

Déterminer le sens d'un verbe dans son cadre prédicatif

Guillaume JACQUET, Jean-Luc MANGUIN, Fabienne VENANT & Bernard VICTORRI

Introduction

Dans son livre *La prédication verbale et les cadres prédicatifs*, Jacques François présente les résultats d'une étude de grande ampleur sur l'analyse sémantico-syntaxique des prédications verbales du français, en s'appuyant sur des modèles formels d'inspiration fonctionnaliste, essentiellement la *Functional Grammar* initiée par Dik (1978, 1997) et la *Role and Reference Grammar* de Van Valin et La Polla (1997). Il commence par proposer la définition suivante de la prédication (François 2003 p. 1):

« Elle [la prédication] est constituée du prédicat verbal et de son cadre prédicatif ('cadres des rôles actanciels', 'schéma de valence sémantique' ou dans le modèle des Principes et Paramètres 'structure argumentale') et elle est le domaine de rattachement des satellites et d'opérateurs spécifiques (par exemple des satellites de localisation spatiale et temporelle et les opérateurs de temps et de modalité objective.) »

Il définit alors 14 classes de « cadres prédicatifs » à partir de la fusion de propriétés aspectuelles (transitionalité et dynamicité) et participatives (relationalité, causativité et agentivité) des procès. Il s'attache ensuite à classer les sens de près de 400 verbes du français, tels qu'ils sont distingués par le *Petit Robert Electronique* (Rey & Rey-Debove 1996), à l'aide de cette classification des cadres prédicatifs. L'un des résultats les plus frappants de ce travail concerne ce qu'il appelle le « taux de distinction classématique », qui indique dans quelle mesure les sens différents d'un même verbe correspondent à des classes différentes de cadres prédicatifs. Pour les 66 verbes polysémiques de la base de données, la moyenne de ce taux s'élève à 81%, « ce qui signifie concrètement que dans quatre cas sur cinq, une distinction de sens identifiée par le *Petit Robert Electronique* correspond à une distinction de classe de prédication » (François 2003, p. 226).

Le programme de recherche que nous présentons ici (cf. Jacquet, Manguin, Venant & Victorri 2010 pour une première ébauche de ce programme) s'inscrit dans le prolongement de ce travail, ce qui n'est pas très étonnant dans la mesure où il a été en partie initié lors d'ateliers qui ont réuni il y a quelques années les quatre auteurs du présent article autour de Jacques François, dans le cadre d'un projet de recherche sur la prédication verbale soutenu par l'Institut de la Langue Française. Il s'agit pour nous de concevoir un système automatique capable, à terme, de calculer le sens d'un verbe dans un énoncé donné en fonction du cadre prédicatif dans lequel il se trouve, puisque, comme l'a montré Jacques François, le cadre prédicatif est la donnée essentielle que l'on doit ajouter à la donnée du verbe lui-même pour calculer son sens en contexte.

Notre approche diffère cependant de celle que Jacques François a développée dans ce livre sur deux points importants. Le premier concerne le matériel de notre étude. Jacques François a analysé les données d'un dictionnaire de langue, alors que nous travaillons directement sur de gros corpus. Cela ne veut pas dire que nous n'utilisons pas du tout de données dictionnaires: notre système comporte un dictionnaire de synonymes verbaux. Mais ce n'est qu'un outil au service de l'étude d'un corpus, qui constitue l'objet principal de notre travail. Nous nous inscrivons ainsi dans le courant que l'on a pris l'habitude d'appeler « linguistique de corpus », même si cela regroupe des travaux très divers et hétérogènes (cf. Cori & David 2008). En ce sens, notre approche est plutôt complémentaire de celle de Jacques François, puisqu'il s'agit de retrouver directement dans l'usage le bien fondé des distinctions de sens opérées par les dictionnaires.

Le deuxième point sur lequel nous nous éloignons de l'approche de Jacques François concerne le rapport à la classification, et, plus généralement, notre refus d'une discrétisation *a priori* des emplois d'un verbe donné: comme nous allons le voir, nous prônons un recours aux mathématiques du continu pour modéliser le sens des prédications verbales, même si nous ne refusons pas la possibilité d'une discrétisation après coup du modèle ainsi construit. L'élément de base de notre modèle sera une notion de distance dans un espace sémantique, cette distance étant définie de manière à modéliser le mieux

possible la notion intuitive de proximité de sens. Une fois que l'on aura calculé cette distance, on peut envisager de regrouper des sens proches dans des classes discrètes, mais ce n'est pas nécessaire au bon fonctionnement du modèle.

Le cadre théorique

On peut caractériser notre approche du sens des prédications verbales par quatre propriétés que nous allons présenter rapidement, de manière à nous situer par rapport à d'autres travaux en traitement automatique des langues portant sur ce domaine (cf. notamment Brent 1993, Briscoe & Carroll 1997, Fabre & Frérot 2002, Andrew, Grenager, & Manning 2004, et Messiant, Gábor & Poibeau 2010).

Une approche entièrement contextuelle

Considérons les énoncés suivants:

J'ai donné un poisson à mon chat

Mon oncle lègue un bateau à mes enfants

Marie va filer de l'argent au gardien

Jacques, peux-tu me passer un verre?

