CHAPITRE VIII

ANNOTER UN CORPUS

Nous ne prétendons pas fournir ici une présentation exhaustive. L'éclatement des réalisations, dispersées dans les publications, l'évolution rapide des outils, les avancées théoriques et pratiques conduisent à un « instantané » fragmentaire. Il est en outre difficile de prévoir les tendances à moyen terme. Notre objectif est de donner une idée des grands axes ... et des difficultés.

Dans la tradition pragmatique anglo-saxonne, les publications concernant les corpus mentionnent souvent les coûts des différentes opérations nécessaires. Ces renseignements permettent de prendre la mesure des moyens à mobiliser pour disposer des corpus réellement adaptés aux recherches linguistiques. À l'échelle de la francophonie, ils donnent une idée de l'ampleur des efforts à fournir. Ces " coûts " sont cependant donnés à titre indicatif. Ils donnent un ordre de grandeur, ils n'autorisent pas vraiment des projections, des comparaisons. À chaque niveau, les types d'annotation diffèrent trop pour qu'une mise en parallèle soit aisée. Pour s'en tenir à l'étiquetage, la taille du jeu d'étiquettes peut changer du tout au tout le coût de la correction.

1. NETTOYAGE ET HOMOGENEISATION

La phase initiale de « nettoyage » et d'homogénéisation des textes collectés sous forme électronique est une étape souvent sous-estimée, alors qu'elle est cruciale. Dans certains cas, les textes à intégrer dans un corpus ont été frappés pour la circonstance : ils contiennent des fautes de frappe ou d'orthographe. Dans d'autres cas, ils sont issus d'une reconnaissance optique : il faut restituer les mots qui ont été répartis entre deux lignes, corriger les erreurs typographiques. Il peut s'agir également de textes déjà saisis pour d'autres fins (bandes de composition de livres ou de journaux), le codage qui y figure doit être pris en compte, pour être

transformé ou supprimé.

Nous ne connaissons pas d'étude spécifique sur les coûts de cette phase. Le compte-rendu du projet AVIATOR (Blackwell, 1993) permet néanmoins d'évaluer les difficultés rencontrées. L'objectif est ici de développer des filtres permettant de « nettoyer » du texte tout-venant pour étudier l'évolution presque au quotidien de l'anglais, dans la perspective d'un corpus de suivi (cf. chapitre VII). Deux millions et demi de mots, provenant du journal *The Times*, sont traités chaque mois. Le titre même de ce compte-rendu donne une idée de l'ampleur du problème : « Des données sales au langage propre ». Comme S. Blackwell le souligne (*ibid.*), la correction de ce qui semble être des erreurs typographiques ne va pas forcément de soi. Une orthographe non standard a parfois pour but d'imiter une prononciation étrangère, dialectale ou idiolectale. Ou bien le mot a été forgé dans une optique ludique[1] (mot-valise, déformations diverses). Il s'agit alors de choix délibérés de la part de l'énonciateur, qui doivent donc être conservés comme tels. Les données comprennent parfois des codes propres au traitement pour lequel les documents étaient destinés au départ (par exemple des indications de photocomposition). Les titres, sous-titres et légendes suscitent aussi un traitement spécifique : quoiqu'ils constituent des unités à part entière, à ne pas mêler au texte qui les environne, ils sont généralement dépourvus de ponctuation finale. Il faut donc distinguer leur début et leur fin.

2. SEGMENTATION

La segmentation consiste à découper une suite de caractères en « unités » : mots simples ou unités polylexicales.

2.1 Repérer les unités

Le repérage des « mots » est délicat[2]. Un certain nombre de caractères, en effet, fonctionnent tantôt comme séparateurs de mots tantôt comme composants de mots. C'est le cas du trait d'union, qui joint deux mots dans *vient-il*, mais pas dans *va-et-vient*[3]. C'est le cas encore de l'apostrophe : séparateur comme guillemet simple, pour signaler l'élision, composant dans *aujourd'hui*, les abréviations et la représentation du langage parlé : *v'la au't chose*. C'est le cas surtout de l'espace, partie intégrante des unités complexes : *une carte bleue*.

Les unités complexes occupent une place importante en français. On estime au cinquième d'un texte la surface qu'elles couvrent. Pour le français, des inventaires extrêmement fournis ont été réalisés au LADL, sous l'impulsion de M. Gross, aboutissant à un dictionnaire électronique de « mots composés » ou DELAC

[1] Cf. (Fiala et Habert, 1989 ; Renouf, 1993).
[2] (Silberztein, 1993, p. 111-136) montre la complexité des phénomènes.
[3] (Mathieu-Colas, 1994) montre l'hétérogénéité extrême des emplois du trait d'union dans les dictionnaires.

(Courtois, 1990 ; Silberztein, 1993, p. 60-108). Ce dictionnaire associe aux séquences retenues des indications sur leurs variations éventuelles (flexion, discontinuités, alternances lexicales) ainsi que leurs propriétés syntaxiques (transformations[4]).

Mentionnons la difficulté à découper automatiquement le texte en phrases : titres, énumérations séparées par des points-virgules, exemples insérés dans le texte et faisant interposition, etc. La ponctuation offre des indices peu fiables[5]. Le point est une marque d'abréviation, un séparateur dans des codes (01.41.13.24.63) ou des nombres (3.13) , un indice d'alignement (dans une table des matières) et une fin de phrase. Or le découpage en phrases est crucial pour de nombreux traitements : examen des cooccurrences, étiquetage et analyse syntaxique ...

2.2 Techniques

Pour isoler les « mots », on écrit des règles qui emploient le contexte pour statuer sur les limites des unités. Par exemple, un trait d'union ayant à sa droite un pronom clitique comme *je*, *tu*, *il* a un statut de délimiteur. Il sépare un verbe de son pronom sujet conjoint (un *t* d'appui peut s'interposer). Ces règles sont combinées avec le recours à des dictionnaires de mots simples ou complexes (par exemple, comprenant la liste des mots français qui incluent en leur sein l'apostrophe, comme *aujourd'hui* ou *prud'hommes*).

Le système INTEX[6] (Silberztein, 1993) est l'exemple d'un segmenteur associant règles et dictionnaires. À partir des dictionnaires électroniques du LADL, il assure le découpage initial d'un texte tout-venant, l'étiquetage des mots simples et la reconnaissance des unités polylexicales. Son approche est basée sur des règles et non sur des probabilités. Il combine deux traitements : la projection sur le texte des dictionnaires, ce qui associe à chaque " mot " la ou les étiquette(s) pertinente(s) ainsi qu'aux suites de mots (éventuellement discontinues) leurs lectures éventuelles comme " mots composés " ou " expressions composées ", puis une désambiguïsation par des " grammaires locales " (*ibid.* p. 154-167). Par exemple, la *phrase Luc a travaillé pour le Ministère de l'intérieur* admet deux interprétations (*ibid.*, p. 139) *: C'est de l'intérieur que Luc a travaillé pour le Ministère* et *C'est pour le Ministère de l'intérieur que Luc a travaillé*. Il y a conflit entre deux unités polylexicales : *Ministère de l'intérieur* et *de l'intérieur*. La représentation produite signale les deux découpages : Luc a travaillé pour le 1[Ministère 2[de l'intérieur]2]1 où les indices identifient les deux possibilités. En l'occurrence, l'ambiguïté n'est pas levée. Dans d'autres contextes, on peut trancher. Des « grammaires locales » élaguent le graphe que constitue le texte dans lequel ont été ajoutées les étiquettes des mots simples et les expressions et mots composés. Elles permettent d'éliminer certains chemins[7]. Par exemple, lorsqu'un mot peut être pronom clitique ou déterminant et qu'il est suivi d'une

[4] Par exemple, *analyse des données* au sens statistique n'accepte pas le pluriel pour *analyse* ni le singulier pour *données* ni le remplacement de *des* par *de*.
[5] Pour le rôle de la ponctuation dans l'analyse syntaxique, voir (Nunberg, 1990).
[6] Les techniques éprouvées des automates et des transducteurs à états finis lui donnent une grande efficacité.
[7] Soulignons l'extrême généralité du traitement effectué. Cela permet d'utiliser INTEX pour d'autres traitements : étiquetage sémantique etc.

forme qui ne peut être qu'un verbe, comme dans : *Max le veut*, l'étiquette {pronom clitique} est éliminée.

