

# Dégrouper les sens : pourquoi, comment ?

Benoît Habert, Gabriel Illouz, Helka Folch

LIR – LIMSI CNRS, BP 133 – 91403 Orsay Cedex – France  
{habert,gabrieli,folch}@limsi.fr

## Abstract

In order to be able to characterize social points of view in a given domain, a first step consists in spotting words which have several meanings (homonymy, polysemy) or which contexts of use differ widely and in identifying the corresponding shades of meaning. The opposition of use among parts of a corpus is detailed.

## Résumé

En lien avec l'objectif global de repérer les points de vue présents dans un domaine, il s'agit de détecter les mots qui ont plusieurs sens ou qui sont employés de manière différente selon les parties d'un corpus et de caractériser leurs emplois. L'application à un corpus partitionné est détaillée.

**Mots-clés :** sémantique distributionnelle, polysémie, homonymie, repérage de points de vue.

## 1. Stabilisation du sens et « mondes sociaux »

L'objectif d'un Web sémantique<sup>1</sup>, proposé par le W3C, le consortium qui gère le Web, est de modifier la division du travail entre l'homme et la machine dans l'accès aux ressources de la Toile par le sens<sup>2</sup>. Il s'agit d'adjoindre aux contenus informels actuels du Web des connaissances formalisées qui puissent être utilisées par des traitements automatiques et qui permettent ainsi, par exemple, de diminuer le temps passé par les utilisateurs à trier dans les résultats profus voire confus des moteurs de recherches. La proposition du W3C empile les couches (Laublet *et al.*, 2002) : i) expression standardisée des méta-données (Floch et Habert, 2004) ; ii) représentation ontologique convergente ; iii) raisonnement. La première couche entend fournir une syntaxe unifiée pour décrire les ressources et les jugements portés sur elles. La seconde couche a pour objectif un accord sur « ce qui existe », les « étants » manipulés : il faut en effet « parler de la même chose ». C'est le niveau ontologique. La troisième couche englobe les traitements, inférentiels en particulier, possibles à partir du moment où l'on dispose à la fois d'un accord sur l'ontologie à manipuler et d'un méta-langage commun.

Une partie des ressources du Web relève sans doute d'une démarche formalisante poussée, contrôlée par une ontologie partagée, rendant possibles inférences et remodelages et limitant l'intervention humaine. C'est le cas de savoirs techniques et scientifiques stabilisés. Une autre partie des ressources du Web fait coexister des points de vue différents<sup>3</sup> ressortissant à des re-

---

Ce travail a bénéficié de discussions avec Eric Gaussier (XRCE), André Salem et Serge Fleury (Syled – Université Paris III), Didier Bourigault, Anne Condamines, Cécile Fabre, Josette Rebeyrolle (ERSS – Université de Toulouse-le-Mirail), Elie Naulleau (Semiosys), Claude Henry (LIMSI) et Pierre Zweigenbaum (STIM – AP-HP).

<sup>1</sup> <http://www.w3.org/2001/sw/>

<sup>2</sup> Voir aussi <http://www.semanticweb.org/>

<sup>3</sup> Le langage de méta-données Topic Maps vise précisément l'articulation de tels points de vue (Floch et Habert, 2002).

présentations semi-formelles où la qualification humaine des informations, l'interprétation, est centrale<sup>4</sup>. La multiplication des forums et des discussions électroniques rend désormais urgent le développement de techniques adaptées à ces fonctionnements sémantiques particuliers<sup>5</sup>. En effet, la désorientation menace aisément face à la multiplication des prises de position dans un domaine donné. La masse même des données fait obstacle à la perception claire des stabilités et des mouvances. Les comptes rendus de réunions publiques et les documents rassemblés par le Centre National du Débat Public<sup>6</sup> sur des thèmes comme la liaison à très haute tension entre l'Espagne et la France<sup>7</sup> sont prototypes de ces débats citoyens soit directement électroniques soit médiatisés par la mise en ligne.

## 2. Dégrouper les sens

Un préalable à la mise au jour des points de vue présents dans un ou plusieurs domaine(s) est le dégroupement automatique de sens : le repérage des mots employés simultanément avec des sens divergents au sein du corpus construit pour ce(s) domaine(s). Ces mots recouvrent des réalités sémantiques (et sociales) très différentes. Il peut s'agir de simples homonymes : *grève* 'arrêt du travail'/'plage de gravier', par exemple. Les contextes d'emploi des homonymes sont souvent différenciés. Les différents sens d'un mot polysémique ont plus de chances d'apparaître dans des contextes proches : il y a souvent continuité d'un sens à l'autre, comme pour *guerre*, de *guerre aérienne* à *guerre médiatique*. C'est le cas en particulier des polysémies régulières, de l'engendrement réglé de nouveaux sens : les mots désignant un instrument de musique peuvent ainsi également renvoyer à la personne qui en joue. Un troisième cas de figure — particulièrement pour le discours social — est celui des mots qui incarnent une divergence plus ou moins grande ou un certain « vague », une indétermination du sens (Hanks, 2000). C'est le cas de *service minimum* pour les transports publics.

Au sein de la sémantique « machinale », le dégroupement automatique<sup>8</sup> a cependant, dans l'immédiat, relativement peu retenu l'attention. Les travaux se sont centrés pour l'essentiel sur la désambiguïsation sémantique (*Word Sense Disambiguation*), c'est-à-dire l'attribution en contexte à un mot du sens pertinent en fonction d'un répertoire de sens prédéterminé<sup>9</sup>. En acquisition sémantique, ce sont la mise en évidence de la sous-catégorisation verbale et les regroupements de mots (la recherche de similarités sémantique et leur organisation par classification hiérarchique par exemple) qui ont surtout été explorés (Manning et Schütze, 1999 : ch. 8). Les obstacles sont partiellement techniques : dégroupement des sens, c'est trouver le moyen de repérer les cas où « un mot en cache un voire plusieurs autre(s) », alors que pour les outils de traitements, il s'agit toujours de la même chaîne de caractères. Les obstacles sont également théoriques. La vision « discrétisante » (les sens sont disjoints) et fixiste du sens domine<sup>10</sup>. Paradoxalement, ce sont plutôt des recherches en bibliométrie qui se sont attachées à la comparaison des dénomi-

<sup>4</sup> Imaginer des Webs sémantiques, aux fonctionnements distincts, c'est généraliser la conception hétérogène du sens défendue dans Kleiber (1999 : 45-51) : la construction du sens articule plusieurs dimensions qui relèvent de sémantiques partiellement disjointes (instructionnelles, référentielles, etc.).