On s'accordera sur le fait que les prédications exprimées par ces quatre énoncés sont sémantiquement proches: il est question dans chacun de ces énoncés de prédiquer un transfert de propriété (en un sens assez large) d'un objet O entre un possesseur initial D (le donneur) et un possesseur final R (le receveur). Comme on peut le constater, cela ne signifie pas que les énoncés eux-mêmes ont le même sens, d'une part parce que ce ne sont pas les mêmes actants O, D et R qui sont évoqués dans chaque énoncé, et d'autre part parce que les valeurs temporelles, modales et assertives diffèrent aussi d'un énoncé à l'autre.

D'où provient alors la proximité sémantique de ces prédications? D'une part du fait que les quatre verbes ont des sens lexicaux proches (ils expriment tous la notion de transfert de propriété), et d'autre part parce que chacun des arguments O, D et R est instancié par des entités d'une même classe sémantique, que l'on pourrait appeler 'objet transférable' pour O, et 'possesseur potentiel' pour D et R. Le point essentiel que nous voulons souligner ici, c'est que ces deux propriétés sont contextuelles:

- Les verbes *donner*, *léguer*, *filer* et *passer* ne sont généralement pas synonymes: ils ne le sont que dans le contexte de cette construction syntaxique et de ce type d'actants. La synonymie disparaît avec d'autres types d'actants: *donner la main à quelqu'un* et *passer la main à quelqu'un* ne sont plus synonymes; et il en est de même si l'on change de construction: *donner quelqu'un* et *filer quelqu'un* ne sont pas non plus synonymes.
- les classes sémantiques auxquelles doivent appartenir les actants sont elles-mêmes contextuelles: *poisson*, *bateau*, *argent* et *verre* ne forment une classe sémantique homogène d'objets transférables que dans ce contexte prédicatif. Pour ne prendre qu'un exemple, si l'on se place dans le contexte du verbe *prendre* + COD, ces mêmes noms ne peuvent plus être regroupés dans la même classe: *prendre un poisson*, *prendre un bateau*, *prendre de l'argent* et *prendre un verre* ont des sens prédicatifs nettement distincts.

C'est la raison pour laquelle nous ne cherchons pas à définir des distances sémantiques entre unités lexicales verbales hors contexte, ni à déterminer *a priori* des classes sémantiques générales de noms. Au contraire, nous n'utiliserons de distance sémantique entre deux verbes que dans un contexte actanciel donné, c'est-à-dire dans une construction syntaxique spécifique pour chacun des verbes (pas forcément la même pour les deux verbes, cf. *jouer de la guitare* et *pratiquer la guitare*), et pour des actants de classes sémantiques données. De même, ces classes sémantiques ne sont pas définies en soi, mais elles sont spécifiques à un contexte prédicatif donné (c'est-à-dire un ensemble de verbes, chacun dans une construction spécifique, présentant des sens voisins).

Notre approche se distingue donc de nombreux autres travaux d'analyse distributionnelle par le refus de chercher à construire une ontologie générale des unités lexicales à partir des données distributionnelles.

Une approche asymétrique

Comme on l'aura remarqué, il y a une certaine circularité dans nos définitions, puisque la distance sémantique entre deux verbes dans un contexte actanciel donné fait appel à une notion de classe sémantique, qui elle-même repose sur la notion de sens voisins de verbes dans une construction donnée. Cette circularité n'est pas forcément rédhitoire: elle est de fait inhérente à la plupart des approches distributionnelles.

Un certain nombre d'approches contextuelles permettent de contourner cette circularité en construisant de manière incrémentale les classes lexicales recherchées. Généralement, la méthode utilisée est symétrique. Par exemple, on regroupe plusieurs noms parce qu'on les a trouvés fréquemment en position de COD d'un verbe donné. Puis on regroupe plusieurs verbes parce qu'ils admettent (toujours dans le corpus) tout ou partie de cet ensemble de noms comme COD. Cela permet d'affiner alors l'ensemble de noms en considérant les COD de tout ou partie de cet ensemble de verbes. Et ainsi de suite, jusqu'à ce que ces va-et-vient permettent de stabiliser une classe de noms et une classe de verbes présentant une affinité suffisante les uns avec les autres.

Ces méthodes symétriques ne conviennent pas dans notre cas. En effet, nous ne cherchons pas à construire des classes de verbes et des classes de noms compatibles dans une construction donnée, nous cherchons à trouver des prédications ayant le même sens (ou des sens proches). D'un côté, notre objectif est plus étroit: il ne suffit pas que deux verbes présentent une affinité forte avec plusieurs noms dans une construction donnée pour qu'ils soient presque synonymes. Par exemple *écrire*, *lire*, *imprimer*, *diffuser* ont tous une affinité pour des compléments d'objet de type *livre*, *journal*, *article*, etc. sans pour autant être synonymes. D'un autre côté, notre objectif est plus large: nous voulons aussi rapprocher des prédications dans lesquelles les verbes n'ont pas la même construction syntaxique, comme *jouer au tennis*, *pratiquer le tennis* et *faire du tennis*, ou encore *abandonner son emploi* et *démissionner de son emploi*.

