Levées d'ambiguïté et domaines d'emploi

Pierre-André Buvet

LEVEE D'AMBIGUITE ET DOMAINES D'EMPLOI

Pierre-André BUVET
Centre Lucien Tesnière-LLI
Université de Franche-Comté

Dans la perspective du traitement informatique des textes en langues naturelles, le Laboratoire de Linguistique Informatique a entrepris le recensement systématique et la description exhaustive des noms simples et composés du français. Les substantifs sont relevés et décrits dans des dictionnaires électroniques - Gross G., 1991 & 1995. L'inventaire des noms composés représente une part considérable de ces travaux du fait de leur importance dans le lexique et de l'impossibilité de les analyser automatiquement à partir de leurs différents constituants. Nous examinerons ici des noms composés du type *N de N* pour constater que les indications de domaines d'emploi qui figurent dans les dictionnaires électroniques peuvent concourir, dans le cadre d'analyses automatiques de textes, à désambiguïser des unités lexicales.

Dans un premier temps, nous présenterons brièvement les données qui sont à la base de notre étude sur la polysémie nominale ; leur système de représentation, i.e. la microstructure d'un dictionnaire électronique, sera également évoqué. Dans un deuxième temps, différents types de polysémies seront examinés. Nous proposerons, ensuite, d'exploiter les domaines d'emploi comme mode de désambiguïsation en analyses automatiques de textes ; un premier état de nos recherches sera présenté.

1. Le dictionnaire électronique des noms composés *X de Y*

1.1 Les noms composés

Un dictionnaire électronique des noms composés, dits *X de Y*, constitués au minimum de deux substantifs situés de part et d'autre de la préposition *de* est en cours d'élaboration ; il comporte :

(i) des unités lexicales du type *N de (Dét + E) N* comme :

cheval de retour
gardien du temple
homme des cavernes
livre de bord
rond de cuir

(ii) des constructions plus complexes dont l'un des constituants, au moins, est soit un substantif caractérisé par un modifieur adjectival ou nominal non prédicable soit un nom composé :

accord-cadre de traité
allocation parentale d'éducation
bande d'arrêt d'urgence
personne de petite taille
première dame de France

Le terme «nom composé» employé ici a une acception large puisqu'il s'applique à des constructions que certains linguistes considèrent comme des «séquences complexes lexicalisées» -Corbin D. 1991. Le parti pris ici d'opposer les groupes nominaux libres (*la voiture de ma soeur*, par exemple) à l'ensemble des constructions plus ou moins figées résulte, d'une part, de l'importance des situations intermédiaires entre les séquences libres et les séquences totalement figées, d'autre part, de la nécessité de prendre en compte ces phénomènes dans la perspective du traitement automatique des langues naturelles - Gross M., 1988 Gross G., 1990. Ces faits de langue et leur corollaire technique expliquent la présence conjointe de suites sémantiquement opaques comme *pied de poule* et de suites lexicalisées par l'usage comme *président de la République* parmi les noms composés *X de Y*[1].

1.2. Forme d'un dictionnaire électronique

Les dictionnaires électroniques du LLI correspondent à des bases de données. Chaque «article» du dictionnaire est donc constitué d'une série de champs contenant chacun un certain type d'information formelle relatif à l'entrée lexicale. Ces informations standardisées sont essentiellement de nature morphologique, d'une part, syntactico-sémantique, d'autre part. Nous nous limiterons ici à une brève présentation des champs qui concernent la description des emplois d'une unité lexicale.