2.3 Difficultés

Les unités polylexicales occupent une place fondamentale dans le lexique. Un segmenteur qui ne dispose pas d'inventaires de ces unités va « émietter » à tort les textes. De multiples techniques ont été testées pour faciliter le repérage automatique de ces mots complexes. Certaines d'entre elles ont été évoquées au chapitre II. D'autres reposent sur le filtrage statistique des mots qui « s'attirent » au sein d'un contexte restreint, d'autres encore sur l'utilisation de patrons syntaxiques (du type [{nom} {préposition} {nom}] comme cadre de vie), d'autres enfin combinent ces deux approches (Daille, 1993). Cependant, nombre de séquences proposées par ces outils ne constituent pas en fait des dénominations (cf. II 3.3)[8]. Les inventaires d'unités complexes réalisés pour le TALN suscitent généralement la perplexité ou la contestation sur la délimitation faite et sur le choix de considérer telle séquence comme une unité dénominative plutôt que comme un syntagme libre. Le risque symétrique de l'« émiettement » est de considérer à tort des suites de mots comme des unités polylexicales.

L'utilisation de dictionnaires comprenant un nombre important d'unités complexes fait naître en outre des ambiguïtés pour les séquences qui fonctionnent comme un tout dans certains domaines et qui sont à considérer comme des syntagmes libres dans d'autres. Dans « l'analyse des données montre que ... », le segment *analyse des données* peut renvoyer à une famille précise de techniques statistiques (présentée dans le chapitre IX), et c'est alors une unité, ou bien il doit être pris « au pied de la lettre », comme un groupe de mots sans lien particulier[9]. Plus les inventaires d'unités complexes s'étendent, plus ils rendent probables ces rencontres de hasard. Il n'est pas toujours sûr qu'il faille faire l'hypothèse, lorsqu'on rencontre une séquence inventoriée, de la présence effective de cette séquence.

3. ÉTIQUETAGE MORPHO-SYNTAXIQUE

Attribuer à chaque mot la ou les étiquettes possibles peut se faire par

[8] Cet excédent s'explique partiellement par le caractère encore fruste des techniques employées. Il tient plus fondamentalement aux limites de nos connaissances sur les mécanismes langagiers de création d'unités dénominatives. Les contraintes sémantiques à l'œuvre sont encore très peu explorées. Enfin, les dénominations possibles constituent un sur-ensemble des dénominations effectives, il n'est pas sûr qu'on puisse modéliser la manière dont une communauté langagière choisit au sein des dénominations possibles.

[9] On ne sait pas attacher de manière fiable à une unité polylexicale une indication de domaine (*analyse de données* : mathématiques, statistiques) et encore moins s'en servir pour n'utiliser que les unités propres au domaine, d'autant que les domaines sont « perméables » : la linguistique peut recourir à l'expression *analyse des données* dans ses deux acceptions.

consultation d'un dictionnaire, où chaque forme est suivie d'une liste de catégories, ou par analyse morphologique, ou par combinaison des deux techniques. Pour lever l'ambiguïté, deux solutions, qui peuvent d'ailleurs être associées, s'offrent alors : le recours à des règles ou l'appel aux probabilités (ce qui est sans doute la tendance dominante)[10].

3.1 Taux d'ambiguïté

Il est nécessaire, pour évaluer la tâche d « désambiguïsation » morpho-syntaxique, c'est-à-dire le choix de l'étiquette correcte parmi les étiquettes possibles, d'évaluer le nombre moyen d'étiquettes pour un mot. M. El Bèze et T. Spriet (1995) donnent les informations suivantes : « [...] une très grosse part de l'ambiguïté syntaxique est détenue par un petit nombre de mots fréquents [..]. De plus, ces mots sont essentiellement des mots outils. Ils appartiennent à des classes fermées et jouent un rôle syntaxique bien cerné dans la littérature. » Ils précisent (*ibid.* p. 58) : " [...] 30 % de l'ambiguïté est détenue par les 8 mots ambigus les plus fréquents[11] (50 % par les 36 premiers) mais il faut traiter 1 825 formes différentes pour lever 90 % de l'ambiguïté[12]. » E. Tzoukermann *et al.* (1996) précisent ce premier constat sur deux ensembles de 94 882 et 200 182 occurrences respectivement, tous deux extraits du journal *Le Monde* (septembre-octobre 1989 et janvier 1990) :

Nombre d'étiquettes	% du corpus de 94 882 mots	% du corpus de 200 182 mots
1	57 %	58 %
2	26 %	25 %
3	11 %	11 %
4	0,5 %	1 %
5	0,9 %	2 %
6	2 %	2 %
7	0,5 %	0,5 %
8	0,5 %	0,1 %

Plus de la moitié des mots ne soient pas ambigus. Le nombre de mots pouvant relever de 4 à 8 étiquettes est très restreint (4.4 % dans le premier cas, et 5.6 % dans le second). Le taux moyen d'ambiguïté par mot se monte alors à 1.72 pour le premier corpus et à 1.81 pour le second[13].

[10] J.-P. Chanod et P. Tapanainen (1995b) les comparent précisément, à partir d'une même segmentation et d'un même analyseur morphologique. Ils donnent l'avantage à l'approche par règles.
[11] Ces 8 formes sont : *la le l' les en un une a*.
[12] Les chiffres de J.-P. Chanod et P. Tapanainen (1995b) concordent globalement.
[13] M. El-Bèze et T. Spriet (1995, p. 52-53) donnent des chiffres proches.

3.2 Désambiguïsation par règles

Certaines suites de catégories sont illicites. Par exemple, deux étiquettes sont possibles pour *le* {déterminant} ou {pronom} et pour *guide* {verbe} ou {nom}. Cependant, toute la combinatoire n'est pas réalisable dans la séquence *le guide*. Des quatre possibilités, seules sont actualisables [{Pronom} {verbe}] (*il le guide*) et [{déterminant} {nom}] *(le guide commence son exposé*). On peut donc écrire une première règle d'élagage qui remplace la combinatoire par les deux seules suites licites de catégories. On utilise alors des règles « négatives ».

D'autre part, certaines formes permettent d'édicter des règles « positives ». Elles imposent en effet des contraintes fortes sur celles qui les précèdent ou les suivent immédiatement. Ainsi, *me* ou *te* sont suivis soit d'un pronom clitique (*il me le donne*) puis d'un verbe soit directement d'un verbe. On peut alors s'appuyer sur cette information pour éliminer des ambiguïtés. Dans *il me le garde, le* ne peut être qu'un pronom clitique et *garde* qu'un verbe. De telles formes servent de levier pour désambiguïser une partie de leur entourage. On parle d'« îlots de confiance ». Les clitiques post-posés et reliés par un trait d'union offrent également de tels appuis (dans *Route-t-il correctement le courrier, route* ne peut être qu'un verbe). Les formes nouvellement désambiguïsées servent à leur tour de point d'appui : les îlots de confiance vont croissant.

Les outils de désambiguïsation sont donc de manière générale des « grammaires locales » (Silberztein, 1993) qui prennent en entrée le graphe correspondant à la projection des différentes étiquettes sur le texte et éliminent une partie des chemins de ce graphe, ou inversement qui rajoutent des chemins (par exemple pour rendre compte des unités complexes comme *bien que* ou *carte bleue*)[14]. Les automates ou transducteurs correspondants ne savent pas traiter les dépendances à longue distance que l'on trouve en syntaxe. C'est également le cas en désambiguïsation probabiliste.