Pour cette raison nous avons opté pour un traitement différencié des verbes et des noms:

- pour les verbes, nous utilisons une ressource lexicographique, à savoir un dictionnaire de synonymes, pour amorcer le calcul de proximité sémantique des prédications;
- en revanche, pour les noms, nous calculons une distance qui ne fait pas appel à la notion de synonymie, mais à une notion de proximité de distribution sélective par rapport à un ensemble de verbes synonymiques dans des constructions données.

Ainsi, notre approche se distingue aussi de la plupart des autres approches distributionnelles contextuelles en traitant de manière différenciée les éléments se trouvant en relation syntagmatique les uns avec les autres.

Une approche géométrique

Notre approche est résolument « continuiste », au sens où le modèle sous-jacent fait systématiquement appel aux mathématiques du continu:

- Le dictionnaire de synonymes permet de définir un espace sémantique des verbes dans lequel chaque verbe occupe une région plus ou moins étendue (selon son degré de polysémie). La distance entre points de cet espace reflète assez fidèlement les différences de sens entre les différents emplois de ces verbes.
- A un verbe dans une construction donnée on associe un espace de sélection distributionnelle dans lequel les distances entre les différents noms pouvant occuper une position actancielle donnée reflètent assez fidèlement les différences de sens qu'ils induisent pour ce verbe.

- La distance entre deux prédications sera calculée à partir de la distance dans l'espace sémantique des verbes et des distances distributionnelles des actants en correspondance dans les deux énoncés. Le fait de disposer d'un modèle utilisant les mathématiques du continu permet d'éviter les difficultés insurmontables des modèles du sens linguistique qui cherchent à organiser les sens dans des structures discrètes.

Pour illustrer ce point, considérons les énoncés suivants:

Il a passé ses microbes à toute sa famille

Il a filé une punition à son fils

Il a donné un coup de pied à son chien

Il a donné un concert à ses amis

Il a légué son caractère de cochon à sa fille

Il est clair que chacune de ces prédications s'éloigne quelque peu du sens de la prédication de transfert de propriété, même au sens large, que nous avons présentée plus haut. Faut-il quand même les regrouper avec ces dernières, en négligeant ces écarts ou bien faut-il au contraire les considérer comme des sens différents, bien que relativement proches? Dans ce dernier cas, combien de sens différents faut-il envisager? Et faut-il hiérarchiser ces sens en les traitant comme des nuances d'un sens général plus vague? Ces questions sont en fait indécidables, car les différentes réponses que l'on peut y apporter sont toutes aussi pénalisantes.

Nous ne cherchons donc pas à construire des classes sémantiques de prédication, mais à situer chaque prédication par rapport aux autres dans un espace sémantique global. Il est bien entendu possible et même probable que ces prédications s'organisent dans cet espace en nuages de points révélant des classes de prédications (presque) synonymes, qui pourraient être calculées par des méthodes automatiques (techniques de *clustering*). Mais nous ne faisons pas l'hypothèse *a priori* de l'existence de telles classes. Notamment, le modèle permet de rendre compte de l'existence de prédications intermédiaires reliant ces classes par des changements graduels de sens. Il permet aussi de rendre compte de nuances de sens sans pour autant multiplier les classes de sens (ou les sous-classes dans les approches hiérarchiques).

Notre approche se distingue donc des autres approches (distributionnelles ou autres) qui supposent une discrétisation, en un sens ou un autre, du sens linguistique.

Une approche opportuniste du contexte

Comme nous l'avons vu dans l'introduction, Jacques François distingue le cadre prédicatif, qui définit la prédication étroite, des satellites et opérateurs temporels ou modaux qui portent sur cette prédication. Pour classique qu'elle soit, cette distinction n'est pas sans poser de problème. En effet, la frontière entre les arguments, qui font partie de la prédication étroite, et les circonstants, qui sont des satellites, n'est pas facile à fixer. Les critères linguistiques sont nombreux (cf. Bonami 1999), mais ils ne sont ni toujours concordants, ni toujours discriminants. Pour ne donner qu'un exemple, prenons le critère de savoir si le complément répond à la question *quoi/qui/à qui/de qui?* ou à la question *où/quand/comment?* et appliquons-le aux exemples suivants:

Le chat saute sur la souris

Le chat saute sur la plage

Le chat saute sur le canapé

Le premier énoncé correspond à la question *sur quoi?* ce qui fait du groupe prépositionnel *sur la souris* un argument indubitable du verbe *sauter*. Dans le deuxième au contraire, c'est la question *où?*, ce qui fait de *sur la plage* un circonstant. Mais qu'en est-il du troisième énoncé? Il semble que les deux questions *sur quoi?* et *où?* conviennent tout autant, ce qui rend le critère inopérant.

