'analyse pour ce type d'opération, car chaque emploi est associé à un domaine distinct. Pour attribuer au substantif le domaine adéquat, il faut d'abord spécifier son emploi en procédant à la désambiguïsation. Avant de traiter ce point, nous présentons le prototype IRIS puis les descriptions formalisées du langage auxquelles il fait appel.

3. Le prototype IRIS

IRIS est un prototype de moteur de recherche d'informations développé à l'Université de Franche-Comté (Moreau 2000). Il vise à la caractérisation des

6 Par simplification, on assimile la langue générale à un domaine, ce qui est inexact d'un point de vue linguistique.

documents en termes de domaines. Les domaines sont encodés d'une façon normalisée dans les dictionnaires électroniques du LLI. Leur projection sur un document permet de le typer compte tenu des fréquences les plus élevées. Nous exposons brièvement le fonctionnement du système puis nous présentons son architecture.

3.1 Fonctionnement

Nous mentionnons les principales étapes qui caractérisent le fonctionnement de IRIS. Un schéma illustre chacune des ces étapes.

Étape 1

En premier lieu, lorsque l'utilisateur a formulé une requête sous une forme nominale, IRIS lui propose de préciser la nature des domaines des documents où apparaît la forme utilisée. Pour ce faire, il consulte les dictionnaires du LLI implémentés et indique tous les domaines associés à l'ensemble des emplois du substantif en question. Supposons qu'un utilisateur tape *effet de serre*. Étant donné que cette forme nominale complexe est associée dans un dictionnaire aux domaines 'écologie', 'économie', 'géoscience' et 'politique', il lui est proposé de choisir parmi ces domaines lequel concerne les documents recherchés.

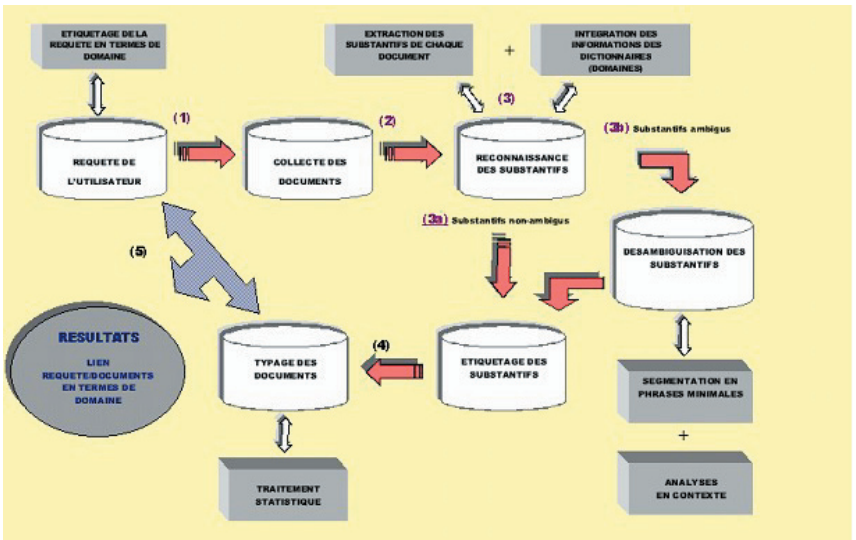

Fig. 1 : Architecture du système Iris

Étape 2

Indépendamment de la première étape, si ce n'est qu'il est tenu compte de la requête formulée sous forme nominale, IRIS fonctionne comme un moteur de recherche classique en allant récupérer sur le Web tous les documents comportant *effet de serre* en tant qu'ensemble de chaînes de caractères.

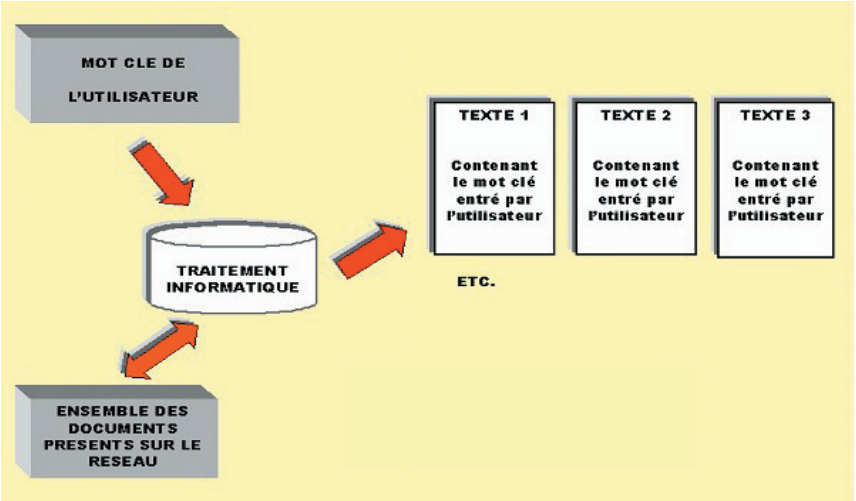

Fig. 2 : Collecte des documents

Étape 3

Chacun des documents ainsi récupérés est ensuite traité par le système INTEX afin d'identifier l'ensemble des substantifs qu'il comprend. La liste des noms ainsi obtenue est ensuite confrontée aux dictionnaires du LLI afin d'associer les substantifs à un ou plusieurs domaines. C'est à ce niveau que se situe le problème de la polysémie, car la confrontation aboutit obligatoirement à des échecs lorsque les noms du texte correspondent à plusieurs entrées d'un dictionnaire. Pour l'instant, cette difficulté est négligée. Le traitement de la polysémie nominale est présenté dans la partie 5.