3.3 Désambiguïsation probabiliste

La désambiguïsation probabiliste s'appuie sur le caractère positionnel de langues comme le français et l'anglais, lequel fournit des contraintes locales fortes. Dans le graphe orienté des étiquettes possibles pour chacun des mots, il s'agit de chercher le chemin de probabilité maximale. Le choix de l'étiquette la plus probable en un point donné se fait au regard de l'historique des dernières étiquettes qui viennent d'être attribuées. En général, cet historique se limite aux deux ou trois étiquettes précédentes, on parle alors de bigrammes ou de trigrammes. Il repose sur des chaînes de Markov (Calliope, 1989, p. 360-370, Mérialdo, 1995, p. 11-13).

Ces méthodes supposent de disposer d'un corpus d'apprentissage. Ce corpus d'apprentissage doit être d'une taille suffisante pour permettre une estimation fiable des probabilités des suites de catégories et des différentes catégories d'un

[14] J.-P. Chanod et P. Tapanainen (1995b) ont ainsi développé un étiqueteur qui comprend 75 règles. E. Tzoukermann *et al.* (1995) donnent des exemples des règles qu'ils ont mises au point pour le français.

mot donné dans ces enchaînements. Le coût de préparation de ce corpus d'apprentissage est important. On procède alors par approximation. Un premier corpus d'apprentissage, relativement court, permet d'étiqueter un corpus plus important. Celui-ci est corrigé, ce qui permet de réestimer les probabilités. Il sert donc à un second apprentissage. Et ainsi de suite.

Les unités polylexicales sont mal prises en compte dans cette approche. Ainsi, pour reprendre l'exemple de M. El-Bèze et T. Spriet (1995), les adjectifs et participes placés immédiatement à droite du nom composé *cour d'appel* s'accordent avec *cour* et non avec *appel*. La probabilité d'un adjectif ou d'un participe passé féminin singulier après un nom masculin singulier comme *appel* sera pourtant donnée comme très faible par le corpus d'apprentissage, à juste titre d'ailleurs. Plus généralement, les désambiguïsations qui reposent sur un contexte large échappent à ce type de méthode. Des ambiguïtés comme première / troisième personne du singulier dans *je ne le pense pas* / *il ne le pense pas* ne sont pas éliminées, parce que ces étiqueteurs probabilistes s'appuient sur le contexte de la catégorie précédente, voire des deux catégories précédentes, pour trancher, et qu'ici il faudrait prendre en compte les trois catégories précédentes (Chanod et Tapanainen, 1995a).

L'approche probabiliste suppose par ailleurs que le corpus d'apprentissage ne présente pas des fonctionnements langagiers trop différents du corpus à étiqueter. Dans le cas de ***BNC***, un certain nombre de mots comme *I*, *well* et *right* etaient mal étiquetés dans la partie orale du corpus dans la mesure où l'apprentissage avait été réalisé sur la partie écrite (Leech *et al.*, 1994).

3.4 Performances

Aucun dictionnaire ne peut être entièrement exhaustif. En outre, les entrées du dictionnaire peuvent être incomplètes (certaines catégories, pourtant possibles, en sont omises). Un analyseur morphologique ne fournit pas non plus d'hypothèses sur la totalité des mots à étiqueter. Il reste donc toujours des « mots inconnus », ne serait-ce qu'en raison des noms propres, des mots empruntés à des langues étrangères ou des néologismes (*débureaucratiser*).

Les taux habituellement cités tournent autour de 95 à 98 % d'étiquettes justes. Ce chiffre paraît encourageant. Cependant, ces performances incluent souvent les ponctuations parmi les formes étiquetées. Or les ponctuations couvrent environ 10 à 15 % de la surface des textes, ce qui diminue d'autant le nombre des formes lexicales qui sont effectivement correctement catégorisées. Par ailleurs, nous l'avons vu, une bonne moitié des formes d'un texte ne relève que d'une catégorie et d'une seule. La désambiguïsation est donc à comptabiliser sur le reliquat seulement, ce qui double le pourcentage d'erreur. Notons enfin que 5 % d'erreur, c'est une étiquette erronée tous les 20 mots, soit plus d'une fois par phrase dans un texte courant. Une telle « performance » handicape un parseur intervenant en aval.

La fiabilité d'un étiqueteur donné est à évaluer à l'aune des tâches qui vont avoir recours par la suite au texte étiqueté : les enjeux ne sont pas les mêmes s'il s'agit d'analyse syntaxique automatique ou d'étude de la répartition de certains patrons

morpho-syntaxiques. Il convient aussi de comparer les résultats affichés avec ceux qui proviennent d'une intervention manuelle. M. Marcus *et al.* (1993) indiquent : « l'étiquetage manuel a pris à peu près deux fois plus de temps que la correction d'un étiquetage automatique, avec un taux de désaccord entre personnes étiquetant à peu près double, et un taux d'erreur presque de 50 % plus élevé. »

Il est en outre extrêmement difficile de comparer les performances : les jeux d'étiquettes, leur taille changent d'un système à l'autre : 37 catégories pour (Chanod et Tapanainen, 1995), 253 pour Tzoukermann *et al.,* 1995) par exemple. Le taux d'ambiguïté d'un étiquetage est en effet proportionnel à la taille du jeu d'étiquettes employé. Il faut également tenir compte de la stabilité des résultats : si le taux d'ambiguïté restant ne varie que faiblement (1.2 %) dans les expériences d'E. Tzoukermann *et al.* (1995) selon qu'ils emploient un jeu de 67 ou de 253 catégories, 2.5 % des formes ont été analysées différemment, (Stein et Schmid, 1995, p. 29), des résultats relativement divergents sont donc fournis. En outre, les ambiguïtés possibles ne sont pas de même nature : on ne peut mettre sur le même plan l'hésitation entre nom et verbe (*porte*) et celle entre adjectif et participe passé. Dans ce cas, la levée d'ambiguïté n'a pas les mêmes conséquences pour les traitements ultérieurs : considérer un mot comme adjectif ou participe passé changera peu la place qui lui sera attribuée dans la structure construite.

3.5 Post-traitement et coûts

Pour un usage linguistique fin, le post-traitement manuel s'avère en tout cas indispensable. Malgré les environnements spécialisés qui ont été développés, la correction reste coûteuse. Dans le cadre de ***BNC***, elle est évaluée (Leech *et al.,* 1994), après le passage d'un étiqueteur probabiliste (CLAWS4, basé sur les chaînes de Markov), au taux de succès de 96 à 97 %, à 40 minutes de travail spécialisé pour 1 000 mots, soit 41 années-homme pour 100 millions de mots. Il faut en outre prendre en compte le nombre d'étiquettes : plus il est grand, plus il rend difficile la correction manuelle. Cette difficulté pousserait à choisir des étiquettes « connues », basées sur le savoir grammatical courant (sur la terminologie grammaticale traditionnelle), pour faciliter le travail des correcteurs et l'utilisation ultérieure par des chercheurs (Greenbaum, 1993).

Pour le corpus de l'université de Lancaster, près de 39 minutes (Black *et al.*, 1994, p. 60) sont nécessaires au traitement de 1 000 mots (pré-traitement, passage de l'étiqueteur probabiliste CLAWS, correction manuelle).

3.6 Evaluation et nouvelles tendances

Eric Brill (1995) résume ainsi les points forts et les faiblesses des deux approches : « [Les] étiqueteurs stochastiques ont bien des avantages sur les étiqueteurs bâtis manuellement, en particulier ils rendent superflue la construction laborieuse de règles manuelles, et saisissent des

informations utiles qui peuvent ne pas avoir été remarquées par l'analyste humain. Cependant, les étiqueteurs stochastiques présentent l'inconvénient que les connaissances linguistiques ne sont capturées qu'indirectement, par le biais de grands tableaux statistiques. »

L'écriture de règles se heurte rapidement à la complexité des interactions effectives entre les règles. En effet, chaque règle agit sur un texte qui a été modifié par les règles précédentes. Il faut donc prévoir autant que faire se peut ces interactions, qui peuvent devenir d'une complexité très grande, voire ne plus être maîtrisables. À l'inverse, la mise au point des règles peut s'appuyer sur l'intuition des locuteurs.