<sup>5</sup> C'est l'objectif du projet exploratoire RNRT Outiller les alliances ([http://www.telecom.gouv.fr/rnrt/index\\_net.htm](http://www.telecom.gouv.fr/rnrt/index_net.htm)) auquel nous avons participé de 2000 à 2003 : fournir des outils pour l'amélioration des débats citoyens organisés par la Fondation pour le Progrès de l'Homme (<http://www.fph.ch>). C'est aussi celui du logiciel Prospéro (Chateauraynaud, 2003).

<sup>6</sup> [http://www.debatpublic.fr/cndp/debat\\_public\\_cpdp.html](http://www.debatpublic.fr/cndp/debat_public_cpdp.html)

<sup>7</sup> <http://debat-liaison-tht-france-espagne.com>

<sup>8</sup> La tâche, dénommée *sense induction* dans Yarowsky (1995), diffère selon qu'on dispose ou non d'une partition préexistante.

<sup>9</sup> Pour un état de l'art, cf. Ide et Véronis (1998) et (Manning et Schütze, 1999 : ch. 6).

<sup>10</sup> *A contrario*, cf. Fuchs et Victorri (1994).

nations par discipline (Losee, 1996) ou aux transferts de dénominations d'un champ à un autre (Losee, 1995) : le mot change de sens en passant ainsi d'un domaine à un autre. Certains travaux en terminologie automatisée ont été consacrés à des phénomènes similaires (Ibekwe-Sanjuan, 1998) : l'évolution d'un groupes de termes au fil du temps par exemple. Le routage d'information, c'est-à-dire l'envoi automatique à l'utilisateur des documents correspondant aux centres d'intérêt qu'il a formulés, rencontre le problème inverse : les mots correspondant aux centres d'intérêt sélectionnés ne sont plus forcément les mêmes.

Malgré les expériences que nous avons déjà menées, notre contribution reste programmatique. Elle vise à présenter de manière raisonnée les différentes facettes du dégroupement de sens : identifier les mots mouvants selon une partition (section 4.1.) ou hors partition (section 4.2.) ; caractériser les directions qui organisent les emplois d'un mot identifié comme mouvant selon une partition (section 4.3.) ou hors partition (section 4.4.). La section 3, liminaire, a pour objectif un ancrage concret et la mise en évidence, par l'exemple, des bénéfices attendus de la direction de travail globale. Pour finir (section 5.1.), nous examinons certaines précautions et certains paramètres à prendre en compte dans la perspective choisie. La section 5.2. conclut la contribution par la question, cruciale mais difficile, des démarches d'évaluation possibles pour les techniques de dégroupement de sens.

### 3. Mots « mouvants » dans un corpus de textes syndicaux

Les écarts de sens peuvent se manifester par une variation des contextes où figure un mot d'une partie à l'autre d'un corpus. Dans Habert *et al.* (1999)<sup>11</sup>, c'est par le biais des fluctuations ou au contraire des stabilités de ces associations <mot, contexte> que nous essayons de progresser vers le repérage automatique des mots qui font consensus relatif et de ceux qui au contraire témoignent de divergences. C'est donc une *sémantique distributionnelle*. Les associations retenues sont les rapports de dépendance élémentaire entre un « gouverneur » et les mots « pleins » qu'il régit (modificateurs ou arguments) au sein de constituants syntaxiques fournis par des analyseurs syntaxiques robustes.

Le corpus utilisé est celui des résolutions générales (RG) des congrès de la CFTC de 1945 à 1964, de ceux de la CFDT et de la CFTC « maintenue » de 1964 à 1992. La partition utilisée pour contraster les distributions n'est pas basée sur les 3 émetteurs repérables : CFTC originelle, CFDT et CFTC maintenue. Elle repose sur l'interprétation des résultats d'une analyse factorielle des correspondances (Habert et Tournier, 1987) : pour la CFDT, ce sont les événements de 1968 plus que la scission qui entraînent un important changement lexical, suivi d'un autre tournant lexical en 1979. Les parties retenues (correspondant aux agglomérats des deux premiers axes de l'AFC), de taille proche, sont les suivantes : 1) les RG de la CFTC jusqu'à la scission de 1964 incluse et de la CFDT d'avant mai 1968 [TC45-64\_DT65-67] (29 704 occurrences) ; 2) les RG de la CFDT maintenue [TC65-90] (31 673 o.) ; 3) les RG de la CFDT « radicalisée » [DT70-76] (25 596 o.) ; 4) les RG de la CFDT « recentrée » [DT79-92] (34 677 o.).

Au sein d'une partie, on peut déterminer pour chaque mot, ses voisins les plus proches. On compare chaque mot aux autres et on détermine pour chaque paire, le nombre de contextes partagés, le nombre de contextes propres à l'un des mots, le nombre de contextes propres à l'autre. Un indice utilise ces quantités et fournit une distance. Les contextes sont les dépendances élémentaires extraites au sein des groupes nominaux fournis par le logiciel d'acquisition terminologique Lexter (Bourigault, 1994). Nous utilisons l'indice de Jaccard<sup>12</sup>. Les voisinages

<sup>11</sup> La revue *Sémiotiques* ayant pris du retard dans ses livraisons, l'article est daté de 1999 alors qu'il est paru en 2002.