Le premier champ syntactico-sémantique est le champ [T:] dans lequel on décrit l'entrée en ayant recours aux traits suivants :

TRAIT	**EXEMPLE**	**CODE**
Animal	*éléphant de mer*	T:ani
événement	*tremblement de terre*	T:év
Humain	*chef d'orchestre*	T:hum
inanimé abstrait	*emprunt d'Etat*	T:ina
inanimé concret	*pierre d'angle*	T:inc
Végétal	*acajou de Madère*	T:vég
Locatif	*maison du peuple*	T:loc

La sous-catégorisation des noms par des traits syntactico-sémantiques est nécessaire mais non suffisante pour rendre compte de leurs emplois. Pour y remédier, les entrées lexicales sont représentées dans le champ [C:] par des classes d'objets ; "il s'agit d'ensembles sémantiquement homogènes [...] qui ont des propriétés syntaxiques spécifiques"- Gross G., 1991 :

CLASSE D'OBJETS	**EXEMPLE**
C:<moyen de transport aérien>	*avion de tourisme*
C:<outil>	*marteau de tapissier*
C:<vêtement>	*tenue de combat*

La notion de domaine d'emploi est rattachée à la description de termes plus ou moins spécialisés. Dans les dictionnaires électroniques, les informations ayant trait au domaine d'emploi et au sous-domaine d'emploi sont reparties respectivement dans les champs [D:] et [SD:]. C'est le degré de spécialité de l'entrée lexicale qui est déterminant pour le codage, ou l'absence de codage, du second champ :

DOMAINE	SOUS-DOMAINE	EXEMPLE	CODE	
médecine		*médecin de famille*	D:méd.	SD:
médecine	génétique	*aberration de nombre*	D:méd.	SD:génét.
transport		*moyen de locomotion*	D:transp.	SD:
transport	Automobile	*roue de secours*	D:transp.	SD:aut.

2. Polysémies

La question de la polysémie a fait l'objet de nombreux travaux linguistiques. C'est la pierre d'achoppement du traitement automatique des langues naturelles. Nous nous contenterons ici d'énumérer quelques situations attestant du statut ambigu de certains noms composés. Nous verrons également comment les dictionnaires électroniques traitent de la polysémie.

2.1 Constructions *N de N* ambiguës

La polysémie d'une séquence de mots peut relever de faits de grammaire. Ainsi, la suite *pas de charge* peut s'interpréter comme :

(i) une séquence **N + Prép + N** :

> *Tous ces sujets, Pierre Auger les a traités au **pas de charge**, comme pressé par le temps.* (*Le Monde* 1993)

(ii) une séquence **Adv + Prép + N** :

> *Enfin, la seule femme, Elvira Sellerio n'a **pas de charge** académique mais elle occupe depuis plus de vingt cinq ans une place de tout premier plan ...* (*Le Monde* 1993)

Les environnements syntaxiques de la suite permettent généralement à un système automatique de distinguer ces deux types de séquences (M. Silberztein, 1993).

Les seules suites nominales peuvent également prêter à confusion. Par exemple, *pommier du Japon* peut correspondre à :

(i) un seul complément:

> *Luc a planté un **pommier du Japon***

La construction *pommier du Japon* est alors considérée comme un nom composé.

(ii) une succession de compléments :

*Luc a ramené un **pommier du Japon***

La phrase accepte alors comme équivalent:

Du Japon, Luc a ramené un pommier[2]

La polysémie ne se limite pas aux seules ambiguïtés syntaxiques ; on observe également des ambiguïtés lexicales. Le traitement automatique de la polysémie par l'intermédiaire des dictionnaires électroniques implique notamment le dégroupement des emplois des mots. Les particularités syntaxiques et sémantiques qui justifient un dégroupement apparaissent aux niveaux des informations formelles standardisées situées à droite de l'entrée lexicale.

2.2 Traitement de la polysémie dans les dictionnaires électroniques

Dans les dictionnaires électroniques, les trois champs essentiels pour décrire les différents emplois d'un substantif sont les champs [T:], [C:], [D:] (Cf. Supra). Ainsi, pour décrire les situations suivantes :

Luc fréquente les bains de vapeur
Luc prend un bain de vapeur

on associe la séquence *bain de vapeur* à deux unités lexicales distinctes . Le champ [T:] rend compte de ce dégroupement de la suite nominale :

ENTRÉE LEXICALE	INDICE	TRAIT
bain de vapeur	1	T:loc
bain de vapeur	2	T:ina

La capacité descriptive des traits syntactico-sémantiques est néanmoins d'une efficacité relative, notamment, pour réduire la polysémie. D'où la nécessité de sous-catégoriser les noms en groupes spécifiques : les classes d'objets. Considérons les situations suivantes :