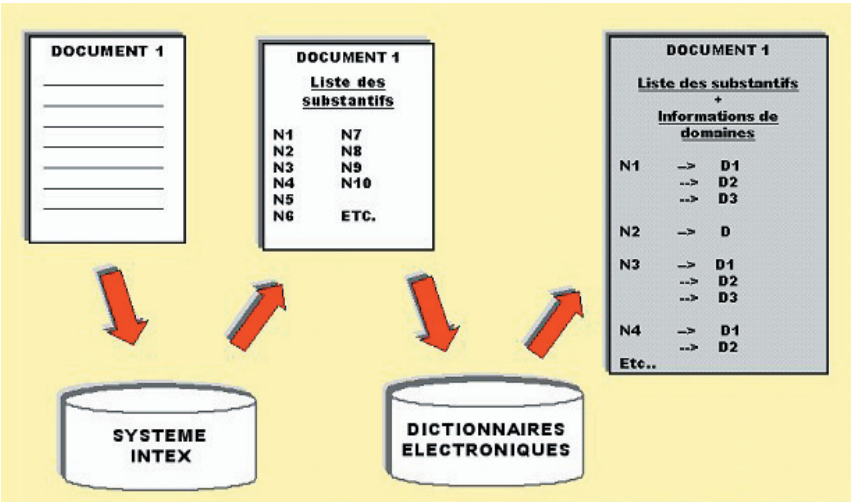

Fig. 3 : Reconnaissance des substantifs

Étape 4

Nonobstant les réserves mentionnées, les documents sont ensuite typés en fonction de l'ensemble des indications de domaine. La conception du prototype IRIS résulte de l'hypothèse selon laquelle la teneur d'un texte est calculable en fonction des domaines des noms qu'il comporte, les plus fréquents étant révélateurs du contenu thématique du texte. Cette hypothèse a été validée expérimentalement sur de nombreux documents. À ce stade, chacun des documents provenant du Web est donc caractérisé par le domaine des noms qui apparaît le plus souvent dans le texte. Si plusieurs domaines prévalent dans un texte, il est fait état de cette variété, à condition qu'elle soit statistiquement significative, pour caractériser le texte.

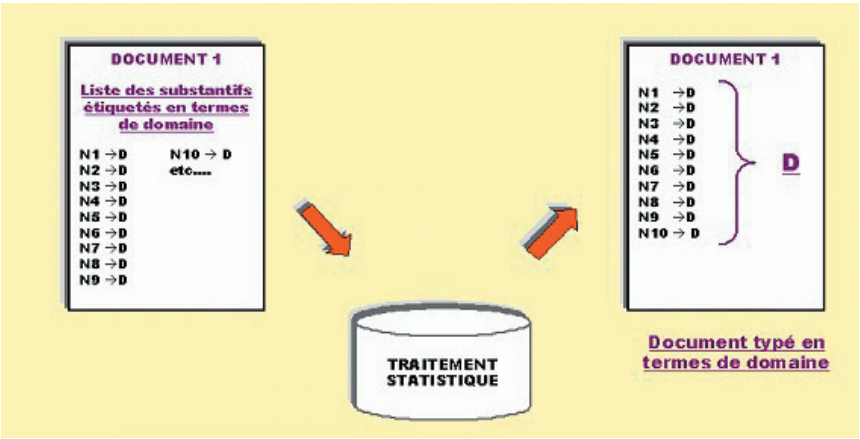

Fig. 4 : Typage des textes

Étape 5

En dernier lieu, les domaines rattachés aux documents sont comparés à celui que l'utilisateur a associé à la forme nominale qui correspond à sa requête. Par exemple, si l'utilisateur accole le domaine 'géoscience' à *effet de serre*, les documents proposés comme résultats de la requête sont uniquement les textes numérisés comportant au moins une occurrence de ce nom composé et interprétés par le prototype IRIS comme des documents relatifs à ce domaine.

