Méthode pour l'analyse automatique de structures formelles sur documents multilingues

THÈSE

présentée et soutenue publiquement le 21 décembre 1998

pour l'obtention du

Doctorat de l'Université de Caen

(spécialité informatique)

par

Emmanuel Giguet

À Marine.

Table des matières

Liminaires

Chapitre 1

Introduction

«Sortez vos cahiers. Aujourd'hui, leçon de grammaire.»

Leçon de grammaire... grammaire... syntaxe... Des mots qui ne laissent personne indifférent. Ils réveillent en chacun de nous des souvenirs si lointains. Les cours de primaire et de secondaire, le stylo à quatre couleurs dans une main, la règle dans l'autre, les heures passées à déterminer la couleur à associer à chaque groupe, à chaque proposition. Les groupes sujets en bleu, les groupes verbaux en rouge, les compléments d'objet direct en vert... Oui, c'est cela. Voilà en quelque sorte le point de départ de cette thèse.

L'analyse linguistique automatique se trouve au confluent de plusieurs disciplines : la linguistique, l'objet d'étude étant la langue, la psycholinguistique, puisque nous étudions une production humaine, et bien sûr l'informatique, l'analyse devant être réalisée automatiquement, voire efficacement. Nous sommes dans le domaine interdisciplinaire de la linguistique informatique, du traitement automatique des langues.

Nous vous proposons un reportage sur l'expédition que nous avons menée durant quatre années. Nous commencerons par quelques promenades épistémologiques qui nous permettront de prendre connaissance du terrain étudié, à savoir l'écrit, et de la démarche scientifique ayant permis son étude, une étude dans laquelle la forme est objectivée. Nous prendrons également connaissance de l'outil que nous utilisons pour l'étudier, à savoir l'ordinateur.

Nous poursuivrons par une excursion dans le multilinguisme au cours de laquelle nous présenterons les documents dans leur dimension multilingue et la nécessité de les traiter comme tels. Nous étudierons leur structure multilingue et expliquerons comment la calculer à l'aide d'un identificateur de langues et d'un segmenteur multilingue.

Nous serons alors armés pour entamer sans préjugé notre voyage dans l'analyse syntaxique automatique. C'est en effet avec un regard neuf qu'il faudra aborder ce voyage car le paysage ne sera plus agrémenté des traditionnels arbres. C'est en terme de flux qu'il faudra dès lors imaginer l'objet. La première étape décrit comment passer d'une vision statique à une vision dynamique de l'objet. La deuxième présente la construction de la structure syntaxique en détaillant successivement la technique d'analyse des syntagmes minimaux et leur mise en relation.

Tout au long de cette présentation, nous mettrons l'accent sur la méthode. C'est elle qui nous permettra d'allier élégamment des concepts linguistiques et informatiques.

Chapitre 2

Promenades épistémologiques

2.1 Circonscription de l'objet

Circonscrire l'objet d'étude paraît être une étape incontournable de l'exposé de ces travaux. Elle permettra au lecteur une meilleure compréhension de la démarche scientifique que j'ai adoptée et des choix théoriques qui ont été les miens. En effet, c'est par une confrontation permanente avec cet objet que nous avons pu définir nos propres concepts. C'est par cette même confrontation que ceux-ci ont évolué et continuent aujourd'hui de s'affiner.

Cet objet, c'est l'écrit. Un objet pluridimensionnel. Un objet pouvant être considéré de si nombreux points de vue que l'infime partie à laquelle nous allons nous intéresser pourra paraître déconcertante. Ces restrictions sont cependant guidées, imposées, par la nature des recherches menées, à savoir l'analyse syntaxique automatique. Examinons précisément les caractéristiques de ce matériau.

2.1.1 L'écrit, un objet observable

L'écrit est un objet concret, résultat d'une production humaine. En tant qu'objet concret, il est observable. En tant que résultat d'une production humaine, sa forme est d'une grande variabilité car elle est soumise à toutes sortes de facteurs extra-linguistiques qui affectent leurs auteurs et vont ainsi modifier les conditions de production.

L'écrit n'en reste pas moins observable et c'est une caractéristique fondamentale dans notre recherche car elle nous permet de construire des outils pour l'explorer et formuler des hypothèses sur sa structure, puis de confronter ces hypothèses à la réalité de l'objet.