L'étiquetage et la désambiguïsation, comme d'autres secteurs de l'annotation des données textuelles, donnent lieu à des approches mixtes, où un étiquetage probabiliste est corrigé *in fine* par des règles du type de celles évoquées ci-dessus, ou vice-versa[15].

Les techniques d'apprentissage sont également mises à contribution. La tentative la plus achevée est actuellement celle d'E. Brill (1995), dont l'étiqueteur est en cours d'adaptation pour le français. Le système dispose d'un dictionnaire associant aux formes les probabilités qu'elles portent telle ou telle catégorie. La catégorie la plus probable est projetée sur le corpus de mise au point. Les erreurs commises ainsi sont repérées par comparaison avec la version étiquetée à la main de ce corpus. Le système propose des règles de correction, assez proches finalement de celles qui ont été évoquées ci-dessus. Elles sont de la forme : changer une étiquette *a* en étiquette *b* si le mot précédent est étiqueté *w*. Elles prennent en compte un contexte étroit : deux positions avant ou après la forme examinée. Sont retenues les règles qui améliorent le plus l'état de la catégorisation, c'est-à-dire qui enlèvent le plus d'erreurs et en ajoutent le moins. Ces règles sont alors appliquées. Une nouvelle comparaison et une nouvelle génération et application de règles sont opérées, jusqu'à ce qu'il ne soit plus possible de corriger le texte sans ajouter davantage d'erreurs qu'on n'en corrige. C'est une autre forme, automatique cette fois, du processus mentionné de « tâche d'huile » autour d'îlots de confiance. E. Brill indique par exemple que son système « apprend » 447 transformations sur un corpus d'entraînement de 600 000 mots avec une exactitude de 97.2 %, mais que les 100 premières suffisent à assurer une désambiguïsation exacte à 96.8 % (*ibid.*, p. 557). Ces règles peuvent s'appuyer soit sur les catégories, éventuellement multiples, soit aussi sur les mots dominés par les catégories.

Pour reprendre les termes de Leech et de ses collègues (1994, p. 61) : « La guerre contre l'erreur est [...] une guerre d'usure, dans laquelle des stratégies variées sont employées, mais où il ne faut pas s'attendre à une solution-miracle. Le rôle de la personne qui corrige *a posteriori* reste crucial, mais l'élimination de l'erreur est une tâche qui est, petit à petit, passée à l'ordinateur. »

[15] Comme l'indiquent M. El-Bèze et T. Spriet (1995, p. 48) : " [...] il suffit d'écrire 4 à 5 règles pour traiter environ 50 % des erreurs commises par un système probabiliste. " E. Tzoukermann *et al.* (1995) constituent comme autant de modules un analyseur morphologique, un ensemble de règles d'élagage et un étiqueteur probabiliste : ils les combinent de diverses manières (en retenant 43 possibilités, jouant sur des seuils et des ordres distincts) et examinent les performances selon les choix, ce qui les conduit à utiliser d'abord les règles puis les probabilités.

4. ANALYSE SYNTAXIQUE

Nous mentionnons avant tout l'analyse syntaxique automatique. L'analyse syntaxique manuelle nécessite surtout de disposer d'un environnement informatique facilitant la tâche de parenthésage et de catégorisation des constituants. Elle rend plus cruciale la vérification de l'homogénéité des résultats.

4.1 Structuration par règles

4.1.1 Règles « négatives »

On retrouve pour le parsage une technique déjà utilisée pour l'étiquetage : l'élagage (*pruning*). Il s'agit dans le domaine syntaxique d'utiliser des règles « négatives », qui ont pour fonction d'éliminer les hypothèses non justifiées. C'est l'approche du parseur ENCG, ce qui amène Voutilainen et Heikkila (1994, p. 190) à parler d'analyseur « réductionniste ». Pour chaque étiquette morphologique d'un mot donné, les fonctionnements syntaxiques possibles sont fournis. Par exemple, un nom peut être sujet, objet, complément prépositionnel, etc. L'élagage élimine les fonctionnements illégitimes en contexte. Ces contraintes syntaxiques (400 dans le cas présent) sont elles-mêmes issues d'études intensives de corpus (Karlsson, 1994, p. 122). En principe, ces règles d'élagage sont indépendantes les unes des autres et n’ont pas besoin d’être ordonnées. Il semblerait cependant qu’une grammaire ENCG reste assez « fragile ».

4.1.2 Règles " positives "

Elles peuvent être de complexité plus ou moins grande. Les grammaires à affixes du projet TOSCA (Nederhof et Koster, 1993, p. 166-170) qui décorent des règles hors contexte d'affixes représentant des paramètres, des attributs ou des traits, permettent une grande finesse de comportement : vérification des accords et des compatibilités sémantiques etc.

4.2 Structuration probabiliste

Les parseurs reposant sur des règles butent sur deux types de problèmes, comme le rappelle M. Rajman (1995, p. 158) : la couverture linguistique et l'ambiguïté. Couverture : les règles mises au point sont soit trop permissives (elles acceptent des énoncés incorrects) soit au contraire trop restrictives (elles refusent des agencements de mots pourtant

valides). Ambiguïté : le nombre d'hypothèses proposées est souvent très important (cf. chapitre II).

L'idée générale du parsage probabiliste[16] est de remplacer la distinction binaire acceptable / non acceptable pour un couple <séquence, structure> par une probabilité, les séquences inacceptables pouvant correspondre alors à une probabilité nulle (*ibid.* p. 159). Les deux problèmes mentionnés trouvent là leur solution. Certains agencements sont reconnus comme rares, mais possibles. D'autres prennent une place centrale, leur probabilité étant forte. La probabilité attribuée à chaque structure pour une phrase donnée permet de classer les structures par probabilité croissante, et de garder la ou les structures de plus forte probabilité. Un corpus arboré de départ sert à l'apprentissage du modèle : la probabilité des différentes réalisations d'un syntagme donné est estimée à partir de sa fréquence dans ce corpus[17]. L'utilisation du modèle sur un corpus plus large permet de vérifier l'adéquation du modèle et de l'améliorer (en accroissant le corpus d'apprentissage).

4.3 Performances et évaluation

Puisque, nous l'avons vu, l'annotation syntaxique peut varier énormément en complexité, il est malaisé de comparer les résultats de différents parseurs. Une des possibilités, encore peu explorée (Atwell *et al.,* 1994), consiste à « aligner » plusieurs représentations syntaxiques d'un même texte. Une version rudimentaire de cette approche (Black *et al.,* 1993, p. 4) consiste à réduire l'annotation aux parenthésages, en éliminant toutes les étiquettes, pour ne garder donc que les découpages structurels et leurs emboîtements. On peut alors aisément comparer deux parenthésages et repérer les désaccords. C'est ce qui est appelé (*ibid.*) le « score de cohérence structurelle » (*structural consistency score*). Une autre optique consiste à soumettre un ensemble de phrases de test à plusieurs analyseurs et à comparer, avant tout manuellement, leurs résultats. Cette deuxième démarche sert plutôt à examiner de manière fine les réactions des parseurs : chaque phrase est centrée autour d'un phénomène syntaxique bien défini, elle est donc souvent relativement simple par rapport aux énoncés effectivement rencontrés par les parseurs dédiés au texte tout venant. On manque en tout état de cause de données comparatives.

Un premier critère d'évaluation est celui de la justesse linguistique des résultats retenus. Elle est difficile à apprécier. On peut tout de même opposer des analyseurs (et partant des corpus arborés) qui visent à un simple dégrossissage et ceux qui, au prix éventuellement d'un post-traitement important, aboutissent à des

[16] (Rajman, 1995) fournit une introduction générale aux modèles probabilistes pour l'analyse syntaxique. (Black *et al.,* 1993) constitue une présentation beaucoup plus détaillée, à la fois en ce qui concerne l'apprentissage des paramètres d'un modèle probabiliste et pour l'interaction entre approche par règles et analyse probabiliste. Ce livre résulte d'une collaboration étroite, pendant cinq ans, entre le centre de recherche IBM Watson et l'université de Lancaster (UCREL - Unit for Computer Research on the English Language).