Luc a cassé un bain-de-pieds
Luc a récupéré le bain-de-pieds de sa tasse

Les deux emplois de *bain-de-pieds* correspondant à deux substantifs inanimés concrets, la possibilité de les distinguer n'apparaît qu'au niveau des champs [C:] et [D:] :

ENTRÉE LEXICALE	INDICE	TRAIT	CLASSE D'OBJETS	DOMAINE
bain-de pieds	1	T:inc	C:< récipient>	D: toil.
bain-de-pieds	2	T:inc	C:<liquide>	D: boiss.[3].

Du point de vue de leur pouvoir distinctif, les domaines sont généralement redondants par rapport aux classes d'objets. On observe cependant des situations où les seules informations pertinentes pour discriminer les emplois d'un mot figurent uniquement dans le champ [D:]. Il en est ainsi, par exemple, de certains emplois de *pied-de-biche* -Mathieu-Colas M. 1994 :

ENTRÉE LEXICALE	INDICE	TRAIT	CLASSE D'OBJETS	DOMAINE
pied-de-biche	1	T:inc	C:<outil>	D:arm.
pied-de-biche	2	T:inc	C:<outil>	D:horl.
pied-de-biche	3	T:inc	C:<outil>	D:manut.
pied-de-biche	4	T:inc	C:<outil>	D:outill. [4]

L'exploitation d'un dictionnaire électronique pour réduire automatiquement des polysémies d'ordre syntactico-sémantique s'appuie principalement sur les classes d'objets - Gross G. 1995. Ce point étant acquis, nous examinons à présent la possibilité de lever automatiquement des ambiguïtés en prenant en compte les domaines.

3. Désambiguïsation et domaines

Nous considérons ici uniquement les informations de domaines rattachées aux substantifs. Nous postulons qu'il est possible de désambiguïser une construction nominale polysémique *N de N* en étiquetant, à l'aide d'un dictionnaire, les domaines des substantifs du fragment de texte où figure la suite ambiguë.

Examinons l'extrait ci-dessous :

L'écurie britannique a opté pour le tout nouveau V10 du motoriste français (le Monde du 16 septembre). Ce moteur tournera au ***banc d'essai*** *fin*

décembre pour être monté sur les monoplaces un mois plus tard. [...].Alain Prost a été champion d'Europe juniors de karting en 1974. Ayrton Senna a été vice-champion du monde de cette discipline en 1979 et 1980. Mais il serait surprenant qu'il ne suive pas de très près les tests du nouveau V10 Peugeot. (*Le Monde* 1993)

Dans ce texte, la suite *banc d'essai* est répertoriée comme une construction lexicalement ambiguë. Néanmoins, on l'interprétera ici uniquement comme un terme d'automobile. Pour l'établir, nous nous proposons de ne pas tenir compte des propriétés syntaxiques[5] de ce nom composé mais d'attribuer cette signification au fait que bon nombre des substantifs de cet extrait relèvent également du domaine de l'automobile. L'étiquetage des noms communs du texte en termes de domaine d'emploi donne le résultat suivant [6] :

nom	**Dom.1**	**Dom.2**	**Dom.3**	**Dom.4**
écurie	*équit.*	*sp.+ aut.*		
motoriste	*aut.*			
banc d'essai	*mécan. ind.*	*aut.*	*langue générale*	
monoplace	*aéron.*	*aut.*		
mois	*métrol.*			
moteur	*mécan. ind.*	*aut.*		
champion d'Europe	*sp.*			
karting	*aut.*			
vice-chamion du monde	*sp.*			
discipline	*éduc.*	*sp.*		
test	*aut.*	*éduc.*	*sp.*	*langue générale*

Des trois interprétations possibles pour *banc d'essai*, c'est celle correspondant à un terme d'automobile qui est la plus plausible du fait que ce

domaine d'emploi est le plus fréquent dans l'environnement textuel de ce nom composé. Cette analyse est renforcée si l'on prend en compte les noms propres (*Alain Prost, Ayrton Sénat, V10, Peugeot*) qui sont tous des termes de l'automobile.