Il ne s'agit donc pas d'un objet immatériel tel que la *compétence*, connaissance qu'un sujet parlant a de sa langue et objet d'étude de Noam CHOMSKY (1965), mais au contraire d'une production bien réelle, d'une production concrète, issue de la *performance*, c'est-à-dire de la mise en acte effective du langage.

2.1.2 Le corpus, une collection de textes

Concrètement, les écrits prennent la forme d'une collection de textes, sans restriction précise sur le style. Nous considérons donc aussi bien les romans que les articles scientifiques, les articles journalistiques que les textes philosophiques. Le recours à des styles très variés permet, par extension, un meilleur aperçu de la couverture des structures conjecturées et donc leur validation empirique. Par ailleurs, sous-estimer l'importance de styles élaborés, tels la poésie, signifie se priver d'une précieuse source d'informations parce que l'on y rencontre souvent de grandes libertés stylistiques qui, comme nous le verrons par la suite, peuvent nous permettre de mieux comprendre la structure de tous les autres types d'écrits.

Bien que les auteurs de textes soient soumis à des facteurs internes tels que l'émotion, le stress, à des facteurs externes, le bruit par exemple, facteurs qui, comme nous l'avons rappelé, engendrent une grande variabilité de la forme des productions, tous restent cependant sous la même influence : celle de *l'activité d'écriture*. C'est cette influence qui nous a conduit à considérer le texte, et non la phrase isolée, comme cadre d'étude pertinent.

Lorsqu'un individu n'est pas plongé dans cette activité d'écriture, par exemple lorsqu'il construit de toutes pièces une phrase illustrant un phénomène linguistique particulier, nous constatons que la nature de l'objet est bien différente de celle des textes que nous étudions. Les structures syntaxiques utilisées sont souvent dégénérées et, sans remettre en cause leur grammaticalité, nous notons que la mise en exergue du phénomène particulier perturbe l'équilibre et nuit à la cohérence de l'énoncé produit. Bien que pouvant présenter, il est vrai, un intérêt linguistique ou pédagogique, ce type d'énoncé, que l'on peut qualifier d'artificiel, est non représentatif des textes que nous souhaitons analyser qui, eux, sont harmonieux. Nous avons donc été amené à les écarter de notre étude.

Le texte, lui, est un ensemble cohérent d'unités plus ou moins complexes, une manifestation particulière de l'activité d'écriture. Chaque unité s'articule avec les autres et contribue à la réalisation d'un équilibre global. Ainsi, la présence ou l'absence d'une ponctuation est significative, la position d'un mot, d'une phrase, d'un paragraphe même, n'est jamais purement fortuite.

Toutes s'inscrivent dans le cadre général d'un processus de construction de textes. C'est précisément ce processus qu'il nous faudra comprendre pour réaliser l'analyse syntaxique.

2.1.3 Les langues, une évolution perpétuelle

Beaucoup de chercheurs se sont intéressés aux mécanismes d'évolution des langues au travers d'études diachroniques. Dans *Éléments de syntaxe structurale* (1959), par exemple, Lucien TESNIÈRE expose la tendance à l'agglutination de mots systématiquement voisins. Il évoque également la disparition de certaines flexions, leur réapparition sous forme de mots grammaticaux. Tout un ensemble de phénomènes qui modifient plus ou moins profondément la forme, la structure des énoncés et qui nous rappellent que la syntaxe d'une langue n'est valide que pour une époque donnée.

Notre étude a porté sur des écrits contemporains. Cependant, nous n'avons pu nous contraindre à une étude purement synchronique car la justification de certaines de nos positions ne peut être établie que par la diachronie. Par ailleurs, en retrouvant l'origine de certains phénomènes linguistiques, il est aussi possible de modéliser au mieux leur trace dans les écrits que nous analysons.

Cette démarche s'est en fait imposée lors de l'analyse des différentes structures de notre objet. Pour concrétiser ce point sans trop entrer dans les détails, on remarquera son utilisation lors de la définition du mot, lors de la modélisation des amalgames, c'est-à-dire pour définir une structure de surface, mais aussi pour définir une structure plus interne, principalement lors de l'analyse de la projectivité de l'ordre structural sur l'ordre linéaire.

2.1.4 Une langue, des langues

Nos recherches en syntaxe ont été appliquées au français contemporain. Cependant, la confrontation permanente avec des corpus de langues variées et la nécessité de travailler sur de véritables unités linguistiques nous ont amené à adopter une démarche résolument multilingue.