En fait, nous pouvons contourner ces difficultés (qui sont décuplées quand il s'agit de traitement automatique), parce que cette distinction entre arguments et circonstants n'est pas vraiment pertinente

pour notre modèle. Si les arguments sont généralement déterminants pour le calcul du sens de la prédication, certains autres compléments, tout circonstanciels qu'ils soient, peuvent aussi jouer un rôle crucial. Ainsi peu nous importe que *à Paris* soit ou non un argument de *monter* dans *Il est monté à Paris*, ou que *comme un champion* soit ou non un circonstant de *jouer* dans *Il a joué comme un champion*: dans les deux cas, le complément est décisif pour déterminer le sens du verbe. De même, en comparant *Ce livre paraît dans deux jours* à *Ce livre paraît intéressant*, ou encore *Il reste pendant deux jours à Il reste fatigué*, on constate qu'un circonstanciel temporel peut, au même titre qu'un argument, déterminer la classe du cadre prédicatif tel que le définit Jacques François, pour qui, rappelons-le, les propriétés aspectuelles (transitionalité et dynamicité) comptent autant que les propriétés participatives.

Nous proposons donc de replacer tout complément rattaché au verbe sur une échelle continue allant du plus influent au moins influent dans la détermination du sens de la prédication. Cela implique d'avoir accès à un moyen de mesurer le « degré d'influence » d'un complément, ce qui n'est pas trivial. Nous faisons l'hypothèse, dans le cadre de notre approche par corpus, que la distribution d'un complément dans le corpus est corrélée à ce degré d'influence. Cela ne veut pas dire que plus un complément est fréquent, plus son degré d'influence est élevé. En revanche, ce sont les écarts relatifs de distribution qui sont significatifs.

Prenons par exemple les syntagmes prépositionnels introduits par la préposition *à* dont la tête nominale fait référence à une personne (*à quelqu'un*, *à ce monsieur*, *à un enfant*, *à M. Untel*, pronom clitique *lui*, etc.). Ces syntagmes seront beaucoup plus présents avec des verbes tels que *donner*, *prendre*, *parler* qu'avec des verbes tels que *travailler*, *écouter*, *manger*. Cette distribution non homogène peut donc nous permettre de discriminer deux ensembles de prédications. C'est cette notion de sélection distributionnelle qui est au cœur de notre modèle. Notons d'ailleurs que cette mesure d'une plus ou moins grande sélectivité d'un type de complément a souvent été utilisée pour distinguer arguments et circonstants (cf. par exemple Fabre & Frérot 2002). Mais comme nous l'avons dit, notre objectif est différent: seul nous intéresse le potentiel discriminant de tel type de complément pour tel verbe, quel que soit le statut de ce complément pour ce verbe. Notons aussi que les techniques que nous utilisons sont assez proches de celles développées dans l'approche de l'analyse sémantique latente (LSA, cf. Schütze 1998), à la différence non négligeable que cette approche, contrairement à la nôtre, ne tient pas du tout compte de l'aspect syntaxique des relations entre unités lexicales en s'en tenant à de simples relations de cooccurrence.

Ainsi notre approche peut être qualifiée d'opportuniste dans la mesure où nous allons chercher les éléments contextuels susceptibles de nous aider dans notre tâche de détermination du sens d'un verbe sans nous préoccuper de leur statut précis. Et si nous ne nous intéressons pour le moment qu'aux sujets et compléments à tête lexicale nominale, c'est essentiellement pour des raisons de faisabilité computationnelle. De fait, les autres éléments contextuels (marques de temps verbal, déterminants des compléments nominaux, adverbes, complétives, etc.) auraient vocation à être eux aussi pris en compte.

Le modèle

Nous allons donner ici les détails pratiques des différentes étapes de la construction de l'espace sémantique de prédication. Cet espace doit rendre compte des différents sens que peuvent prendre les unités étudiées, mais aussi de la topologie sémantique définie par ces différents sens. Rappelons que notre méthode nécessite deux opérations distinctes dont les résultats vont être ensuite combinés:

- La construction d'un espace sémantique des verbes dans lequel chacun occupe une région plus ou moins étendue (selon son degré de polysémie), à partir d'un dictionnaire des synonymes. La distance entre points de cet espace reflète assez fidèlement les différences de sens entre les différents emplois de ces verbes.
- La construction d'un espace de sélection distributionnelle associé à chaque verbe dans chacune de ses constructions. Deux noms seront représentés par deux points proches dans cet espace s'ils occupent la même position dans la construction et qu'ils contribuent à sélectionner des sens voisins de la prédication.

- Le calcul de la distance entre deux prédications à partir de la distance dans l'espace sémantique des verbes et des distances distributionnelles des noms en correspondance dans les deux énoncés.

L'espace sémantique des verbes

Ploux et Victorri (1998) ont mis au point une méthode de construction automatique des espaces sémantiques. Cette méthode utilise la relation de synonymie comme accès aux informations lexicosémantiques. L'étude approfondie des relations de synonymie permet en effet de mettre en évidence à la fois le fonctionnement des unités polysémiques prises individuellement, et leur place dans l'organisation globale du lexique.

La construction des espaces sémantiques repose sur un constat: un synonyme ne suffit pas en général pour définir un sens lexical. Prenons l'exemple du verbe *abandonner*, qui va nous servir à illustrer toute cette présentation. Dans sa synonymie avec *abandonner*, *laisser* est à la fois synonyme de *quitter* et de *confier*, qui correspondent à deux sens différents de *abandonner*. L'idée est donc de caractériser un sens par un ensemble de synonymes.