<sup>12</sup> D'ailleurs peu approprié puisqu'il « éloigne » les mots qui rentrent dans des nombres de contextes très

observés peuvent différer fortement d'une partie à l'autre. Nous en fournissons deux exemples, où les 5 plus proches voisins d'un mot sont classés par proximité décroissante avec ce mot au sein de la partie :

Pivot	TC45-64_DT65-66	DT70-76	DT79-92	TC65-90
<i>action</i>	<i>lutte, organisation, représentant, participation, syndicalisme</i>	<i>lutte, organisation, stratégie, mouvement, pratique</i>	<i>lutte, intervention, mobilisation, négociation, revendication</i>	<i>mesure, plan, orientation, programme, objectif</i>
<i>travailleur</i>	<i>salarié, pays, peuple, classe ouvrière, organisation</i>	<i>classe ouvrière, masse, peuple, classe, forces</i>	<i>salarié, masse, ensemble, population, syndicat</i>	<i>salarié, personne, entreprise, homme, famille</i>

Les décalages sont très sensibles pour *travailleur* : le voisinage pour TC65-90 est indéniablement d'inspiration chrétienne (*personne, homme, famille*). Le partage de *classe ouvrière* dans les deux premières parties, son absence dans la troisième sont également significatifs. Les voisins sont plus « neutres » pour DT79-92 par rapport aux parties précédentes. Le partage de *lutte* comme voisin d'*action* par les 3 premières parties, mais l'irruption de *négociation* dans la troisième manifestent également des évolutions sémantiques.

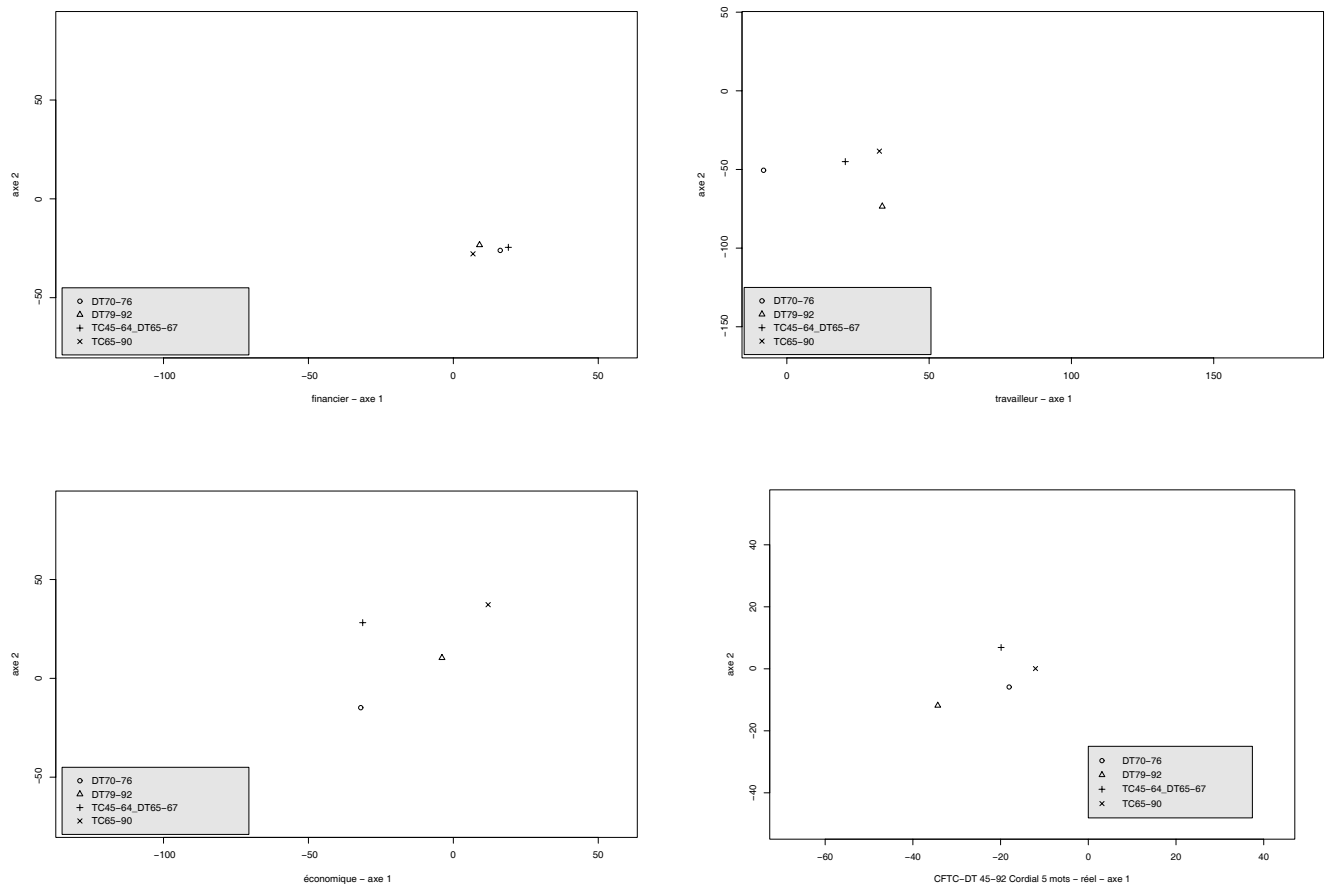
Examiner les voisins d'un mot ne permet pas de détecter les mots dont les contextes changent significativement. Pour avancer dans ce sens, nous retenons, au sein des 100 noms et des 100 adjectifs les plus fréquents dans le corpus, les 26 noms et les 32 adjectifs effectivement partagés par les 4 parties. Dans l'espace à  $n$  dimensions<sup>13</sup> de l'union des contextes employés par les  $k$  mots partagés considérés, on remplace chaque mot partagé par autant d'hétérographes artificiels que de parties. *Travailleur* est ainsi remplacé par *travailleurDT70-76*, *travailleurTC65-90*... On peut alors examiner les convergences/divergences de contextes selon les parties pour un mot partagé. On utilise une méthode d'analyse multidimensionnelle (Sammon, 1969) pour projeter dans un plan les 4 points correspondant à un même mot. Ces quatre points peuvent être assez ou très rapprochés dans le plan : on peut penser que l'emploi du mot est stable d'une partie à l'autre. Dans le cas contraire, il peut s'agir d'un mot « chahuté ». La figure 1 manifeste une telle opposition entre deux mots pourtant proches sémantiquement « en langue » : *financier* et *économique* (à gauche, de haut en bas). Les emplois d'un mot peuvent faire « bande à part » dans une partie, isolée, alors que les autres sont regroupées. C'est le cas figure 1 de *travailleur* (en haut à droite), où les emplois correspondant à l'époque « radicale » de la CFDT (1970-1967) se distinguent, et de *réel* (en bas à droite) où c'est cette fois la CFDT « recentrée » qui se démarque.