Après étude du matériau, les premières hypothèses sur sa structuration syntaxique ont émergé, hypothèses que nous avons confrontées au corpus, mais manquant cependant d'une validation multilingue rigoureuse. Dans cet esprit, les travaux sur corpus de Hervé DÉJEAN portant sur la découverte des structures formelles des langues se sont révélés essentiels puisqu'en rejoignant les conclusions de nos propres investigations, ils ont contribué à la validation empirique de ces hypothèses. C'est ainsi que la notion de syntagme minimal, unité pivot de la structure syntaxique de la phrase dans notre théorie, se trouva consolidée par la mise en évidence de sa stabilité dans de nombreuses autres langues. Il en fut de même du concept de proposition qui, bien qu'écarté pendant un temps de notre approche, s'est finalement avéré incontournable.

Le choix de la langue française a été guidé par mon identité. Cette justification peut paraître *a priori* fragile car l'étude de la syntaxe d'une langue ne doit être basée que sur des critères formels. Aussi, lorsque cette étude est menée sur sa propre langue, elle peut se voir plus facilement perturbée par l'utilisation de critères extra-formels, tels l'interprétation. Dans notre recherche, l'approche orientée corpus et l'utilisation de l'informatique minimisent cependant considérablement en participant à l'objectivation : d'une part, l'approche orientée corpus laisse une place beaucoup plus réduite à l'introspection, d'autre part, le passage obligatoire par le filtre informatique, dans lequel il est très difficile d'introduire le sens et qui ne s'appuie par conséquent que sur la forme, restreint considérablement l'influence de l'interprétation.

2.1.5 L'écrit, un objet dynamique

Les sections précédentes nous ont permis d'évoquer le fait que l'écrit était le résultat d'un processus mental, une mise en acte effective du langage. Dès lors, on ne peut totalement passer sous silence certaines propriétés que l'écrit partage avec l'oral, notamment celle qui constitue l'essence même de notre recherche à savoir leur forme.

Il est commun dans notre domaine de s'offrir la liberté de voir l'écrit comme un objet unidimensionnel, c'est-à-dire en laissant de côté son découpage logique (en livres, chapitres, sections, paragraphes). De ce point de vue, l'écrit et l'oral sont des objets *unidimensionnels*. Des objets unidimensionnels véhiculant un message structuré, donc *pluridimensionnel*. La présentation sous forme linéaire du message, et donc de sa structure, détermine l'utilisation d'un codage. Le codage de la structure est réalisé par l'intermédiaire de traces laissées par le producteur du message pour permettre à l'interprétant (le lecteur ou l'auditeur) sa recomposition et son interprétation. Ce sont bien sûr ces traces, qui une fois repérées et analysées, nous permettront de tendre vers notre objectif, l'analyse syntaxique automatique.

La plupart de ces traces sont aujourd'hui bien identifiées même si la sémantique qu'elles véhiculent n'est pas toujours maîtrisée. Certaines marques se retrouvent aussi bien dans l'écrit que dans l'oral. On compte parmi elles des mots structurants tels que les conjonctions, les prépositions et les pronoms relatifs, des accords en genre, en nombre, en personne. D'autres sont spécifiques à l'une des deux formes : la ponctuation, la casse, le style des caractères pour l'écrit, la prosodie et la pause pour l'oral. Ces traces affectent la forme des énoncés et leur étude est le centre des recherches en syntaxe.

Nous nous sommes cependant arrêté sur l'étude d'une trace d'une toute autre nature, liée au caractère unidimensionnel de l'écrit et de l'oral. D'une toute autre nature car, contrairement aux précédentes, elle n'est pas visuelle ou auditive. Elle contraint cependant la forme des énoncés et mérite donc d'être considérée. Cette trace est issue de la *dynamique* de leur déroulement : le support confère certes une vision statique à l'écrit mais il n'en reste pas moins, tout comme l'oral, dynamique dans son déroulement. Plus précisément, dans l'oral aussi bien que dans l'écrit, le producteur et l'interprétant ont une activité dynamique : celle du producteur s'inscrit dans l'objet sous forme d'une trace, une trace du processus de création de l'énoncé, celle de l'interprétant permet sa restructuration. Cette activité dynamique détermine l'orientation de notre support unidimensionnel que l'on peut dès lors considérer comme un flux. L'orientation de ce flux impose, aussi bien lors de la production que lors de l'interprétation du message, une construction incrémentale de sa structure : le rattachement à la structure d'une nouvelle unité introduite dans le flux est contraint par la structure formée des unités précédentes, et la nouvelle structure ainsi définie va contraindre le rattachement des unités à venir.