[17] En principe, ce corpus doit être aussi vaste que le permettent les moyens rassemblés. La précision des estimations qu'il autorise en dépend. La collaboration IBM Watson - Université de Lancaster a abouti par exemple à l'analyse manuelle de 800 000 mots (Black *et al.,* 1993, p. 16).

analyses vérifiées et cohérentes au sein du cadre théorique choisi et qui peuvent servir de pierre de touche à des recherches linguistiques fines. Pour le système TOSCA, H. van Halteren et N. Oostdijk (1993, p. 155) indiquent que, pour les textes de fiction, dans 88 % des cas, l'analyse juste fait partie des résultats produits par le parseur, alors que cette proportion tombe à 56 % pour les textes qui ne relèvent pas de la fiction. Malheureusement, ils ne fournissent pas d'hypothèses sur les raisons de ce décalage. Les textes « informatifs » comprennent-ils des phrases plus longues, des constructions spécifiques (par exemple propres à des disciplines scientifiques ou techniques) qui ne se rencontreraient pas dans les textes de fiction ? Selon A. Voutilainen et J. Heikkila (1994, p. 194), le parseur ENCG donne l'étiquette syntaxique correcte d'un mot dans 96 % des cas (85 % environ des mots n'ont plus qu'une seule étiquette syntaxique à la fin du processus d'émondage, mais avec un taux d'erreur de 3 %). Les constats de (Black *et al.,* 1993, p.2-5), voici quelques années, sont plus sévères. Les auteurs parlent de « déplorable état de l'art » (*ibid.* p. 2) et citent trois expériences peu encourageantes. Dans la première, trois des auteurs chercheurs à IBM Watson ont procédé de manière indépendante, en 1990, à l'évaluation de quatre parseurs importants pour l'anglais, sur 35 phrases de 13 mots extraites au hasard de dépêches (2 millions de mots) de l'agence *Associated Press*. Les avis concordaient : un des systèmes analysait 60 % des phrases correctement. Les scores des trois autres parseurs allaient de 35 à 40 % de résultats justes. Deuxième expérience : en 1992, le concepteur d'un parseur important a pris 50 phrases de 13 mots dans ***Brown***, en variant les genres choisis. Il a indiqué les frontières de constituants à la main, préparant ainsi la « bonne réponse ». Il a ensuite utilisé son parseur : les résultats étaient corrects dans 30 % des cas seulement. Troisième expérience : la comparaison en 1992 des résultats de sept parseurs sur 100 phrases de longueur variable (de 4 à 69 mots avec une moyenne de 22 mots) tirées au hasard d'un million de mots du *Wall Street Journal*. La correction moyenne du simple parenthésage (sans prendre en compte les étiquettes) ne dépassait pas 22 %, et les résultats s'étalaient de 16 % à 41 % de résultats structuralement corrects.

Un second critère d'appréciation, concernant les parseurs et les grammaires qu'ils utilisent, est la réutilisation possible ou effective de l'approche soit sur d'autres secteurs de la même langue soit pour d'autres langues. C'est ainsi que le parseur ENCG développé pour l'anglais a été adapté au suédois, au danois et au basque (Voutilainen et Heikkila, 1994, p. 191).

Un troisième critère, lié au précédent, mais plus difficile à apprécier, parce que moins factuel, est celui de la " coloration théorique " des conventions d'annotation. À quel cadre théorique sous-jacent renvoient-elles ? Notons tout de même que la tendance est plutôt, sinon à des notations consensuelles, ce qui n'a pas grand sens, du moins à des pratiques évitant les distinctions controversées et les parti-pris méthodologiques trop marqués[18]. C'est nécessaire pour que le corpus puisse être réutilisé (Black *et al.,* 1993, p. 37).

Il est enfin un critère que nous écarterons, celui du temps nécessaire au parsage lui-même[19]. D'abord parce qu'il est difficile de donner des informations

[18] Une exception au moins : le corpus de 65 000 mots d'oral transcrit (enfants de 6 à 12 ans) analysé manuellement (Polytechnic of Wales) qui s'inspire étroitement de la Grammaire Fonctionnelle Systémique de Halliday.

[19] A titre anecdotique, deux chiffres, empruntés à Hindle (1994, p. 116) : avec Fidditch, de l'ordre de 6 heures pour analyser un million de mots, et presque deux semaines pour analyser 44 millions de mots de dépêches de l'agence *Associated Press*.

comparables (les langages informatiques utilisés, la taille des mémoires, leur configuration changent notablement le sens des mesures). Ensuite parce le temps de calcul n'est plus une ressource rare, et qu'en outre l'amélioration des performances des machines le réduit continuement. Enfin, parce que l'optimisation des parseurs est un art fructueux[20], mais qu'il faut probablement attendre une plus grande maturité du domaine pour qu'elle soit vraiment à l'ordre du jour pour les corpus arborés.

4.4 Post-traitement et coûts

C'est la phase de « nettoyage » manuel des résultats fournis par le parseur utilisé.

Il peut s'agir, comme pour le système TOSCA, de choisir entre les analyses alternatives proposées (Halteren et Oostdijk, 1993, p. 157-159). Sont utilisées des forêts partagées (*shared forests*), qui mettent en facteur commun les sous-arbres partagés. L'annotateur examine la phrase en contexte et sélectionne à chaque point d'ambiguïté le sous-arbre approprié.

A l'inverse, dans le cas de ***Penn Treebank***, où le parseur déterministe *Fidditch* (Hindle, 1994), fournit une analyse syntaxique unique pour chaque phrase, mais laisse des constituants non rattachés, la tâche des annotateurs est d'attacher les constituants « orphelins ». Voici pour la phrase *Battle-tested industrial managers here always buck up nervous newcomers with the tale of the first of their countrymen to visit Mexico, a boatload of warriors blown ashore,* l'état des traitements fourni dans (Marcus *et al.,* 1993, p. 322-325) :

1) Analyse syntaxique automatique produite par *Fidditch* :

Les constituants non attachés débutent par ?. Les syntagmes prépositionnels commençant par *of* sont attachés à un nom s'ils en suivent un (c'est le cas pour *tale of*, *boatload of*), et restent non attachés dans le cas contraire (*first of*). Les virgules, qui peuvent jouer le rôle de conjonctions, fragmentent aussi l'ensemble d'arbres.

```
( (S
  (NP (NBAR (ADJP (ADJ "Battle-tested/JJ")
                  (ADJ "industrial/JJ"))
            (NPL "managers/NNS")))
  (? (ADV "here/RB"))
  (? (ADV "always/RB"))
  (AUX (TNS *))
  (VP (VPRES "buck/VBP"))
  (? (PP (PRES "up/RP")
         (NP (NBAR (ADJ "nervous/JJ")
```

[20] F. Karlsson indique ainsi (1994, p. 142) qu'une réécriture du parseur ENCG a fait passer le temps d'analyse de 3 à 5 mots seconde à 400 à 500 mots seconde ...

```
                (NPL "newcomers/NNS")))))
        (? (PP (PREP "with/IN")
             (NP (DART "the/DT")
               (NBAR (N "tale/NN")
                  (PP of/PREP
                    (NP (DART "the/DT")
                      (NBAR (ADJP
                           (ADJ "first/JJ")))))))))
        (? (PP of/PREP
             (NP (PROS "their/PP\$")
               (NBAR (NPL "countrymen/NNS")))))
        (? (S (NP (PRO *)
                (AUX to/TNS)
                (VP (V "visit/VB")
                  (NP (PNP "Mexico/NNP")))))))
        (? (MID ",/,"))
        (? (NP (IART "a/DT")
             (NBAR (N "boatload/NN")
                (PP of/PREP
                  (NP (NBAR
                       (NPL "warriors/NNS"))))
                (VP (VPPRT "blown/VBN")
                  (? (ADV "ashore/RB"))
                  (NP (NBAR (CARD "375/CD")
                         (NPL "years/NNS")))))))
        (? (ADV "ago/RB"))
        (? (FIN "./.")))
```