VisuSyn : *abandonner* (116 unités, 223 cliques) - composantes 1 et 2

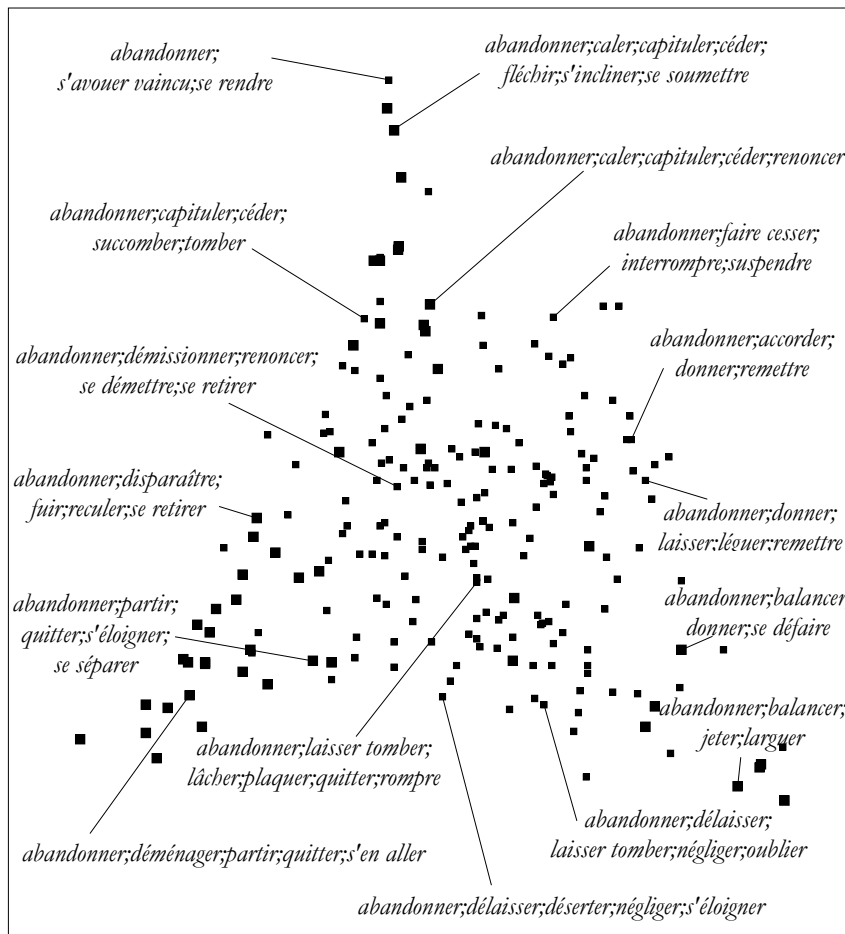


Figure 1: Représentation bidimensionnelle de l'espace sémantique des verbes restreint au verbe *abandonner*

Plus précisément, l'algorithme utilise les cliques du graphe de synonymie. Une clique est un ensemble le plus grand possible de mots deux à deux synonymes. L'idée sous-jacente à la construction des espaces sémantiques est qu'une clique correspond, en première approximation, à une nuance de sens

possible pour les mots considérés. Sans entrer dans les détails (voir pour cela Ploux & Victorri 1998 et Victorri & Venant 2007), disons simplement que les points de cet espace sont les cliques du graphe de synonymie, et qu'il est muni de la distance du χ^2 , bien connue en analyse des données. La figure 1 montre l'espace sémantique restreint à *abandonner* et ses synonymes, tel qu'il a été obtenu à partir du dictionnaire des synonymes du laboratoire de linguistique de l'Université de Caen (CRISCO 1998). Il s'agit d'une vision partielle puisque l'espace étant de grande dimension, il faut le projeter pour pouvoir en obtenir une représentation bidimensionnelle (nous avons fait figurer ici la projection selon les deux premiers axes d'une analyse en composantes principales).

On peut confronter cette représentation aux différents sens de *abandonner* relevés par les dictionnaires de langue. Par exemple, le *Trésor de la langue française informatisé* (TLFI 2001) énumère sept sens principaux (pour ne pas alourdir l'exposé, on ne prend pas en compte les emplois pronominaux):

1. Renoncer à un pouvoir, à des droits, à la possession d'un bien ou à l'utilisation d'une chose.
2. Quitter un lieu, ne plus l'occuper.
3. Cesser de défendre une cause, renoncer à des principes, à une idée en la rejetant ou simplement en s'en séparant
4. Renoncer à poursuivre une action, une recherche, renoncer à une entreprise, un projet
5. Quitter quelqu'un, s'en séparer, le laisser à lui-même
6. Laisser à quelqu'un la possession ou le soin d'un bien (ou d'une personne), laisser quelque chose à l'entière disposition de quelqu'un
7. Laisse quelque chose ou quelqu'un en proie à quelque chose (généralement une force hostile)

On pourra observer que ces sept sens sont bien présents dans la représentation géométrique, et que, de plus, les relations de proximité entre ces sens sont respectées. Par exemple, en se déplaçant du haut de la figure vers le bas, on passe progressivement du sens 4 (*s'avouer vaincu, se rendre, caler, céder*) au sens 5 (vers le bas: *jeter, larguer*), soit en passant par le sens 1 (au centre: *démissionner, se retirer*), soit sur la droite via la notion de renoncement (sens 3) et de dépossession (sens 6). Puis en se dirigeant vers la partie gauche, on glisse graduellement vers le sens 2 (*s'en aller, fuir*) en passant par le sens 7 (*laisser, délaissé, négliger*).