Ce sont ces deux axes qui, pris en compte simultanément vont guider notre démarche scientifique.

2.2 Théorisation de l'objet

Cette section explicite la méthode de construction de notre théorie. Nous y présentons la technique d'élaboration des hypothèses de structures syntaxiques ainsi que les moyens de validation ou de réfutation auxquels nous avons eu recours. Après une analyse critique de cette démarche, nous présentons comment ces hypothèses deviennent concepts. Avant de commencer cet exposé, nous allons rappeler quels sont les objectifs motivant la théorisation de notre objet.

2.2.1 Objectifs des recherches

Bien que ma formation ait été effectuée dans une filière purement informatique, les travaux que je présente constituent une recherche en linguistique, sur corpus, assistée par ordinateur ; c'est donc une recherche en linguistique informatique. L'objectif en est la découverte des structures nécessaires à l'élaboration d'une théorie de la syntaxe de notre objet. Dans ce cadre, nous nous intéressons à la modélisation des structures et également à celle des processus permettant la génération de ces structures.

Nous n'oublions cependant pas qu'un analyseur syntaxique est un des modules-clés de tout système de traitement des langues. L'objectif secondaire sera donc de produire un analyseur utilisable dans un tel système. Par objectif secondaire, il faut entendre qu'un analyseur syntaxique incorporable dans un système de traitement des langues ne sera qu'un produit dérivé, une retombée opératoire de notre recherche première, en syntaxe, sur corpus, assistée par ordinateur.

Il ne faut en effet pas perdre de vue que la qualité des résultats d'un analyseur syntaxique placé dans un système de traitement des langues, que sa performance et que la qualité de sa conception seront entièrement dépendantes de la qualité des recherches menées en syntaxe. Plus les concepts utilisés seront en accord avec la nature de l'objet et meilleures seront la qualité et l'efficacité de l'analyseur. Plus la modélisation des concepts mettant en évidence les universaux et les spécificités de chaque langue sera précise et plus son architecture finale sera pertinente.

2.2.2 Exposé de la démarche scientifique

La figure 2.1 page ci-contre présente de manière schématique la démarche scientifique que nous avons suivie pour théoriser notre objet. Cette démarche peut être analysée en quatre étapes : (1) l'élaboration d'hypothèses de structures syntaxiques, qui passe par l'observation du matériau, (2) l'élaboration d'un système de règles et son informatisation, (3) la confrontation des hypothèses au corpus, (4) l'évaluation de l'analyseur linguistique et sa comparaison avec d'autres systèmes. Nous allons détailler ces quatre étapes et nous décrirons également la nécessaire prise de recul par rapport aux travaux effectués.

Élaboration d'hypothèses de structures

L'élaboration d'hypothèses de structures syntaxiques, appelées hypothèses par la suite, est la première étape de notre travail scientifique. Ces hypothèses naissent de plusieurs séries d'observations faites sur le matériau. Les observations sont réalisées à l'œil nu ou bien assistées par ordinateur : observations à l'œil nu lorsqu'il s'agit d'écrits traditionnels, livres, revues, magazines, observations assistées par ordinateur lors d'explorations de corpus électroniques.

C'est uniquement après cette phase de familiarisation avec le matériau, cette phase de découverte du matériau, qu'apparaissent les premières régularités. Régularités qui, une fois généralisées, nous permettent de former, au travers de propositions de structures, des hypothèses tentant d'expliquer le phénomène observé. Afin d'être validées, ces hypothèses sont soumises au contrôle de l'expérience, démarche tout à fait classique en sciences expérimentales. Dans notre recherche, le contrôle par l'expérience se réalise par une confrontation des hypothèses avec le corpus, via l'outil informatique.

Le corpus utilisé pour l'observation est du texte de langue française mais c'est pendant cette même étape que des études sont menées sur la validité multilingue des structures conjecturées. Ces études multilingues sont également menées sur corpus.