#### 4. Facettes du dégroupement de sens

Les expériences de la section 3 avaient pour objectif de « faire sentir » certaines des dimensions du dégroupement de sens. Nous allons nous appuyer sur elles pour une présentation plus large, qui oppose le repérage des mots mouvants (sections 4.1. et 4.2.) à la caractérisation des (pôles de) sens sous-jacents (sections 4.3. et 4.4.), selon que l'on s'appuie ou non sur une partition du corpus.

différents, même si l'un des mots n'emploie que des contextes utilisés également par l'autre.

<sup>13</sup> 1 400 pour les 26 noms et 763 pour les 24 adjectifs.



#### 4.1. Identifier les mots mouvants selon une partition

Des « visualisations » intuitives comme celles de la section 3 permettent un premier repérage des stabilités et mouvances. On peut chercher à leur substituer/associer un indice de dispersion facilitant l'examen et le classement. C'est la démarche de Aussenac-Gilles *et al.* (2003). Les données de l'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) permettent, sur un corpus en ingénierie des connaissances opposant deux périodes, d'isoler, via des indices appropriés, les termes dont le comportement comme tête/dépendant de groupe syntaxique (nominal ou verbal) a fortement changé d'une période à l'autre. Ces termes sont regroupés manuellement en catégories pour mieux caractériser l'évolution du domaine.

#### 4.2. Identifier les mots mouvants hors partition

Dans le cadre d'une représentation vectorielle, le vecteur correspondant à un mot ayant plusieurs sens est l'« assemblage » des vecteurs représentant les sens sous-jacents. Avant de prétendre extraire ces vecteurs sous-jacents (section 4.4.), il faut repérer les vecteurs-assemblages. On peut faire plusieurs hypothèses. Dans une perspective proche de celle de Salton *et al.* (1997), il est possible que les mots à sens multiples contribuent à rapprocher les unités textuelles utilisées (documents, phrases)<sup>14</sup>. Il s'agit alors de détecter les mots qui, lorsqu'on les enlève, laissent place à un « nuage » d'unités textuelles plus dilaté (en recherche d'information, ce sont au contraire les mots qu'on souhaite éliminer). Une autre hypothèse est qu'un mot à sens multiple

<sup>14</sup> En analyse factorielle des correspondances, ce serait un sous-ensemble des formes qui contribuent **le moins** à créer les oppositions majeures.

aurait des voisins moins proches entre eux qu'un mot plus univoque. C'est le développement de l'intuition exemplifiée pour *travail* et *action* en section 3. Il s'agit alors de développer des métriques de dispersion du nuage des voisins, en tenant compte de la répartition inégale du nombre de voisins selon le mot examiné.

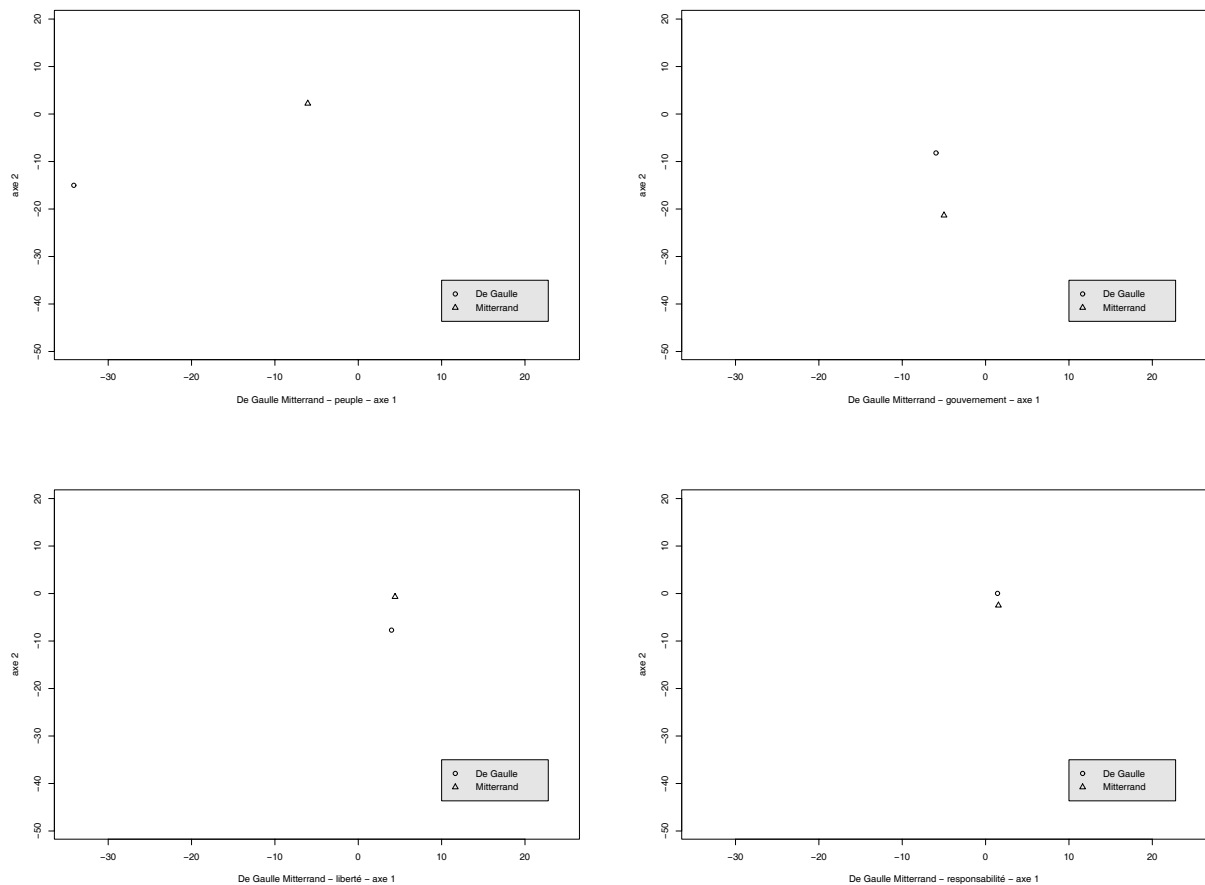
#### 4.3. Contraster les emplois selon une partition