Pour vérifier la validité de cette hypothèse, nous avons interrogé un corpus informatique regroupant tous les articles du journal *Le Monde* parus en 1993 à l'aide du système automatique INTEX - Silberztein M. 1993. Nous présentons maintenant les résultats de nos requêtes.

3. 1. Résultats

Nous avons sélectionné dans le dictionnaire électronique des noms composés *X de Y* une série de substantifs ambigus. Le système INTEX a relevé les occurrences de ces mots dans le corpus du journal *Le Monde*. L'environnement textuel des unités lexicales étudiées est d'environ 250 caractères typographiques. Les premiers résultats sont partiellement probants.

SÉQUENCE AMBIGUË	FRÉQ	HV1	HV2	CH	REBUT
ballon d'oxygène	45	0 %	100 %	0 %	0 %
banc d'essai	19	31,6 %	42,1 %	10,5 %	15,8 %
banc des accusées	28	17,8 %	67,9 %	3,6 %	10,7 %
coussin d'air	5	100 %	0 %	0 %	0 %
force de frappe	32	21,8 %	53,1 %	9,4 %	15,6 %
pas de charge	28	0 %	92,85 %	7,15 %	0 %
ligne de fond	0				
queue de poisson	3	33,33 %	33,33 %	33,33	0 %
tableau de famille	2	0 %	50 %	50 %	0 %

Ce tableau est représentatif de l'état actuel de nos investigations. Dans la première colonne, on trouve différentes séquences *N de N* analysées par INTEX. La colonne suivante (**FREQ** : fréquence) correspond au nombre d'occurrences relevées par INTEX. Les autres colonnes font état de la

validité ou non de notre hypothèse relative à l'exploitation des domaines pour procéder à des désambiguïsations :

- dans la colonne **HP1** (hypothèse validée 1), on indique en pourcentage les occurrences où l'interprétation de la séquence *N de N* est corrélée aux domaines des substantifs de son environnement textuel comme dans l'exemple suivant :

Sa raquette produit peu souvent la caresse d'une amortie ou le claquement d'un service-volée. Ses matches ne soulèvent que très rarement de subtiles émotions. Ils sont des instants de sensations brutales, où chacun frissonne en comptant les coups et désire le voir pulvériser son adversaire. Comme il y a une ivresse des profondeurs qui menace le plongeur imprudent, il y a sur le court une ivresse de la ***force de frappe****. Jim Courier sait donner ce vertige à ses adversaires comme naguère Jimmy Connors. Thomas Muster n'a jamais eu assez de lucidité pour ne pas se laisser aspirer dans ces profondeurs. Après que l'Autrichien lui eut tenu tête une manche, l'Américain a entrepris ce travail d'aspiration vers le fond. Sur la terre battue, Jim Courier semble le seul à savoir qu'une rencontre n'est terminée qu'après le deuxième rebond de la balle de match.*(*Le Monde* 1993)

- dans la colonne **HP2** (hypothèse validée 2), on signale les constructions *N de N* dites, par convention, métaphoriques. Notre hypothèse fonctionne ici *a contrario* dans la mesure où cette interprétation est autorisée par l'absence de domaines spécifiques reliés à l'un des emplois du nom composé considéré. Ainsi, on peut distinguer quatre emplois spécialisés (dans les domaines de l'armement nucléaire, de la guerre, du tennis et de la musique), d'une part, un emploi métaphorique, d'autre part. Dans l'exemple ci-dessous, on ne relève aucun terme rattaché à l'un de ces quatre domaines ; c'est donc l'interprétation métaphorique qui prévaut :

Mais seul un juge, une fois saisi, pourrait établir l'existence d'un groupe Hersant qui, à travers des participations diverses, exercerait sa maîtrise sur un ensemble de titres. Or la désignation d'un juge transformerait l'affaire en poudrière politique. Le gouvernement actuel peut-il prendre le risque de heurter un groupe de presse qui lui a donné une dizaine de députés et qui peut manier une vraie ***force de frappe*** *éditoriale ? Si une telle nomination n'intervient pas, la FFSJ et le SNJ envisagent de faire directement appel aux tribunaux*. (*Le Monde* 1993)