Ainsi, cette méthode permet effectivement d'obtenir automatiquement une représentation continue du sens des verbes qui répond bien aux attendus de notre cadre théorique.

Les espaces de sélection distributionnelle

Les différents sens de *abandonner* ne mettent en jeu que trois constructions syntaxiques:

GN₁ abandonne.

GN₁ abandonne GN₂.

GN₁ abandonne GN₂ à GN₃.

Les informations fournies par la construction syntaxique ne sont donc pas suffisantes pour la discrimination des sens au sein de l'espace sémantique. Il faut aussi de caractériser sémantiquement les arguments du verbe. Ainsi, très approximativement, si le GN₂ désigne un lieu, *abandonner* prend le sens 2, s'il est de type 'humain', on obtient les sens 5 ou 7, avec le type 'objet' ce sera les sens 1 ou 6, avec le type 'idée' le sens 3, avec le type 'action' le sens 4, etc. Notons d'ailleurs qu'il y a des cas intermédiaires: si le GN₂ est de type 'animal domestique', il est impossible de trancher entre la séparation affective (sens 5) et la dépossession (sens 1 et 6)... Notons aussi que la présence et la caractérisation du GN₃ sont nécessaires, notamment pour distinguer les sens 1 et 6, et les sens 5 et 7. Enfin en l'absence de complément, c'est en examinant le type de GN₁ (sportif, homme politique, etc.) que l'on peut espérer distinguer les sens 1, 3 et 4. Bien entendu ces indications restent très grossières: dans un certain nombre de cas, seul l'examen d'un contexte plus large permettrait de déterminer le sens précis du verbe. Mais nous faisons l'hypothèse que la présence massive de cas typiques dans une analyse quanti-

tative de corpus permettra de dessiner le paysage général des emplois de ce verbe avec suffisamment de netteté, malgré les erreurs inévitables sur un nombre restreint d'occurrences.

La technique que nous utilisons pour la caractérisation sémantique des arguments de la prédication s'inscrit dans le cadre bien connu de l'analyse distributionnelle « à la Harris ». Elle est exploitée depuis longtemps dans la communauté du traitement automatique des langues pour la construction de bases de connaissances ou de ressources terminologiques à partir de textes (cf., entre autres, Greffens-tette 1994, Habert & Nazarenko, 1996; Aussenac-Gilles, Biébow & Szulman 2000, Lin & Pantel 2001). Elle est entièrement automatique. Elle ne fait appel à aucune modélisation préalable de connaissances sémantiques sur le corpus et elle utilise les rapports de dépendance syntaxique élémentaires entre unités lexicales. Mais rappelons que contrairement à la plupart des travaux dans cette lignée nous ne cherchons pas à créer des classes de noms ayant le même sens ou faisant partie d'une même classe sémantique générale, mais des ensembles de noms qui influencent de la même façon le sens d'un verbe donné dans une construction donnée. Autrement dit si nous voulons regrouper des noms comme *pouvoir*, *emploi*, *mission*, *mandat*, *charge*, *fonction*..., ce n'est pas pour caractériser le sens de *pouvoir* ou de *charge* dans l'absolu, mais pour caractériser le sens de *abandonner* dans *abandonner son pouvoir* ou *abandonner sa charge*.

En quelques mots (cf. pour plus d'information Jacquet & Venant 2005 et Jacquet, Venant & Vic-torri 2005), notre méthode consiste à calculer une distance entre mots sur la base de contextes syn-taxiques partagés. Grâce à cette distance, on peut construire l'espace de sélection distributionnelle associé à un verbe et une construction. Le corpus est analysé automatiquement et ce sont les sorties d'un analyseur syntaxique robuste qui constituent les données de base pour la construction de l'espace distributionnel. Lors de nos précédentes expérimentations nous avons utilisé les analyseurs SYNTAX (Bourigault & Fabre, 2000) et XIP (Ait, Chanod & Roux 2002).

Les données sont séparées en deux types: d'une part les mots lexicaux du corpus, d'autre part les contextes lexico-syntaxiques de ces mots lexicaux. Un contexte lexico-syntaxique est une paire consti-tuée par un mot lexical et une relation syntaxique avec un mot lexical donné, par exemple 'tête nomi-nale de complément direct de *abandonner*', 'tête nominale de complément prépositionnel de *abandon-ner* introduit par la préposition *à*' ou encore 'verbe régissant un complément direct de tête nominale *charge*'. A chaque mot lexical et à chaque contexte lexico-syntaxique de ce mot sont associées leurs fréquences dans le corpus: cela constitue ce que nous appelons la fiche distributionnelle de ce mot.