Le dégroupement de sens est dans ce cadre guidé (*supervisé*) par une partition préexistante qui fournit le cadre d'observation et de détection des contrastes distributionnels.

Cette partition peut reposer sur une signalétique externe : datation, émetteur, etc. C'est la démarche suivie par Folch (2002). Dans le cadre du projet *Scriptorium*, consacré au sein de la Direction des Etudes et Recherches d'EDF à l'étude des débats autour de thématiques internes à l'entreprise, un corpus a été constitué autour de la notion de *service public* et de sa rédefinition avec l'ouverture à l'Europe. L'opposition entre émetteurs (les différents syndicats, la direction) fonde la partition au sein de laquelle sont observées, par les méthodes exposées en section 3, les convergences et divergences d'emploi des mots. Nous avons appliqué également ces méthodes à la mise en regard des interventions radio-télévisées de De Gaulle et de Mitterrand<sup>15</sup>. On observe ainsi sur la figure 2 un grand écart entre De Gaulle et Mitterrand dans l'emploi de *peuple* (en haut à gauche), ce qui ne surprend guère, et une grande proximité dans celui de *gouvernement* (en haut à droite), ce qui n'étonne pas non plus, tandis que le rapprochement sur *liberté* (en bas à gauche) et *responsabilité* (en bas à droite) est plus inattendu. Nous avons également examiné quelques pôles (cf. figure 3) dans les discours de Mitterrand en opposant la cohabitation avec ce qui précède (en reprenant le partage proposé par D. Labbé). La proximité sur *Europe* (en haut à gauche) s'oppose à l'éloignement pour *France* (en haut à droite). Il en va de même pour *gouvernement* (en bas à gauche) et *ministre* (en bas à droite). On retrouve en filigrane l'opposition entre le premier ministre et le président de la France, de la cohabitation, alors que ni l'Europe ni le gouvernement ne sont présentés de manière sensiblement différente.

---

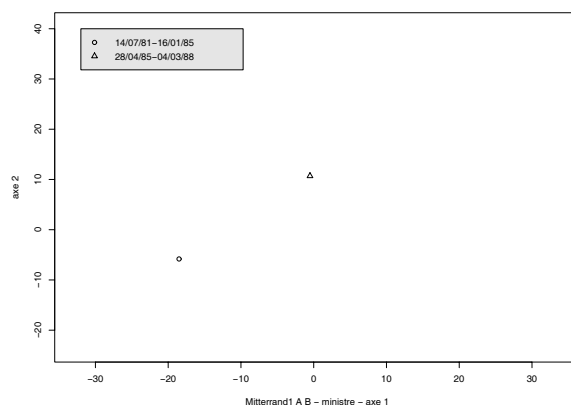
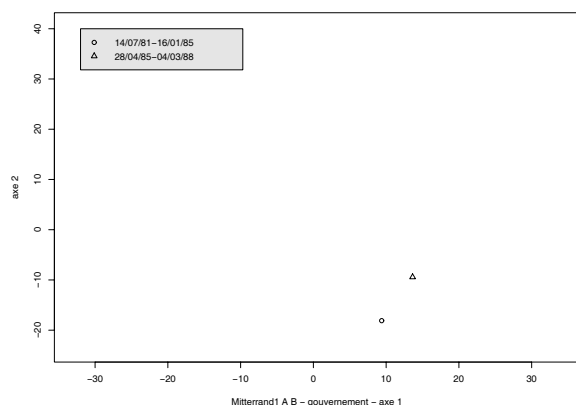
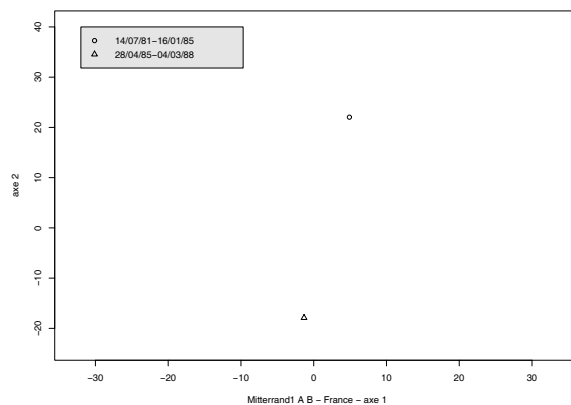
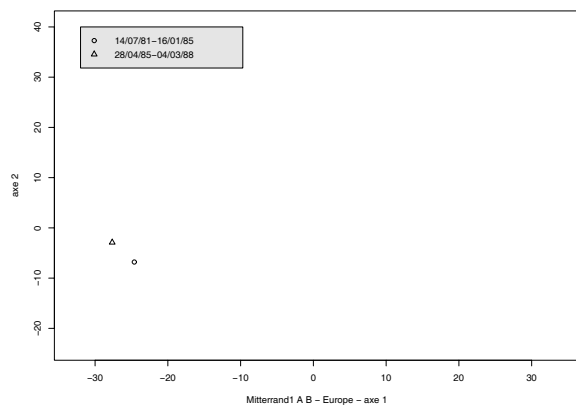
<sup>15</sup> Le corpus nous a été fourni par D. Labbé, que nous remercions.