2) Après simplification automatique et avant correction manuelle :

La représentation est simplifiée pour faciliter la tâche des annotateurs en rendant le résultat visuellement plus clair et en éliminant des distinctions mineures (nom propre / nom commun, par exemple).

```
( (S
  (NP (ADJ Battle-tested industrial)
    managers)
  (? here)
  (? always)
  (VP buck)
```

```
(? (PP up
      (NP nervous newcomers)))
(? (PP with
      (NP the tale
         (PP of
            (NP the
               (ADJP first))))))
(? (PP of
      (NP their countrymen)))
(? (S (NP *)
     to
     (VP visit
        (NP Mexico))))
(? ,)
(? (NP a boatload
      (PP of
         (NP warriors))
   (VP blown
      (? ashore)
      (NP 375 years))))
(? ago)
(? .))
```

3) Après correction manuelle :

L'environnement utilisé permet d'attacher un constituant, de changer sa position dans l'arbre, de modifier son étiquette ... Grâce à des notations spécifiques, on peut d'une part indiquer qu'une séquence est un constituant majeur mais que sa catégorie syntaxique est sujette à discussion, et d'autre part rendre compte des ambiguïtés réelles : c'est le cas pour *blown ashore 375 years ago* qui peut modifier soit *warriors* soit *boatload*, d'où l'indication *pseudo-attach*.

```
( (S
   (NP Battle-tested industrial managers
      here)
   always
   (VP buck
      up
      (NP nervous newcomers)
      (PP with
```

```
            (NP the tale
               (PP of
                  (NP (NP the
                        (ADJP first
                           (PP of
                              (NP their countrymen)))
                        (S (NP *)
                          to
                          (VP visit
                             (NP Mexico))))
                      ,
                      (NP (NP a boatload
                            (PP of
                              (NP (NP warriors)
                                (VP-1 blown
                                    ashore
                                    (ADVP (NP 375 years)
                                        ago)))))
                        (VP-1 *pseudo-attach*)))))))
    .)
```

4.5 Coûts

Pour l'insertion manuelle d'arbres syntaxiques rudimentaires (parenthésage et étiquetage des constituants), la vitesse peut atteindre une phrase par minute (Black *et al.,* 1993, p. 20). La moyenne pour l'analyse syntaxique manuelle effectuée à l'université de Lancaster est de 51 minutes pour 1 000 mots : cela comprend pré-traitement, parenthésage et étiquetage grossier dans un environnement informatique spécifique et post-traitement (*ibid.* p. 60).

D'après (Marcus *et al.,* 1993, p. 323), la correction des résultats du parseur utilisé pour ***Penn Treebank*** suppose un temps d'apprentissage (de l'ordre de deux mois) plus long que le nettoyage de l'étiquetage. La vitesse moyenne de correction est alors de l'ordre de 475 mots l'heure (voire 575 ou 675 quand les sorties du parseur sont simplifiées avant correction). L'évaluation faite est la suivante (*ibid.*) : « À un taux moyen de 750 mots par heure, une équipe d'annotateurs à temps partiel travaillant 3 heures par jour devrait arriver à 2,5 millions de phrases analysées corrigées en un an, chaque phrase étant corrigée une seule fois. »

Il faut en outre prévoir le temps de familiarisation avec les conventions d’annotation syntaxique. (Black *et al.,* 1993) indique ainsi qu'il a fallu attendre

six mois d'apprentissage en moyenne avant que le travail d'un annotateur devienne optimal.

4.6 Difficultés

Tout ne ressortit pas à un format d'arbre. C'est le cas des éléments parenthétiques qui forment des structures autonomes, non reliées au reste de la phrase. Cela suppose que le parseur puisse suspendre l'analyse englobante, effectuer celle d'un tel élément, et reprendre l'analyse de plus haut niveau (Briscoe, 1994, p. 98). À supposer que l'on arrive à analyser automatiquement de telles structures, il reste à disposer des notations adéquates.

La distinction entre les arguments d'un verbe et ses simples modifieurs s'avère extrêmement délicate à ajouter de manière cohérente. Le dessein, ***dans Penn Treebank***, était d'ajouter manuellement cette information. La difficulté rencontrée a conduit à faire machine arrière. De la même manière, ***Susanne*** n'a pas réussi, malgré des efforts soutenus des annotateurs, à intégrer un classement des compléments en termes de grammaire de cas, à la Fillmore : « la nature des relations logiques que des prédicats variés entretiennent dans l'usage réel avec leurs arguments s'est avérée trop diverse pour un tel traitement, et l'équipe croit avoir 'testé jusqu'à épuisement'[21] l'hypothèse selon laquelle la structure propositionnelle de base en anglais peut être adéquatement décrite grâce à un ensemble limité de 'cas » (Sampson, 1994, p. 185). Les relations entre les pronoms et leurs antécédents n'ont pas non plus été ajoutées à ***Susanne***, probablement moins par peur de déboucher sur des apories que faute de moyens.

Toute grammaire « fuit », pour reprendre une image souvent employée dans la communauté du parsage robuste. L'idée de rendre compte de l'ensemble des phénomènes syntaxiques de la langue (on parle de la « couverture » de la grammaire utilisée par un parseur) est un fantasme, stimulant certes, comme tous les mythes, mais illusoire, comme le soulignent du point de vue linguistique J.-M. Marandin (1993) et du point de vue du TALN T. Briscoe (1994, p. 100). Une raison de fond : la langue varie. Dans le temps d'abord. Mais aussi selon les genres discursifs et les domaines d'emploi. À la différence des langages formels utilisés en logique ou en informatique, l'ensemble des règles n'est pas donc fini. Ce constat, classique pour le lexique, soulève plus de réticences en syntaxe.

5. ÉTIQUETAGE SEMANTIQUE

L'une des grandes méthodes d'analyse sémantique de corpus suppose des connaissances préalables et consiste à projeter ces connaissances sur le corpus pour en faire ressortir certaines propriétés. C'est sur ce principe que repose le travail de M. Sussna (1993) et la plupart des

[21] *tested to destruction*

recherches en matière de désambiguïsation lexicale.

Le principe général de cette méthode est simple. On étiquette le corpus pour l'enrichir d'informations sémantiques. Pour ce faire, on exploite généralement des données lexicales et non contextuelles, connaissances générales sur les sens d'un mot, le concept ou le thème auquel il renvoie. Ceci permet alors d'observer le fonctionnement du mot en contexte. De multiples expériences ont été menées dans cette optique[22] : elles diffèrent par le jeu d'étiquettes utilisé et par la méthode d'étiquetage.

Toutefois, les données lexicales initiales font parfois défaut. C'est même souvent le cas lorsque le corpus à traiter relève d'une langue spécialisée. Il faut alors commencer par construire les catégories sémantiques devant servir à étiqueter le corpus.

5.1 Construire des catégories sémantiques

La difficulté de réutiliser les bases lexicales spécialisées, l'inadéquation des bases lexicales générales et plus fondamentalement le manque de ressources lexicales, notamment pour le français (cf. chapitre III), soulèvent la question de l'acquisition des connaissances lexicales. La construction manuelle de ce type de base de données requiert l'expérience d'un lexicographe et, pour les langues spécialisées, celle d'un expert du domaine. Le coût et la difficulté de ces entreprises ont mis à l'honneur les méthodes automatiques ou semi-automatiques qui considèrent les corpus comme des sources de connaissances pour la construction de catégories sémantiques, dans l'idée qu'elles puissent servir ensuite à étiqueter des corpus.