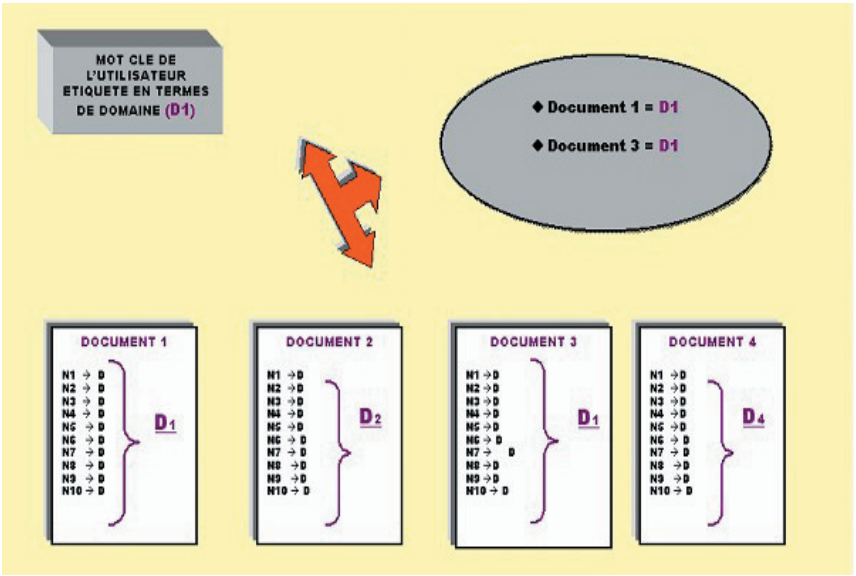

Fig. 5 : Présentation des documents à l'utilisateur

3.2 Architecture

Le prototype IRIS comporte six modules de traitement. Cinq d'entre eux sont opérationnels (Moreau 2000) : les modules de requête, de collecte, de reconnaissance des substantifs, d'étiquetage des substantifs et de typage des documents. Un seul module reste à mettre en œuvre : celui de désambiguïsation des substantifs. Dans la dernière partie, une expérimentation relative au fonctionnement de ce module est présentée afin de prouver la faisabilité du projet.

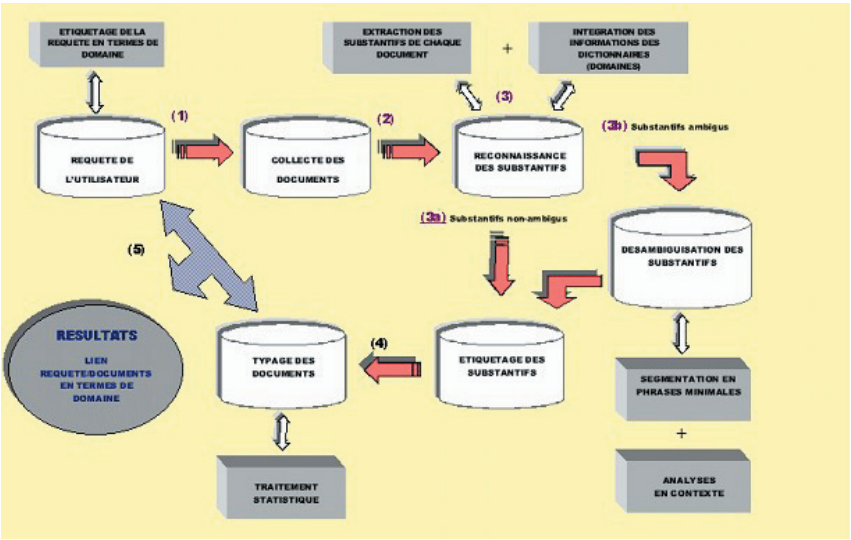

Fig. 6 : Architecture du système IRIS

Avant de présenter l'expérimentation mentionnée, nous discutons des descriptions formalisées qui sont implémentées.

4. Les descriptions formalisées du LLI

Les descriptions formalisées du LLI sont des dictionnaires électroniques et des grammaires locales. Dans un premier temps, le cadre théorique des travaux du LLI est évoqué. Il est question ensuite des principales particularités de ces descriptions : les dictionnaires sont présentés en fonction de leur type et de leur mode de structuration; nous montrons également que les classes d'objets correspondent à des grammaires locales.

4.1 Le cadre théorique

L'analyse n'est évidemment pas approfondie. Il s'agit ici uniquement d'indiquer les points importants par rapport à la conception des dictionnaires et à l'élaboration des grammaires locales.

Le projet central du LLI est de réaliser des dictionnaires de grande ampleur destinés aux divers systèmes opérant sur des données linguistiques. Ces dictionnaires visent à une couverture exhaustive du français, entre autres langues, au regard de propriétés explicites et reproductibles qui puissent faire l'objet de procédures informatisées.

Ce parti pris de description complète et robuste est nécessairement fondé sur une théorie générale du langage : en l'occurrence le modèle des classes d'objets (cf. Gross 1995, 1996, 1998; Le Pesant et Mathieu-Colas 1998). Un discours donné étant assimilé à un ensemble de phrases, il s'agit de rendre compte de la combinatoire entre les mots dans le cadre de phrases élémentaires, d'une part, de phrases complexes, d'autre part. Un tel programme de recherche nécessitant un principe organisateur, il est postulé que toute phrase est constituée d'un prédicat et de ses arguments et que les autres unités linguistiques ressortissent à l'actualisation. Dans cette modélisation de la phrase, bien que les prédicats dominent structurellement les arguments, les prédicats sont définis par leur schéma d'arguments. Quant aux différents actualisateurs, leur condition d'apparition est subordonnée aux différentes relations entre les prédicats et les arguments constitutives de phrases.