La partition peut résulter d'outils de partitionnement (*clustering*) qui regroupent les unités textuelles choisies sur la base des traits jugés pertinents (formes brutes, formes racinisées – *stemming*, lemmes, lemmes moins mots-outils, etc.). Le corpus peut imposer cette démarche. Les archives d'un forum électronique peuvent ainsi permettre une répartition par émetteur (auteur) qui s'avèrera en réalité inutilisable si l'« émiettement » en une multitude d'auteurs rend imperceptibles les fluctuations des associations. C'est une démarche de regroupement préalable de cet ordre qui a été suivie dans l'expérience rapportée en section 3. Le partitionnement par interprétation des résultats d'une analyse factorielle préalable permettait en particulier de « lisser » les décalages de taille importants entre les documents utilisés (de 357 à 14 100 o.). Il aboutissait également à un nombre « raisonnable » de parties, où les contrastes et ressemblances étaient plus aisément saisissables. Dans Sébillot (2002), les 9 500 paragraphes d'un corpus de 200 articles du Monde diplomatique (des années 1987-97) sont représentés par les 165 noms les plus fréquents. Une classification hiérarchique de ces 165 noms débouche sur 80 classes thématiques, réduites aux 27 jugées pertinentes (et « nommées ») par l'accord de 4 évaluateurs sur les 5 mis à contribution. Si deux mots d'un thème sont présents dans un paragraphe, le paragraphe est placé dans ce sous-corpus thématique. Un paragraphe peut apparaître dans plusieurs thèmes<sup>16</sup>. Au sein de chaque thème, les noms les plus fréquents sont classés. Les traits sont les noms, verbes et adjectifs dans une fenêtre de plus ou moins 5 mots. Il est possible alors d'examiner pour un mot donné ses voisins dans deux thèmes, les partages et les divergences. Lorsqu'on considère les thèmes Négociations et Territoire, les voisins *état*, *parti*, *économique*, *nouveau*, *place*, *arrivée*, *politique* sont partagés, tandis que *accession*, *an*, *armée*, *concentra-*

<sup>16</sup> Les 27 classes thématiques ne constituent donc pas une partition de l'ensemble des paragraphes.

*tion, pays, coalition, contrôle, gouvernement, partage, achat, central, public* sont propres au premier thème et *local, soviétique, année, exécutif, prise, public, président, central* au second. Une visualisation des rapprochements/divergences serait probablement plus confuse avec au moins théoriquement jusqu'à 27 points pour un mot donné.



Au total, recourir à une partition « donne à voir » des oppositions. En revanche, cela tord éventuellement les contrastes, puisqu'on ne peut plus percevoir les décalages internes à une partie ni d'autres oppositions et rapprochements qui naîtraient d'autres partitions possibles (par exemple pour le corpus syndical de la section 3 entre CFTC maintenue, CFDT et CFTC jusqu'en 1964). C'est pourquoi il est sans doute préférable dans tous les cas de « rebattre les cartes », en testant plusieurs partitions, mais aussi en rassemblant hors partition (section 4.4.) les contextes d'un mot détecté comme « mouvant » et en les soumettant, de manière éventuellement couplée (Lebart *et al.*, 1997 : 185-206), aux techniques de classification et d'analyse de correspondances.

#### 4.4. Contraster les emplois hors partition

Rapp (2003) part de l'hypothèse que ce qui est appelé un mot ambigu est l'« assemblage » des vecteurs représentant les sens sous-jacents : le vecteur normalisé correspondant à *bank* devrait être plus similaire à celui résultant de la somme des vecteurs normalisés pour *money* et *river* qu'à la somme des vecteurs normalisés de toute autre paire de mots. Rapp cherche les vecteurs sous-jacents aux 12 mots dont la désambiguïsation est examinée dans Yarowsky (1995) (*axes* : *grid/tools*; *bass* : *fish/music*; *crane* : *bird/machine*; *drug* : *medicine/narcotic*; *duty* : *tax/obligation*; *motion* : *legal/physical*; *palm* : *tree/hand*; *plant* : *living/factory*; *poach* : *steal/boil*; *sake* : *benefit/drink*; *space* : *volume/outer*; *tank* : *vehicle/container*). Il se can-

tonne à deux sens possibles seulement par mot ambigu<sup>17</sup>. Une fenêtre de  $\pm 1$  mot plein, sur les 100 millions des mots du *British National Corpus* (BNC), fournit les contextes sous-jacents aux vecteurs. L'algorithme cherche les 10 mots les plus fortement associés à un mot ambigu (les contextes « majeurs »). Il produit les 90 paires possibles issues de ces 10 mots. Il calcule enfin la similarité entre ces paires et le mot ambigu. La plus ou moins grande similarité entre les vecteurs est mesurée par la distance de Manhattan. Pour *bank*, les 10 premières associations, par similarité décroissante, sont *account/river*, *accounts/river*, *accounts/manager*, *account/manager*, *account/accounts*, *loans/river*, *accounts/holiday*, *account/loans*, *account/holiday*, *central/account*. Malgré l'absence de lemmatisation, qui produit des néo-doublons (*account/river* ; *accounts/river*), les paires renvoient pour l'essentiel aux deux sens de *bank*. Pour les 2 premières paires correspondant aux 12 mots choisis, 9 paires sur 24 (37.5%) correspondent aux sens distingués par Yarowsky. Il est également possible d'appliquer des méthodes de classification aux contextes associés aux mots ambigus, avec un risque : *les grands regroupements (top-level cluster partitions) basés uniquement sur l'information distributionnelle n'entrent pas forcément en correspondance avec les acceptions généralement reconnues* (Yarowsky, 1995 : 195).