La construction de ces catégories sémantiques — qu'il s'agisse de classes de synonymes, de groupes de mots relevant d'un même champ sémantique ou d'un même thème — suit toujours le même principe général. La démarche consiste à :

- définir le contexte d'un mot, de manière à identifier les mots qui cooccurrent avec lui, l'ensemble des mots qui figurent dans le même contexte et qui, dans une approche distributionnelle de la sémantique en décrivent le sens ;
- définir une mesure de similarité entre les mots deux à deux, chaque mot étant représenté par les relations de cooccurrence dans lesquelles ils entrent ;
- exploiter cette mesure de similarité pour construire des classes de mots considérés comme équivalents selon le point de vue considéré (par exemple, des synonymes ou des mots relevant du même domaine…).

À ces trois étapes correspondent trois « ordres d'affinité » (Grefenstette,

[22] Une variante de cette méthode consiste à projeter des connaissances non pas sous la forme d'étiquettes destinées à enrichir le texte, mais sous la forme de patrons qui permettent de sélectionner de manière ciblée des données considérées comme pertinentes. Nous ne développons pas cet aspect ici. (Hearst, 1992) exploite, par exemple, cette méthode pour rechercher des relations hyponymiques dans un corpus destinées à enrichir un thesaurus existant.

1994b), trois niveaux de relations entre les mots[23] : les relations de cooccurrence, de similarité et d'équivalence[24]. Le travail de G. Grefenstette présenté au chapitre IV suit cette démarche générale. Nous nous appuyons sur cet exemple dans ce qui suit.

5.1.1 Définir un contexte

Le choix de la nature du contexte dépend du corpus exploité et des relations sémantiques recherchées. G. Grefenstette retient le syntagme nominal pour identifier les noms sémantiquement voisins et le document pour construire les familles de mots (cf ; chapitre IV, section 2). Trois grandes classes de contextes peuvent être identifiées : les contextes graphiques, syntaxiques et documentaires. L'extrait de ***Menelas*** suivant montre la différence, pour le mot *épisode*, entre une fenêtre de 7 mots (encadrée) et le contexte syntaxique tel que le définit (Grefenstette, 1994) (en italiques) :

> Depuis cette époque on ne note aucune récidive d'angor jusqu'à il y a 8 jours où il a *présenté* un *épisode* de *précordialgie survenant* à l'effort, durant environ 45 minutes, sans irradiation[25].

Les contextes graphiques se définissent comme des fenêtres de mots : deux mots cooccurrent s'ils figurent à moins de *x* mots de distance[26] dans l'ordre linéaire du texte. La taille de la fenêtre dépend des relations sémantiques que l'on recherche, les cooccurrences à petite, moyenne et grande distance tendant respectivement à faire ressortir des expressions figées ou semi-figées (*prendre pour*, *avoir faim*), des contraintes de sélection (*boire* / *vin*) et des mots appartenant au même champ sémantique (Lafon, 1981; Church et Hanks, 1990). Le calcul des fenêtres graphiques ne nécessitant qu'un corpus segmenté, elles sont souvent privilégiées pour le traitement de gros corpus.

L'apparition de corpus arborés permet désormais de définir des contextes syntaxiques. Seuls les mots appartenant au même syntagme ou, mieux, en relation de dépendance syntaxique sont alors retenus comme cooccurrents. Pour étudier les contraintes de sélection, on considère ainsi les relations sujet-verbe ou verbe-objet (Church et Hanks, 1990 ; Hindle, 1990) tandis qu'on prend le groupe nominal comme contexte pour repérer les classes d'adjectifs (Assadi et Bourrigault, 1995). Cette approche syntaxique suppose de disposer d'un corpus arboré ou partiellement arboré et généralement désambiguïsé sur le plan morpho-

[23] Nous ne considérons ici que les relations entre mots, mais les affinités peuvent être calculées pour d'autres unités : on a vu (en III-2) que G. Grefenstette calcule des similarités entre des expressions, en l'occurrence des groupes nominaux (1993).

[24] Nous généralisons le propos de G. Grefenstette en décrivant le troisième ordre d'affinité comme celui des relations d'équivalence plutôt que comme celui des axes sémantiques qui nous semblent avoir un statut intermédiaire entre la similarité et l'équivalence.

[25] Nous n'avons pas considéré ici que les groupes prépositionnels *durant 45 minutes* et *sans irradiation* devaient être rattachés à *épisode*. Pour l'anglais, G Grefenstette résout le problème du rattachement du groupe prépositionnel par des règles *ad hoc* (1994).

[26] En général, les relations de cooccurrence ne sont pas orientées et l'ordre dans lequel figurent les mots est indifférent.

syntaxique[27], mais elle engendre moins de bruit que l'approche graphique[28] : les contextes linguistiquement aberrants (l'association *jours – épisode* dans l'exemple ci-dessus) sont éliminés. Cela rend cette approche bien adaptée aux corpus de taille moyenne (Basili *et al.*, 1993a ; Bouaud *et al.,* 1997).

Les contextes documentaires, enfin, sont définis à partir d'une unité textuelle (paragraphe, partie, article, chapitre, document…). C'est ce type de contexte que G. Grefenstette définit pour le calcul des variantes.

De nombreux auteurs ne retiennent par ailleurs que les contextes les plus significatifs. Ce filtrage *a posteriori* des contextes préalablement extraits est le plus souvent statistique[29] : on ne retient comme cooccurrents que les mots figurant « anormalement » souvent dans les mêmes contextes[30].

5.1.2 Calculer des similarités

Une fois définie la notion de contexte, on peut calculer pour un mot l'ensemble de ses cooccurrents, sa distribution. Cette distribution sert alors à représenter les mots et permet de les comparer entre eux. C'est l'approche suivie par G. Grefenstette et décrite au chapitre IV. Concrètement, cela signifie qu'un mot se représente par un vecteur sur l'ensemble des cooccurrents possibles, *i.e.* sur l'ensemble des mots du corpus. La similarité entre deux mots est mesurée comme une distance entre les vecteurs représentant chacun de ces mots[31].

Ces mesures de similarités sont difficiles à exploiter en tant que telles. Les scores obtenus ne s'interprètent pas dans l'absolu mais seulement relativement les uns aux autres. Par ailleurs, les mesures ou les classements obtenus résistent à l'interprétation. On a souvent besoin de savoir sur quels critères deux mots sont rapprochés

Le problème vient plus fondamentalement de ce qu'une liste triée des

[27] On peut toutefois proposer des méthodes de pondération des analyses concurrentes en cas d'ambiguïté syntaxique. Voir par exemple (Grishman et Sterling, 1994).

[28] « [N]on seulement les associations syntaxiques reflètent une information fonctionnelle, ce que ne font pas les paires rapprochées sur une base graphique, mais la méthode d'extraction de ces associations syntaxiques est aussi *plus efficace*, le nombre d'associations utiles détectées étant considérablement plus élevé que ce qu'on obtient par des méthodes reposant sur une distance graphique. » (Basili *et al*., 1993a, p. 154). L'analyse syntaxique fonctionne en effet comme un premier filtre.

[29] Ce n'est cependant pas le seul type de filtrage possible : pour la recherche de collocations, F. Smadja (1993) filtre les collocations sur une base syntaxique, ou même en fonction de leur degré de figement.

[30] Voir par exemple (Lafon, 1981), (Church et Hanks, 1990) ou (Justeson et Katz, 1996). D'autres auteurs, visant la construction de classes sémantiques plutôt que la recherche de collocations, considèrent au contraire que le seul fait qu'un contexte soit attesté une fois suffit à le rendre significatif (Bensch et Savitch, 1995 ; Bouaud, 1997). Signalons par ailleurs qu'un filtrage statistique ne peut s'effectuer que sur un volume important de données.