Le modèle a une conséquence importante : la division des unités linguistiques en fonction de leur statut de prédicat, d'argument élémentaire[7] ou d'actualisateur ne recoupe pas celle qui a trait aux parties du discours traditionnelles. Ainsi, les prédicats peuvent correspondre, entre autres, à :

a) des verbes (p. ex. *chérir* dans *Luc chérit Léa*)
b) des adjectifs (p. ex. *épris* dans *Luc est épris de Léa*)
c) des noms (p. ex. *béguin* dans *Luc a le béguin pour Léa*)

De même, les verbes peuvent être :

a) prédicatifs (p. ex. *gifler* dans *Léa a giflé Luc*)
b) supports (p. ex. *donner* dans *Léa a donné une gifle à Luc*)[8]

Le cadre théorique permet de prendre en compte la polysémie car, pour une forme donnée, il y a autant d'emplois prédicatifs qu'il y a de schémas

7 Les arguments élémentaires sont des noms qui ne sont jamais prédicatifs.

8 Il s'agit alors d'actualisateurs.

d'arguments[9]. Le principe des classes d'objets joue un rôle déterminant dans le traitement de la polysémie. Les classes d'objets constituent des ensembles d'items sémantiquement homogènes définis à l'aide de propriétés syntaxiques. On distingue les classes d'arguments, d'une part, les classes de prédicats, d'autre part. Les premières résultent d'une sous-catégorisation syntacticosémantique des substantifs correspondant aux arguments élémentaires[10]. Les secondes ont trait essentiellement à des adjectifs, des noms et des verbes. Ce classement est pris en compte dans la description des prédicats en termes de classes. Il s'agit cependant d'une catégorisation effectuée sur la base de leurs propriétés syntaxiques et sémantiques remarquables. De ce fait, la caractérisation syntacticosémantique d'un prédicat prime sur ses particularités morphologiques lorsqu'il recouvre deux de ces formes, voire les trois[11]; ainsi, dans *Luc ressent de la fatigue* et *Luc est fatigué,* on a affaire à un unique prédicat de la classe <sensation physique> saturé par le même argument *Luc*.

Le figement est une autre caractéristique majeure des langues naturelles dont il faut absolument tenir compte dans l'optique du TALN. Depuis de nombreuses années, le LLI a entrepris de recenser les catégories figées du français. Il en a résulté dans un premier temps la mise au point d'une typologie des noms composés fondée sur leur structure interne. Cette typologie a montré que la structure interne des noms composés est beaucoup plus riche qu'on le pensait. Plus de 700 moules de formation ont été recensés. Les composés relatifs à chacun de ces moules ont été systématiquement inventoriés. Plus de 100 000 noms de la langue générale ont ainsi été recueillis. Parallèlement, un travail similaire est en cours pour les langues de spécialités. D'autres travaux portent sur les adjectifs et les verbes composés (à ce jour 10 000 adjectifs et 30 000 verbes ont été décrits). Les recensements exhaustifs et systématiques entrepris au LLI recouvrent une grande partie des constructions figées du français. Leur spécification dans les dictionnaires électroniques est une contribution majeure au traitement du figement.

---

9 Ainsi, la combinatoire de *prendre* avec des compléments sémantiquement distincts permet de dissocier les différents sens du verbe : 1) *prendre un steak*, 2) *prendre le train* 3) *prendre l'autoroute*, etc.

10 Par exemple, les noms de <pays> comme *Allemagne*, *France*, etc. ou les noms d'<outils> comme *marteau*, *scie*, *tournevis*, etc.

11 par exemple, les phrases *Luc a désiré cela, Luc a été désireux de cela* et *Léa a eu le désir de cela* sont considérées comme strictement équivalentes dans la mesure où les arguments *Luc* et *cela* se rapportent à un même prédicat qui recouvre soit la forme verbale (*désirer*), soit la forme adjectivale (*désireux*), soit la forme nominale (*désir*).

4.2 Typologie et structuration des dictionnaires

Les réflexions théoriques qu'on vient d'évoquer ont été déterminantes pour la conception des dictionnaires (Mathieu-Colas 1994). Dans la perspective du TALN, leur élaboration implique une formalisation, une représentation systématique et structurée des données lexicales. Dans un premier temps, il est question des différentes sortes de dictionnaires mis en œuvre et, dans un deuxième temps, de la façon dont sont agencés les principaux descripteurs afférents à chaque entrée lexicale.

Les prédicats du français sont décrits dans des dictionnaires électroniques distincts selon qu'il s'agit de verbes, d'adjectifs ou de substantifs. Les prédicats étant définis par leur structure d'arguments, les dictionnaires font état pour chaque entrée non seulement de la classe sémantique du prédicat mais aussi de celles de son sujet et de ses éventuels compléments. D'autres informations relatives à l'emploi constituant l'entrée sont également précisées; entre autres, les synonymes et antonymes possibles du prédicat. Dans les cas de polymorphie prédicative, un système de pointeur permet de relier entre eux les différents dictionnaires concernés. Un dictionnaire des noms élémentaires a également été réalisé. Sa spécificité tient à la caractérisation syntacticosémantique de chaque entrée en termes de classes d'objets. Un dictionnaire des noms propres est en cours de réalisation. Dans les différentes applications rattachées au TALN, ces substantifs représentent une part non négligeable du lexique à traiter. Il convient donc de les prendre en compte systématiquement afin d'améliorer les diverses procédures d'analyse de documents. Le recensement des noms propres dans un dictionnaire électronique permet de les décrire sémantiquement en faisant appel au principe des classes d'objets.