## 5. Paramétrages, précautions et évaluation

### 5.1. Paramètres et précautions

Dans le cadre d'une sémantique distributionnelle, un obstacle partagé par les différentes tâches (acquisition, désambiguïsation) est l'émiettement et le déséquilibre des distributions des sens d'un mot. A titre d'exemple, les 1 345 phrases contenant *vendre* ou un de ses dérivés extraites du corpus PAROLE de 14 millions de mots de numéros extraits aléatoirement des années 1987, 1989, 1991, 1993 et 1995 du journal *Le Monde* ne contenaient aucun exemple du sens 'trahir', pourtant présent dans tous les dictionnaires. Le dégroupement de sens bute sur le même obstacle. Dans Rapp (2003), qui s'appuie pourtant sur les 100 millions de mots du BNC, les deux premières paires pour *bass* (*fish/music*) sont *guitar/treble* (soprano, clef de sol) et *guitar/string* : ce échec est attribué par Rapp à la mauvaise représentation du sens *fish* dans le corpus. Il en va de même pour *tâche* dans Aussenac-Gilles *et al.* (2003) dont les deux sens ('structure de représentation'/'tâche prescrite') sont très inégalement représentés. Cela doit conduire à veiller non seulement à une taille minimale des corpus utilisés mais à leur diversité thématique. Il est par ailleurs possible (section 4.3.) qu'un nombre limité de parties facilite les mises en évidence de sens divergents.

Par ailleurs, la répartition des traits permettant de classer les mots est souvent très éparpillée : les matrices résultantes sont fortement creuses. Une des approches possibles est le calcul de similarités de second ordre (Grefenstette, 1994a). Dans Sebilot (2002), les mots pleins qui servent à classer les noms sont ainsi remplacés dans un deuxième temps par des regroupements de ces mots pleins. On passe d'une matrice 383x8 000 à une matrice 383x544 : l'espace des traits est divisé par 15.

Dans l'esprit de Grefenstette (1996) et de Curran et Moens (2002), on peut chercher la définition la plus opératoire des traits utilisés, les contextes, en faisant varier leur taille (empan limité de  $k$  mots / phrases, paragraphes) et leur nature (formes graphiques, lemmes, positions syntaxiques) et en tenant compte de la plus ou moins grande adéquation des textes traités aux outils de segmentation, d'étiquetage ou de structuration disponibles. Les deux contraintes à concilier

<sup>17</sup> Rapp pense que l'algorithme se généralise à 3...  $k$  sens. On peut en douter. En premier lieu, augmenter les sens multiplie les tuples à examiner et atténue les écarts entre la somme des vecteurs de ces tuples. En second lieu, savoir le nombre de sens effectivement présents reste un problème à part entière, à supposer même qu'un tel but soit accessible.

sont d'une part de disposer de suffisamment de contextes pour rapprocher/différencier les mots de manière fiable (ce qui privilégie les contextes larges et les formes graphiques) et d'autre part de bénéficier de contextes plus précis et aisément interprétables (ce qui avantage les collocations restreintes et les dépendances syntaxiques). On peut d'ailleurs découpler les deux problèmes, en utilisant des contextes larges et « grossiers » pour repérer les mots mouvants/stables sur le plan sémantique et en utilisant des contextes restreints et plus parlants pour aider à l'interprétation.

Enfin, la constitution des unités textuelles au sein desquelles opérer les dégroupements peut se révéler délicate. Dans bien des cas, les unités « naturelles » que sont le document ou le paragraphe conviennent. Il peut s'avérer nécessaire néanmoins de disposer d'un grain plus fin. Un compte rendu de débat suppose d'isoler les tours de parole et de les rattacher à des émetteurs. Le mécanisme des reprises et réponses dans un mail inséré dans un fil de discussion de forum implique éventuellement de rattacher des fragments d'un mail donné à des émetteurs distincts, pour éviter les fausses proximités liées aux propos qu'un acteur rapporte mais qu'il ne prend pas à son compte. Le nombre parfois élevé des intervenants dans un forum ou dans un débat peut demander en aval de cet éclatement de regrouper les émetteurs en catégories, pour que les variations de sens soient tout simplement perceptibles.

## 5.2. Évaluer ?

La désambiguïsation sémantique a permis le développement de méthodes d'évaluation, via les campagnes SensEval (Kilgariff et Palmer, 2000). Un répertoire de sens prédéterminé, basé sur un dictionnaire existant, sert à étiqueter manuellement — dans le corpus d'entraînement et dans le corpus sur lequel seront jugés les systèmes en compétition — l'ensemble, relativement restreint, de mots à désambiguïser. En acquisition de catégories sémantiques, le rattachement de deux mots à une classe (*cluster*) est considéré comme juste s'ils figurent dans une même catégorie de thesaurus (Grefenstette, 1994), thesaurus qui peuvent être éventuellement combinés (Curran et Moens, 2002).

L'emploi de dictionnaires ou de thesaurus existants pour le dégroupement de sens peut permettre, pour des mots « de langue générale », d'examiner la corrélation entre le nombre de sens distingués dans ces ouvrages de référence et le « degré de mouvance » fourni par une méthode développée pour le dégroupement<sup>18</sup>. Ces références<sup>19</sup> ne conviennent plus forcément pour les divergences de point de vue en veille sociale. Une évaluation a posteriori des listes de mots univoques / « mouvants » fournies par les algorithmes est peu fiable : les facteurs conditionnant la satisfaction ou l'insatisfaction observées sont peu contrôlables. La présence d'experts du domaine peut permettre d'envisager une comparaison des listes obtenues avec les jugements formulés a priori par ces experts sur les mots les plus fréquents du corpus. On risque néanmoins de se heurter au paradoxe mis en évidence par les expériences rapportées par Véronis (2004) : la tâche de repérage de mots polysémiques semble facile aux annotateurs mais elle débouche sur de faibles taux d'accord.