[31] Nous préférons parler ici de similarité entre les mots plutôt que de distance comme le font les travaux de classification automatique. Le terme de distance sémantique est d'ordinaire employé pour désigner des distances calculées à partir d'une taxonomie ou d'un réseau (cf. *supra*). G. Grefenstette (1994) ou P. Bensch et W. Savitch (1995) s'inspirent de la mesure de Jaccard ou Tanimoto mais la littérature sur les méthodes de classification présente de multiples mesures de similarité (Saporta, 1990 ; Lebart et Salem, 1994) et différentes mesures sont employées en acquisition de connaissances sémantiques.

similaires d'un mot donné n'est pas une classe : ces listes sont centrées autour d'un mot pôle et ce n'est pas parce que *ship* (*navire*) et *truck* (*camion)*, par exemple, sont tous les deux similaires à *boat* (*bateau*) (Hindle, 1990) que les deux relations de similarités sont comparables ni que *ship* et *truck* sont nécessairement similaires entre eux. Partant de ce constat, G. Grefenstette (1994) propose de structurer cette liste des similaires d'un mot selon ses différents axes sémantiques, ce qui revient à distinguer différents types de similarités. J. Bouaud et ses collègues (1997) choisissent de représenter un ensemble de relations de similarités sous la forme d'un graphe qui situe un mot dans un réseau de similarités et fait ressortir des zones denses, riches en similarités croisées. Pour aller plus loin dans cette voie, il faut construire des classes sémantiques à partir d'une relation d'équivalence entre les mots. C'est là pour nous le véritable troisième ordre d'affinité.

5.1.3 Construire des classes de mots

Cette étape n'est pas abordée dans le traitement lexicographique de G. Grefenstette (1993), mais cette piste est explorée par d'autres auteurs, pour la modélisation d'un domaine, notamment[32]. En interprétant le score de similarité entre les mots comme une mesure de distance entre des objets, on peut appliquer les méthodes de classification automatique pour construire des classes de mots. Il s'avère cependant que les classes induites à partir de corpus sont difficiles à exploiter. Les méthodes purement inductives produisent des regroupements de mots hétérogènes. Pour construire des catégories sémantiques cohérentes, il faut corriger ces premiers résultats en fusionnant ou en scindant certaines classes pour obtenir une granularité régulière, en éliminant les intrus, parfois en reconstituant « à la main » des classes complètement éclatées.

Pourtant, si l'on considère l'ampleur et la difficulté de la tâche consistant à donner une description lexicale de l'ensemble des mots d'un corpus, et d'un corpus spécialisé notamment, il s'avère que les connaissances lexicales induites à partir de corpus, aussi bruitées et imparfaites soient-elles, sont précieuses. Ce sont des ébauches qui proposent une première organisation du matériau lexical et permettent d'amorcer le travail de description. A. Mikheev et S. Finch (1995) soulignent par exemple l'intérêt de ces méthodes de classification pour la modélisation des connaissances d'un domaine : « [l]a construction de classes sémantiques de mots à partir de corpus permet au cogniticien de repérer les principales catégories ou principaux types sémantiques existant dans le domaine en question et d'organiser le lexique en regard de ces types. ».

[32]Voir, entre autres, (Assadi et Bourrigault, 1995), (Bensch et Savitch, 1995), (Mikheev et Finch, 1995), (MacMahon et Smith, 1994) ou (Bouaud *et al*., 1997).

5.1.4 Procéder par itérations

La construction de catégories sémantiques repose généralement sur une alternance d'induction de connaissances à partir de corpus et d'interprétation, *i.e.* de projection de connaissances extérieures au corpus. Une première classification permet d'identifier une ou plusieurs classes cohérentes qui peuvent être figées puis projetées sur le corpus sous la forme d'un étiquetage partiel. Seuls les mots de ces premières classes porteront une étiquette de classe, mais ils constituent des îlots de confiance à partir desquels une nouvelle classification peut être construite[33]. Cette méthode incrémentale est donc une méthode mixte consistant à induire des connaissances même parcellaires que l'on peut ensuite projeter sur le corpus pour en induire de nouvelles.

Une variante de cette démarche incrémentale part non des premières classes induites mais d'un étiquetage grossier du corpus. C'est ce que font R. Basili *et al*. (1993b) ou R. Grishman et J. Sterling (1994) mais aussi Z. Harris (voir chapitre VII).

5.2 Projeter des catégories sur un corpus

5.2.1 Segmentation en unités sémantiques

Déjà présente au niveau morpho-syntaxique, la question de la segmentation du corpus se pose d'autant plus au niveau sémantique que la tradition fait davantage défaut. Quelle unité de sens faut-il retenir ? On considère souvent le mot, par solution de facilité parce que les sources lexicales utilisées sont elles-mêmes structurées autour des mots, aux expressions polylexicales et mots composés près. Dans certains cas, cependant, les unités inférieures sont à étiqueter : pour une étude thématique de ***Enfants***, les préfixes négatifs doivent être comptés au même titre que les adverbes de négation, lesquels comportent au contraire généralement plusieurs mots (*ne… pas*). Il est par ailleurs souvent difficile d'identifier les mots qui, dans un syntagme ou dans une phrase, doivent porter une étiquette donnée. Dans ***Enfants***, les expressions *difficultés financières*, *pas assez d'argent*, *considérations financières* ont toute une connotation négative, mais à quel mot associer cette étiquette négative ?

[33] C'est la démarche adoptée par Bouaud *et al.* (1997) ou P. Bensch et W. Savitch (1995, p. 12) : « quand on applique notre technique de classification […] à un corpus réel, elle identifie un ensemble de catégories qui paraissent naturelles, sans toutefois classer beaucoup de mots dans ces catégories. Mais, il s'est avéré que ce petit nombre de mots classifiés dans un premier temps pouvait servir de point de départ pour classifier d'autres mots. ».

5.2.2 Désambiguïsation sémantique

Si les problèmes d'ambiguïté sont négligés — dans la langue de spécialité notamment —, l'étiquetage peut se faire hors contexte, sur la liste des formes du texte. C'est l'approche de (Basili *et al*., 1993c) semble-t-il. Pourtant, l'objectif est généralement de désambiguïser le corpus et l'étiquetage doit être fait en contexte.

L'étiquetage manuel est envisageable pour les corpus de taille moyenne (en deçà du million de mots) s'il faut choisir parmi quelques étiquettes générales parce que les cas ambigus sont rares et faciles à trancher : « Une fois qu'une classe sémantique est clairement définie, avec l'aide d'une interface conviviale, l'étiquetage à la main d'un mot est l'affaire de quelques secondes. Nous avons résolu de simplement sauter les mots pour lesquels le choix d'une étiquette n'est pas évident[34] ou pour lequel aucune étiquette ne paraît adaptée.» (*ibid.*, p. 346-347). « On n'a pas forcément besoin de faire appel à un linguiste pour l'étiquetage, [même si] on a besoin d'un linguiste pour établir un jeu d'étiquettes approprié. » (Basili *et al*., 1993a, p. 157).

S'il faut procéder à un étiquetage fin en revanche, la procédure manuelle devient sujette à erreur, difficile à homogénéiser et surtout trop coûteuse. « [L]a partie du corpus Brown qui est étiquetée par les classes de mots de WordNet, un exemple de corpus important , disponible et désambiguïsé à la main, montre clairement combien il est difficile d'obtenir des données 'satisfaisantes'. Ce corpus est relativement petit (de l'ordre de quelques centaines de milliers de mots) en comparaison de la taille des corpus actuels (plusieurs millions ou dizaines de millions de mots) ; la méthode d'annotation qui a été utilisée est très coûteuse en temps de travail […] ; et la qualité des résultats reflète la difficulté de la tâches standards actuels (les annotateurs sont en désaccord dans environ 10% des cas […]). » (Resnik, 1995).

D'où le besoin de méthodes automatiques robustes de désambiguïsation de corpus et l'intérêt des travaux qui, comme (Sussna, 1993), cherchent à les mettre au point.

[34] C'est-à-dire s'il prend plus de « 30 secondes » (Basili *et al*., 1993a) (NDA).