Les différents types de dictionnaires du LLI ont donc en commun la particularité d'associer systématiquement leurs entrées respectives à des informations syntacticosémantiques formalisées. Par conséquent, leur implémentation dans les divers systèmes peut donner lieu à un étiquetage sémantique des textes complémentaire de l'étiquetage morphosyntaxique usuel. Cette façon de procéder devrait améliorer de façon significative le traitement des documents numérisés. Des informations relatives à leurs domaines d'emploi sont également rattachées d'une façon systématique aux unités lexicales dans les dictionnaires.

Les dictionnaires électroniques sont autant de bases de données où les informations sont enregistrées en fonction de champs préalablement définis. La structuration normalisée de ces informations est indispensable à l'implémentation des dictionnaires dans des systèmes opérant sur des données linguistiques. Une base de données est alors une série d'enregistrements uniformisés du point

de vue des champs qui régissent chaque type d'information donné. Le premier champ correspond à une unité lexicale, éventuellement lemmatisée. Il s'agit d'un champ clef, soit celui par lequel s'effectue l'accès aux autres champs. Ces champs comportent nécessairement des informations standardisées qui sont de deux ordres : chaque entrée est associée, d'une part, à des spécificateurs morphosyntaxiques et, d'autre part, à des spécificateurs syntacticosémantiques. Parmi ceux du premier type figure notamment un marqueur flexionnel. C'est également ici que la dichotomie entre les mots simples et les mots composés est spécifiée, de sorte que pour les composés, il est indiqué quel est leur type formel. Les indications relatives aux classes sémantiques ou aux domaines constituent l'essentiel du second type. C'est principalement à ce niveau qu'apparaissent les différences entre les dictionnaires de prédicats et les dictionnaires d'arguments dans la mesure où les premiers comportent des champs spécifiques, relatifs aux structures d'arguments et au mode d'actualisation des prédicats. Chaque emploi est décrit comme une unité lexicale autonome, tant pour la clarté de la présentation que pour sa souplesse d'utilisation, dans la perspective d'un traitement automatique. Une telle décomposition, dont l'intérêt est évident du point de vue sémantique, donne en même temps plus de rigueur à la description morphologique; ainsi, il est possible d'établir les corrélations susceptibles d'apparaître entre les propriétés morphologiques (p. ex. les variantes) et l'emploi des unités.

4.3 Classes d'objets grammaires locales

La description syntacticosémantique des unités lexicales s'effectue essentiellement au niveau des classes d'objets puisque les principales particularités de leur combinatoire sont stipulées dans les différentes classes auxquelles elles sont rattachées. La hiérarchisation des classes d'objets permet de rendre compte d'autres aspects de leur combinatoire en faisant état de leur compatibilité avec d'autres prédicats que ceux qui leurs sont appropriés. Cette hiérarchisation n'est pas une arborescence telle que les relations donnent lieu à une structuration extrêmement pyramidale. Le parti pris de définir les unités lexicales par leurs propriétés linguistiques justifie la prise en compte des héritages multiples[12]. Un autre point essentiel est qu'à un même niveau de la hiérarchie, une même unité lexicale ressortit éventuellement à plusieurs classes puisqu'une telle caractérisation permet de rendre compte des cas de la polysémie régulière[13].

12 Par exemple, la classe <boisson> est rattachée à deux classes superordonnées: <aliment>, d'une part, <liquide>, d'autre part. Il s'ensuit la possibilité de faire état de compatibilités argument-prédicat distinctes (cf. *Luc a dégusté un vin* et *Le vin s'est renversé*).

13 La prise en compte de ce phénomène permet de résoudre automatiquement l'anaphore suivante: *Ce roman est épais mais il est captivant*. (Le Pesant 1998).

La description d'une classe d'objets peut s'apparenter à l'élaboration d'une grammaire locale (Buvet et Blanco 2000). Dans la mesure où le prototype IRIS fait appel au système INTEX, cette grammaire est implémentée sous la forme d'un transducteur (Silberztein 1993, 1999). Lorsqu'il s'agit d'une classe d'arguments, le transducteur rend compte de la combinatoire de l'ensemble des substantifs avec les différents prédicats appropriés, les définitionnels et les non définitionnels; les différents éléments pris en compte sont étiquetés du point de vue de leurs propriétés syntacticosémantiques. Dans le cas où on a affaire à une classe de prédicats, le transducteur fait état également de la structure argumentale des différents prédicats et de leur mode d'actualisation (cf. Fig. 8).

L'implémentation des dictionnaires du LLI dans le système INTEX est relativement aisée. En revanche, celle des classes, sous forme de transducteurs, est loin d'être achevée, car elle nécessite toutes sortes de traitements préalables. Dans l'expérimentation ci-dessus, les grammaires locales utilisées ont été élaborées à cette occasion.