Comme Yarowsky, on peut aussi engendrer des « pseudo-mots » : fusionner en une étiquette arbitraire deux mots dont on sait qu'ils étiquètent le sens de deux homonymes ou d'un mot polysémique. C'est inverser la démarche de Rapp : engendrer *bank-river*, *grid-tools*, etc. On

<sup>18</sup> Bien que le traitement de la polysémie et de l'homonymie varie sensiblement d'un dictionnaire à l'autre, la ligne de partage devra être conservée. La polysémie est sans doute plus difficile à détecter. On notera d'ailleurs que les 12 mots ambigus traités dans Rapp (2003) correspondent à des homonymes. Les résultats se dégraderaient probablement avec des mots polysémiques.

<sup>19</sup> Malheureusement non aisément accessibles pour la recherche sur le français (à la différence de WordNet et de multiples thesaurus en ligne pour l'anglais).

peut alors examiner si ces pseudo-mots sont effectivement repérés comme « mouvants ».

### 5.3. Interpréter

Les distinctions sémantiques en dégroupement de sens sont labiles et sujettes à caution. Dans le même esprit que Aussenac-Gilles *et al.* (2003), il nous paraît crucial de ne pas céder aux mirages d'indices opaques et de revenir systématiquement aux contextes, en s'appuyant sur des architectures de gestion de corpus et de traitement adaptées (Folch, 2002).

## Références

- Aussenac-Gilles N., Bourigault D. et Teulier R. (2003). Analyse comparative de corpus : cas de l'ingénierie des connaissances. In *14èmes journées francophones d'ingénierie des connaissances (IC 2003)* : 67-84.
- Bourigault D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse en mathématiques, informatique appliquée aux sciences de l'homme. Paris, École des Hautes Études en Sciences Sociales.
- Bourigault D. et Fabre C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, vol. (25) : 131-151.
- Chateauraynaud F. (2003). *Prospéro : une technologie littéraire pour les sciences humaines*. CNRS Éditions.
- Curran J.R. et Moens M. (2002). Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)* : 231-238.
- Folch H. (2002). *Articuler les classifications sémantiques induites d'un domaine*. Doctorat en informatique. Université Paris XIII.
- Folch H. et Habert B. (2002). Articulating conceptual spaces using the Topic Maps standard. In Wood L. (Éd.), *Proceedings XML 2002*.
- Folch H. et Habert B. (2004). Langages de méta-données pour le Web sémantique : RDF et Topic Maps. In Ihadjadène M. (Ed.), *Outils et méthodes en recherche d'information*. Hermès. À paraître.
- Fuchs C. et Victorri B. (Eds) (1994). Continuity in linguistic semantics. *Linguisticae Investigationes Supplementa*, vol. (19). John Benjamins.
- Grefenstette G. (1994a). Corpus-derived first, second and third order affinities. In *Proceedings of EURALEX*.
- Grefenstette G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher.
- Grefenstette G. (1996). Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. In Boguraev B. and Pustejovsky J. (Eds), *Corpus Processing for Lexical Acquisition. Language, Speech and Communication* : 205-216.
- Habert B., Folch H. et Illouz G. (1999). Sortir des sens uniques : repérer les mots « mouvants » dans le domaine social. *Sémiotiques*, vol. (17). *Dépasser les sens iniques dans l'accès automatisé aux textes*, Habert B. (resp.) : 121-151.
- Habert B. et Tournier M. (1987). La tradition chrétienne du syndicalisme français aux prises avec le temps. Évolution comparée des résolutions générales CFTC, CFDT et CFTC-maintenue (1945-1985). *MOTS. Presses de la Fondation Nationale des Sciences Politiques*, vol. (14) : 21-46.
- Hanks P. (2000). Do word meanings exist ? *Computers and the Humanities*, vol. (34/1-2) : 205-215.
- Ibekwe-SanJuan F. (1998). Terminological variation, a means of identifying research topics from texts. In *Proceedings of ACL-COLING'98* : 654-661.
- Ide N. et Véronis J. (1998). Introduction to the special issue on word sense disambiguation : the state of the art. *Computational Linguistics*, vol. (24/1) : 1-40.
- Kilgariff A. et Palmer M. (Eds) (2000). *Senseval : Evaluating Word Sense Disambiguation Programs. Computers and the Humanities*, vol. (34). Kluwer.

- Kleiber G. (1999). *Problèmes de sémantique : la polysémie en question. Sens et structures*. Presses Universitaires du Septentrion.
- Laublet P., Reynaud C. et Charlet J. (2002). Sur quelques aspects du Web sémantique. In *Actes du GDR I3*.
- Lebart L., Morineau A. et Piron M. (1997). *Statistique exploratoire multidimensionnelle*. Dunod.
- Losee R.M. (1995). The development and migration of concepts from donor to borrower disciplines : Sublanguage term use in hard and soft sciences. In *Proceedings of 5th International Conference on Scientometrics and Informetrics* : 265-274.
- Losee R.M. (1996). Text windows and phrases differing by discipline, location in document, and syntactic structure. *Information Processing and Management*, vol. (32/6) : 747-767.
- Manning C.D. et Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- Rapp R. (2003). Discovering the meanings of an ambiguous word by searching for sense descriptors with complementary context patterns. *Terminologie et Intelligence Artificielle* : 145-155.
- Salton G., Wong A. et Wang C.S. (1997). A vector space model for automatic indexing. In Sparck Jones K. et Willett P. (Eds), *Readings in Information Retrieval* : 273-289. Article publié en 1975.
- Sammon J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computing*, vol. (18) : 401-409.
- Sébillot P. (2002). *Apprentissage sur corpus de relations lexicales sémantiques. La linguistique et l'apprentissage au service d'applications du traitement automatique des langues*. Habilitation à diriger des recherches. Université de Rennes I. (IRISA Documents d'habilitation 41).
- Véronis J. (2004). Quels dictionnaires pour l'étiquetage sémantique ? In Fuchs C. et Habert B. (Resp.), *Le français moderne. Traitement automatique des langues et linguistique*.
- Yarowsky D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting*. Association for Computational Linguistics : 189-196.