L'espace distributionnel associé à un verbe dans l'une de ses constructions est composé des diffé-rents contextes lexico-syntaxiques du verbe dans cette construction trouvés dans le corpus. Cet espace est muni d'une distance: les noms qui se trouvent dans la même position dans la construction syn-taxique sont plus ou moins proches selon que leur fiche distributionnelle est plus ou moins similaire. Ainsi, dans l'espace distributionnel associé à la construction ' GN_1 abandonne GN_2 ', les noms *pouvoir*, *emploi*, *mission*, *charge*, etc. pourront être relativement proches les uns des autres parce que leurs fiches distributionnelles sont plus semblables entre elles que celles des noms *idée*, *opinion*, *espoir*, *jugement*,...

Ainsi les espaces distributionnels permettent de regrouper les noms qui jouent un même rôle dans la sélection du sens d'un verbe dans une construction donnée.

L'espace sémantique des prédications

L'espace sémantique des prédications est obtenu en combinant les données provenant de l'espace sé-mantique des verbes et celles provenant des espaces distributionnels. Chaque prédication peut en effet être localisée à la fois dans l'espace sémantique des verbes en déterminant le sens précis du verbe dans cet espace, et dans l'espace distributionnel pertinent en déterminant pour chaque nom intervenant dans la prédication sa position dans cet espace.

La distance sémantique entre deux prédications est alors calculée en combinant la distance entre les sens des deux verbes dans l'espace sémantique des verbes et des distances entre noms dans les espaces distributionnels concernés.

Prenons l'exemple de la prédication *Le trésorier a abandonné sa charge*. En analysant les synonymes de *abandonner* qui acceptent des contextes lexico-syntaxiques voisins, on localise le sens de *abandonner* dans cette prédication au voisinage de cliques comme *abandonner*, *démissionner*, *renoncer*, *se démettre*, *se retirer*. Dans l'espace distributionnel associé à la construction 'GN₁ abandonne GN₂', on place d'une part *trésorier* dans le voisinage de *président*, *comptable*, etc. et d'autre part *charge* au voisinage de *pouvoir*, *mission*, *emploi*, etc. Ces opérations s'effectuent automatiquement en utilisant les données sur les fiches distributionnelles des mots lexicaux impliqués. Soit alors une autre prédication, par exemple *Le directeur va démissionner de son poste*. On effectue les mêmes opérations: détermination du sens de *démissionner* dans l'espace sémantique des verbes et localisation de *directeur* et de *poste* dans l'espace distributionnel associé à la construction 'GN₁ démissionne de GN₂'. On peut alors déterminer la distance entre ces deux prédications. D'une part la distance entre les sens de *abandonner* et de *démissionner* va être faible dans l'espace sémantique des verbes. D'autre part les distances entre *trésorier* et *directeur* et entre *charge* et *poste* vont être elles aussi faibles dans les deux espaces distributionnels concernés (celui de la construction 'GN₁ abandonne GN₂' et celui de la construction 'GN₁ démissionne de GN₂'). On en déduit donc que ces deux prédications ont des sens très proches.

Ainsi, cette méthode permet, comme on le souhaitait, de comparer des prédications en tenant compte à la fois de la proximité sémantique des verbes et de celles des têtes nominales des groupes régis par les verbes. Et le fait d'utiliser une distance pour effectuer cette comparaison permet de rendre pleinement compte du caractère continu des différences de sens. Pour illustrer ce dernier point prenons l'énoncé suivant:

Jean a dû abandonner son chien à la SPA [Société Protectrice des Animaux]

Cette prédication sera à la fois assez proche des prédications du type 'transfert de propriété', que nous avons évoquées au début de cette article, via la synonymie de *abandonner* avec *donner* et *léguer*, mais aussi de prédications de type 'séparation affective', via la synonymie avec *quitter*, *larguer*, *rompre*. On peut donc penser qu'elle sera à égale distance des cas prototypiques de ces deux types de prédication, sans que l'on ait à choisir entre l'un ou l'autre. Il faut d'ailleurs noter que ce cas est intermédiaire à deux niveaux: non seulement le sens du verbe est intermédiaire entre le sens de transfert de propriété et celui de séparation affective, mais le statut syntaxique du complément prépositionnel *à la SPA* est aussi intermédiaire: statut d'argument du receveur dans le scénario d'un transfert de propriété, et statut de circonstant (lieu de la séparation) dans le scénario d'une séparation.

Conclusion

Le programme de recherche que nous venons de présenter est encore à un stade exploratoire. Nous avons conscience que l'intérêt de cette approche ne pourra être validé que quand le système sera entièrement implémenté et opérationnel. Les premières expérimentations que nous avons réalisées nous ont montré que nous aurons à surmonter un certain nombre de difficultés inhérentes à ce genre de traitement à large couverture: notamment, il nous faut développer des procédures spécifiques pour les expressions figées, pour les collocations, pour les verbes support, etc.