5. Levées d'ambiguïté avec le système INTEX

Il est prévu que le module de désambiguïsation du prototype IRIS utilise également le système INTEX et les descriptions formalisées du LLI pour traiter les cas de polysémie nominale. Parmi les différentes fonctionnalités d'INTEX, la fonctionnalité dite de transformation permet, d'une part, de restructurer une configuration donnée et, d'autre part, de remplacer certains de ses éléments par d'autres. Cette fonctionnalité contribue à l'analyse morphosyntaxique (cf infra) et syntacticosémantique des textes (cf. infra et Buvet 2000). Nous montrons ici comment le second type d'analyse permet de lever automatiquement des ambiguïtés. Les procédures que nous présentons seront les procédures auxquelles fera appel IRIS.

5.1 INTEX

INTEX (Silberztein 1993) est un logiciel de traitement de corpus qui donne à l'utilisateur la possibilité de développer (c.-à-d. d'éditer, de déboguer, de gérer) des ressources linguistiques à large couverture, sous la forme d'un nombre potentiellement important de dictionnaires et de grammaires électroniques, puis de les appliquer à des textes de taille importante (jusqu'à 1 Go sur des ordinateurs de type PC).

L'expérimentation a nécessité le développement d'une version étendue d'INTEX, qui a ensuite servi de base à la nouvelle version 4.32[14]. Ces versions

14 Désormais disponible à partir du site Internet : laseldi.univ-fcomte.fr.

contiennent en particulier un module morphologique qui possède des opérateurs de flexion et de dérivation, ce qui permet d'effectuer des opérations comme :

a) voler_Kfp => volées
b) cousine_fp => cousines
c) manger_Aable_ms => mangeable
d) émetteur_V_N => émettre_N => émission

Pour effectuer l'opération a), INTEX utilise un dictionnaire de type DELAF (Courtois, Silberztein et coll. 1990) dans lequel chaque verbe est associé à un transducteur qui décrit sa conjugaison. L'opération b) est similaire, mais INTEX lemmatise «cousine» avant d'effectuer la flexion. L'opération c) a été rendue possible, car INTEX dispose maintenant d'un dictionnaire dans lequel la possibilité ou non d'adjectivation en *-able* des verbes est décrite (Leeman et Meleuc 1990). De même, la description de la nominalisation des verbes dans le dictionnaire permet l'opération composée d) (Giry-Schneider 1978, 1987); noter que le nom *émetteur* est codé V_N0 tandis que le nom *émission* est codé V_N dans le DELAF étendu.

Le nouveau module morphologique, utilisé par les transducteurs étendus (à variables) d'INTEX, donne la possibilité d'effectuer des analyses et générations transformationnelles. Par exemple, à partir du texte analysé : (N0 *cette affaire*) (ETRE *est*) (ABLE *risible*), la règle : **On peut $ABLE_V de $N0** produit le résultat : *On peut rire de cette affaire*.

5.2 Opérations effectuées

Pour notre expérimentation, nous avons choisi un texte numérisé comportant plusieurs occurrences de *effet de serre* et dont le thème est la 'géoscience'. L'objectif est d'établir que la désambiguïsation des substantifs contribue à une meilleure caractérisation automatique du document.

On ouvre le texte à analyser avec INTEX ; on applique ensuite les procédures par défaut de préanalyse («preprocessing») : reconnaissance des phrases, normalisation des contractions et élisions, analyse morphologique et application des dictionnaires de mots simples et composés. On applique ensuite un graphe pour reconnaître les structures argumentales suivantes :

NPred de NArgA (E + de NArgB)

=: ($_{\text{NPred}}$ *les émissions*) *de* ($_{\text{NArgA}}$ *gaz*) *des* ($_{\text{NArgB}}$ *pays en voie de développement*)

NPred (E + de NArgA) PREP PArg

=: ($_{\text{NPred}}$ *la capacité*) *de* ($_{\text{NArgA}}$ *ce moteur*) *à* ($_{\text{PArg}}$ *surchauffer*)

La Fig. 7 (infra) produit un texte normalisé, dans lequel les prédicats et leurs arguments sont écrits sous une représentation fonctionnelle, par exemple :

acquis (Sommet de la Terre); adhésion (États-Unis); augmentation (gaz à effet de serre); capacité (penser); concentration (CO2); croissance (Économie mondiale); croissance (émissions); doublement (concentration, CO2); doublement (quantité, carbone)

Les constructions verbales correspondantes, par exemple *les pays industrialisés émettent de plus en plus de CO2*, sont normalisées de la même façon grâce à l'opérateur de nominalisation $V_N.

Ces représentations fonctionnelles sont ambiguës, dans la mesure où les prédicats reconnus peuvent appartenir à plusieurs domaines sémantiques. L'opération suivante consiste donc à formaliser les contraintes syntacticosémantiques associées à chaque classe de prédicats, afin de lever le maximum d'ambiguïtés; en d'autres termes, chaque classe de prédicats doit être associée à un ensemble de graphes qui formalisent la combinatoire de ses éléments. Ainsi, le graphe suivant (cf. Fig. 8) fait partie de l'ensemble des graphes de la classe de prédicats «propriétés physiques».

Par exemple, si le nom *émission* est suivi d'un nom qui appartient à une des classes d'arguments <gaz>, <corpuscule> ou <rayonnement>t, alors il sera associé à la classe de prédicat <propriété physique>.