- dans la colonne **CH** (contre-hypothèse), on précise en pourcentage les situations qui contreviennent à notre modèle. Ainsi, malgré de nombreux termes de musique *force de frappe* ne relève pas de ce domaine dans :

En ouverture du concert des toujours fringants That Petrol Emotion, programmé lundi 11 octobre au Passage du Nord-Ouest à Paris, les Skippies ont confirmé que le rock dur forgé en France n'avait plus à pâlir des comparaisons. Eux, comme d'autres (Deity Guns, Colm, Burning Heads), savent à présent muscler un son, le propulser avec cohérence, revendiquer un chant anglophone qui jadis nous plongeait dans l'embarras. Cette ***force de frappe*** *n'est pas encore le signe d'une originalité exclusive, cette remise à niveau ne masque pas l'évidence des emprunts et des citations. On pourrait reprocher ainsi à ce quintette rennais au patrimoine sautillant de trop bien reproduire les schémas en vogue. Soit la mise en chanson d'une énergie frénétique combinant arrogance punk, savoir-faire pop et impact métallique*. (*Le Monde 1993*)

3.2. Conclusion provisoire

Les résultats obtenus démontrent qu'un système automatique peut lever des ambiguïtés lexicales à partir des informations de domaines telles qu'elles figurent dans les dictionnaires électroniques . Toutefois, les contre-exemples relevés montrent que des dispositifs effectuant des analyses morpho-syntaxiques et distributionnelles fines n'en sont pas moins indispensables pour réduire automatiquement la polysémie.

Un système performant destiné à lever automatiquement des ambiguïtés dans des textes peut donc adjoindre à ses outils d'analyse, un procédé d'étiquetage des domaines d'emploi des substantifs par le biais d'un dictionnaire électronique de ces unités lexicales.

RÉFÉRENCES

CORBIN D. (1991) «La formation des mots : structures et interprétations», *Lexique* n° 10, Villeneuve d'Ascq, P. U. L..

GROSS G. (1990) : «Définition des noms composés dans un lexique-grammaire», *Langue française* n° 87, Paris, Larousse.

GROSS G. (1991) : «Forme d'un dictionnaire électronique», *Actes du colloque La station de traduction de l'an 2000* , Mons.

GROSS G. (1995) : «Une sémantique nouvelle pour la traduction automatique Les classes d'objets», *La tribune des industries de la langue et de l'information électronique* n° 17-18-19, Paris.

GROSS M. (1988) : "Sur les phrases figés complexes du français", *Langues française* n°77, Paris, Larousse.

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

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

---

[1] Le sens de l'expression *pied de poule*(il s'agit d'un type de tissu) n'est pas déductible à partir de ces différents constituants. Par contre, *président de la République* n'est pas sémantiquement opaque ; cependant, l'association des différents composantes de la suite n'est pas libre :

*Chirac est le nouveau (président + *chef) de la République*

*Mitterrand est l'ancien président de (la République + *l'Etat)*

[2] Dans cet exemple, la seconde interprétation n'est pas contradictoire avec la première :

De Paris, Luc a ramené un pommier du Japon

[3] Les codes **toil.** et **boiss.** symbolisent respectivement les dénominations **toilette** et **boissons**.

[4] Les codes **arm.**, **horl.**, **manut.** et **outill.** symbolisent respectivement les dénominations **armement**, **horlogerie**, **manutention** et **outillage.**

[5] Ces propriétés syntaxiques sont de nature structurale et distributionnelle.

[6] Dans ce tableau, plusieurs domaines peuvent être rattachés à une même forme nominale. Chaque domaine (dom1, dom2, dom3) correspond à un emploi spécifique de l'entréeLes codes **aéron.**, **éduc.**, **mécan. ind.**, **métrol**. et **sp**. symbolisent respectivement les dénominations **aéronautique**, **éducation**, **mécanique industrielle** et **sport**.