Les perspectives ouvertes par ce programme de recherche, s'il aboutit à des résultats convaincants, sont nombreuses. En premier lieu, la construction d'un espace sémantique de prédications pourrait rendre de grands services dans des tâches classiques de traitement automatique des langues telles que les systèmes d'extraction d'information ou les systèmes de question-réponse. Il est en effet essentiel dans ces systèmes de pouvoir reconnaître les différentes formulations possibles d'un même événement ou d'une même propriété. On peut aussi penser à d'autres applications, notamment dans le domaine didactique. Ainsi on pourrait fournir une ressource permettant à des apprenants s'exprimant maladroitement en français de trouver la formulation communément employée par un locuteur natif pour exprimer ce qu'ils veulent dire. Le système proposerait par exemple à quelqu'un cherchant à dire quelque chose comme *monter un sommet* les formulations *gravir une montagne* ou *atteindre un sommet*.

Mais avant tout, l'intérêt de ce programme est, à nos yeux, d'ordre théorique. Comme on l'a vu avec l'exemple de *abandonner*, il s'agit de valider une approche continue à la fois du sens des unités lexicales et des relations syntaxiques entre ces unités, seul cadre théorique capable à nos yeux de rendre compte de toute la richesse du phénomène de la prédication, au cœur de l'interface entre syntaxe et sémantique, un thème cher à Jacques François.

Bibliographie

- Aït S., Chanod J.P., Roux C. (2002), Robustness beyond shallowness: incremental dependency parsing, *Natural Language Engineering*, 8(2/3), pp. 121-144.
- Andrew G., Grenager T., Manning C. (2004), Verb Sense and Subcategorization: Using Joint Inference to Improve Performance on Complementary Tasks. *Conference on Empirical Methods on Natural Language Processing, EMNLP 2004*, pp. 150-157.
- Aussenac-Gilles N., Biébow B., Szulman N. (2000), Revisiting Ontology Design: a method based on corpus analysis, *Actes de 12th International Conference on Knowledge Engineering and Knowledge Management*.
- Bonami O. (1999), *Les constructions du verbe: le cas des groupes prépositionnels argumentaux*, Thèse de doctorat, Université Paris 7.
- Bourigault D., Fabre C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, pp. 131-151.
- Brent M.R. (1993), From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax, *Computational Linguistics*, 19, pp. 203-222.
- Briscoe T., Carroll J. (1997), Automatic Extraction of Subcategorization from Corpora *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 356-363.
- Cori M. & David S. (2008): Les corpus fondent-ils une nouvelle linguistique?, *Langages*, 171, pp. 111-129.
- CRISCO (1998), *Dictionnaire électronique des synonymes*, Laboratoire CRISCO, Université de Caen, <http://www.crisco.unicaen.fr/>
- Dik S. (1978), *Functional Grammar*, Dordrecht, Foris.
- Dik S. (1997), *The theory of Functional Grammar*, Berlin, Mouton De Gruyter.
- Fabre C., Frérot C. (2002). Groupes prépositionnels arguments ou circonstants: vers un repérage automatique en corpus, *Actes du Colloque Traitement automatique des langues naturelles, TALN-2002*.
- François J. (2003), *La prédication verbale et les cadres prédicatifs*, Bibliothèque de l'Information Grammaticale 54, Louvain, Peeters.
- Grefentstette (1994), *Explorations in Automatic Thesaurus Discovery*, London, Kluwer Academic Publishers.
- Habert B., Nazarenko A. (1996), La syntaxe comme marchepied de l'acquisition des connaissances: bilan critique d'une expérience, *Actes des Journées sur l'acquisition des connaissances*, Association Française d'Intelligence Artificielle, pp. 137-142.
- Jacquet G., Manguin J.-L.; Venant F., Victorri B. (2010), Construction dynamique du sens: application à la prédication verbale, *Actes des Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels, Rochebrune-2010*.
- Jacquet G., Venant F. (2005), Construction automatique de classes de sélection distributionnelle, *Actes du Colloque Traitement automatique des langues naturelles, TALN-2005*.
- Jacquet G., Venant F., Victorri B. (2005), Polysémie lexicale, in P. Enjalbert (éd.), *Sémantique et traitement automatique des langues*, Hermès, pp. 99-132.
- Lin D., Pantel P. (2001), Induction of Semantic Classes from Natural Language Text, *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*.
- Messiant C., Gábor K., Poibeau T. (2010). Acquisition de connaissances lexicales à partir de corpus: la sous-catégorisation verbale en français, *Traitement automatique des langues*, 52:1, à paraître.
- Ploux S., Victorri B. (1998), Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39:1, pp. 161-182.
- Rey A., Rey-Debove J. (1996), *Le Petit Robert, Dictionnaire de la langue française*, édition sur CD-Rom, Paris, LIRIS interactive.
- Schütze, H. (1998). Automatic word sense discrimination, *Computational Linguistic*. 24:1, pp. 7-124.
- TLFI (2001), *Trésor de la Langue Française informatisé*, laboratoire ATILF, CNRS-Université Nancy2, <http://atilf.atilf.fr/>.
- Van Valin R.D., La Polla R.J. (1997), *Syntax: Structure, Meaning, Function*, Cambridge, Cambridge University Press.
- Victorri B., Venant F. (2007), Représentation géométrique de la synonymie, *Le Français Moderne*, 2007:1, pp. 81-96.