Après avoir appliqué ces graphes, les structures argumentales sont en principe désambiguïsées. Pour cette expérience, les ambiguïtés résiduelles sont dans leur majorité dues au fait que nous n'avons construit que les graphes correspondant à quelques-unes des milliers de classes d'objets élaborées au LLI. Mais il est inévitable que certaines ambiguïtés ne seront pas levées, ne serait-ce que parce que le contexte définitionnel de chaque classe de prédicats n'est pas toujours explicitement présent dans la phrase, par exemple dans : *Ces pays ont trop émis ces dernières années*[15]. Dans ces cas, il faut travailler au niveau du texte, plutôt qu'au niveau de la phrase. C'est ce que nous avons simulé, en prenant en compte les noms (prédicats et arguments) désambiguïsés au niveau du texte : en comptant la fréquence de chaque domaine associé à ces noms, nous obtenons les domaines les plus fréquents :

Domaine = **geosc** (géosciences) : 11 occurrences

Domaine = **environurb** (environnement et urbanisme) : 7 occurrences

Les résultats obtenus sont conformes à ceux qui étaient attendus : les domaines des noms qui sont prédominants correspondent au thème principal du document.

15 L'analyse de la phrase ne permet pas ici d'indiquer s'il s'agit d'une émission de gaz ou d'une émission radiotélévisée.

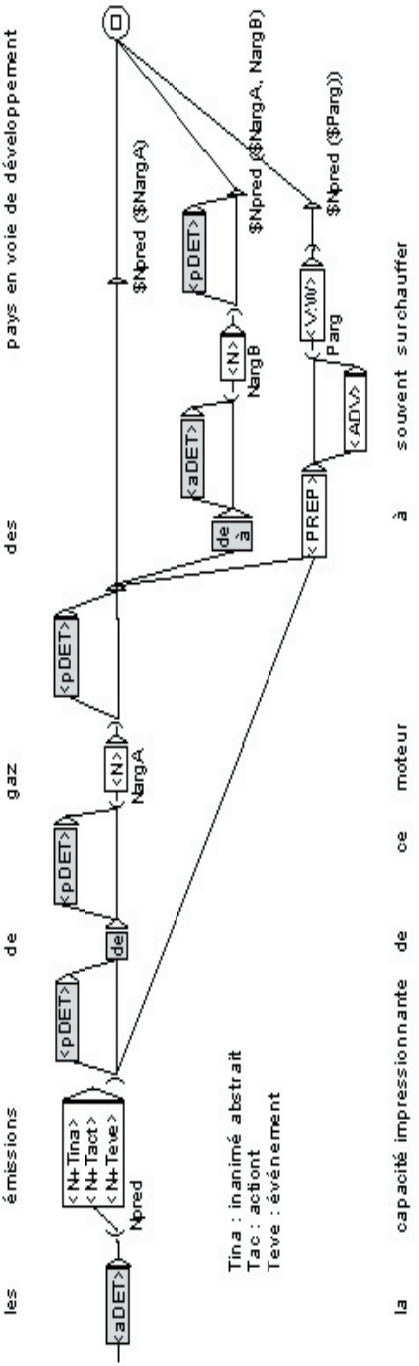

Fig. 7 : Normalisation des structures argumentales

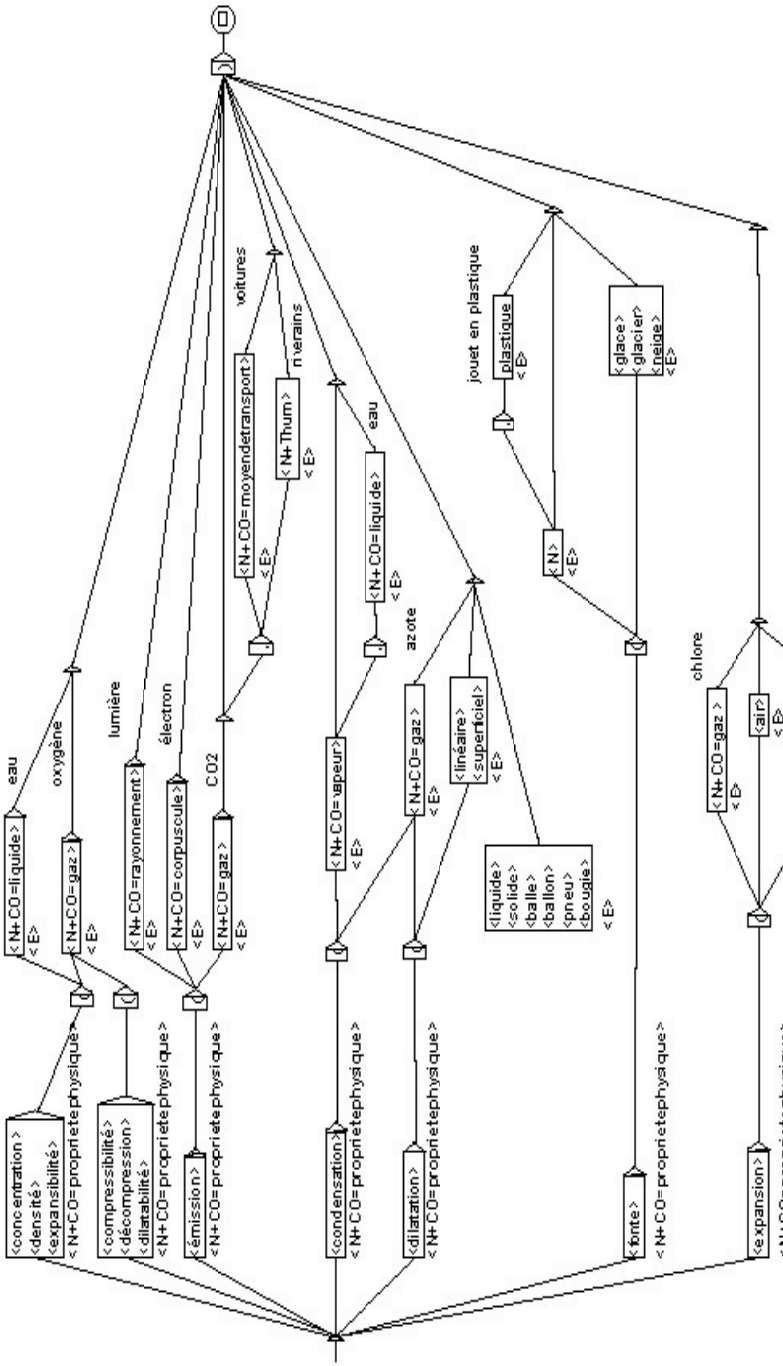

Fig. 8 : Un graphe de la classe <propriété physique>

6. Perspectives

Il faut évidemment réaliser d'autres expérimentations que celle effectuée pour prouver définitivement l'intérêt d'intégrer dans IRIS un module de désambiguïsation fonctionnant selon les principes mentionnés. Néanmoins, les résultats déjà obtenus nous semblent suffisamment probants pour envisager la mise en œuvre de ce module. Le prototype IRIS prouverait alors qu'il est possible de rendre les activités d'indexation et de recherche plus intelligentes lorsqu'il est fait appel à des outils développés en TALN. Un autre intérêt de ce module serait son éventuelle intégration dans d'autres systèmes opérant sur des données linguistiques, par exemple ceux qui ont trait à la traduction automatique.

Références

Buvet, P.-A. 2000 «Représentations métalinguistiques de phrases simples à l'aide de transducteurs», *Revue Informatique et Statistique dans les Sciences Humaines* 36 : 85-99, Liège, CIPL.

Buvet, P.-A. et M. Mathieu-Colas 1999, «Les champs domaine et sous-domaine dans les dictionnaires électroniques», *Cahiers de Lexicologie* 75-2 : 173-191, Paris, Didier.

Buvet, P.-A. et X. Blanco 2000 «De l'analyse syntactico-sémantique du lexique à la traduction automatique», *BULAG* 25 : 69-87, Besançon, PUFC.

Courtois, B., M. Silberztein et coll. 1990, *Dictionnaires électroniques du français, Langue française* 87, Paris, Larousse.

Giry-Schneider, J. 1978 *Les nominalisations en français*, Genève, Droz.

Giry-Schneider, J. 1987, *Les prédicats nominaux en français*, Genève, Droz.

Gross, G. 1994 «Classe d'objets et description des verbes», *Langages* 115 : 15-30, Paris, Larousse.

Gross, G. 1995 «Une sémantique nouvelle pour la traduction automatique : les classes d'objets», *La Tribune des Industries de la langue et de l'informatique Électronique* 17-18-19 : 16-19, Paris, Observatoire des industries de la langue.

Gross, G. 1996 «Prédicats nominaux et compatibilité aspectuelle», *Langages* 121 : 54-72, Paris, Larousse.

Gross, G. 1998 «Pour une véritable fonction synonymie dans le traitement de texte», *Langages* 131 : 103-114, Paris, Larousse.

Leeman, D et S. Meleuc 1990 «Verbes en tables et adjectifs en –able», *Langue française* 87 : 30-51, Paris, Larousse.

Le Pesant, D. 1998 «Utilisation des propriétés des anaphores dans la définition des relations lexicales», *Langages* 131 :115-126, Paris, Larousse.

LE PESANT, D. et M. MATHIEU-COLAS 1998 «Introduction aux classes d'objets», *Langages* 131 : 6-33, Paris, Larousse.

LERAT, P. 1995 *Les langues spécialisées*, Paris, PUF.

MATHIEU-COLAS, M. 1994 *Les mots français à trait d'union*, Paris, Didier.

MOREAU, F. 2000, *INTEX et la recherche d'informations spécialisées*, Mémoire de DEA, Université de Franche-Comté, Besançon.

POIBEAU, T. 2003, *Extraction automatique d'information, du texte brut au Web sémantique*, Paris, Hermès.

QUEMADA, B. et coll. 1984, *Néologismes du français actuel Datations et Documents lexicographiques Matériaux pour l'histoire du vocabulaire français* 24 :2, Paris, Klincksieck.

SILBERZTEIN, M. 1993 *Dictionnaires électroniques et analyse automatique de textes Le système INTEX*, Paris, Masson.

SILBERZTEIN, M. 1999 «INTEX: a Finite State Transducer toolbox, in Theoretical Computer», Science 231-1: 33-46, Saint-Louis, Elsevier.