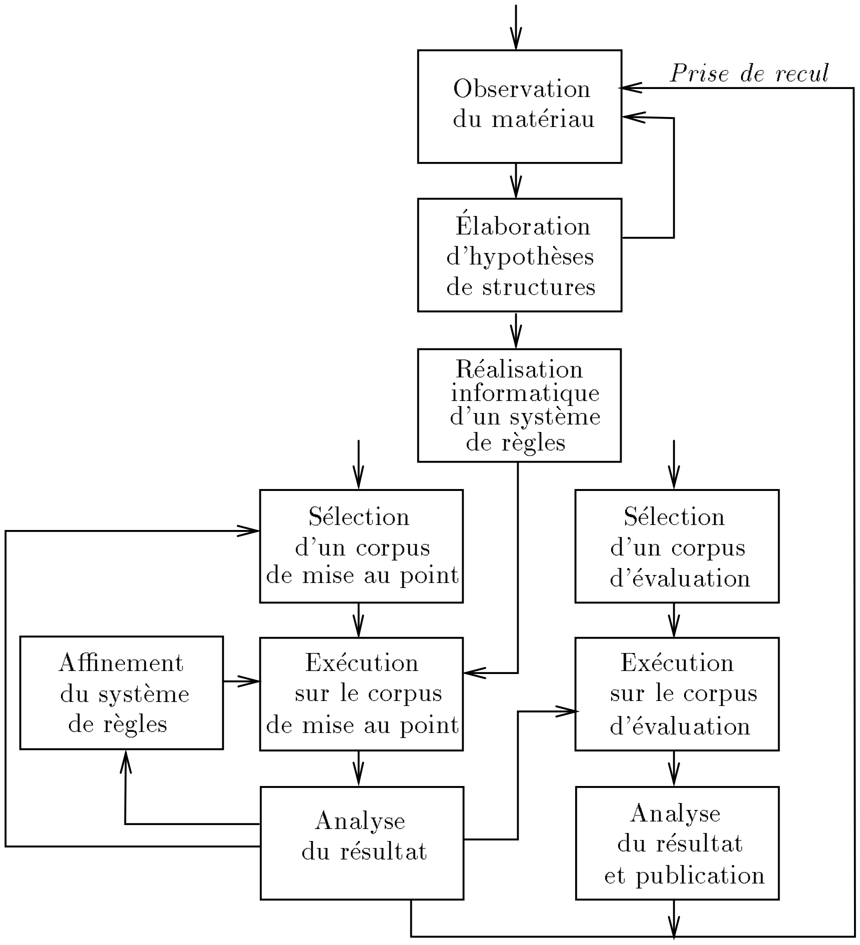


FIG. 2.1 - *Démarche scientifique*

Après avoir formulé une hypothèse, nous tentons de la valider ou de la réfuter par une confrontation avec le corpus. Cette confrontation passe par une réalisation informatique qui nécessite l'élaboration d'un système de règles.

Élaboration d'un système de règles

La confrontation des hypothèses au corpus via une réalisation informatique est caractéristique de l'approche orientée corpus. La contrainte de l'utilisation de l'outil informatique proposée par cette approche se révèle ici très enrichissante du point de vue linguistique : pour valider une hypothèse de structure sur corpus, il faut en exhiber toutes les réalisations concrètes, c'est-à-dire toutes les instances, dans un corpus, et cela, de manière automatique.

La construction automatique de toutes les instances d'une hypothèse à l'aide d'un ordinateur n'est pas immédiate. Il faut fournir à la machine, sous forme d'un programme, une méthode de calcul de ces instances. Cette méthode de calcul est appelée *système de règles*. Elle doit décrire en terme opératoire la construction de toutes les instances d'une hypothèse pouvant être réalisées dans un corpus. En effet, l'ordinateur est une machine qui ne sait qu'effectuer des calculs, qu'exécuter des processus.

La spécification du système de règles en vue d'une réalisation informatique exige la plus grande rigueur méthodologique car l'explicitation du processus doit être complète et cohérente. Elle demande donc un haut degré d'étude et de compréhension du phénomène linguistique.

Confrontation des hypothèses au corpus

La confrontation au corpus est une épreuve de vérité, c'est celle qui va permettre de vérifier si les hypothèses modélisées par le programme sont valides.

Cette confrontation passe par la *sélection d'un corpus de mise au point*, c'est-à-dire d'un corpus de travail. Contrairement à la phase d'élaboration des hypothèses dans laquelle l'objet d'étude était le matériau aussi bien au format papier qu'au format électronique, la confrontation requiert, elle, l'utilisation d'un corpus électronique.

Après sélection d'un corpus de mise au point, il est alors possible d'*exécuter le programme* informatique sur ce corpus. L'application du processus informatique est objective, systématique, régulière et en plus reproductible à l'identique, elle ne peut être remise en cause. La rigueur de la spécification du système de règles requise par l'informatique, alliée à son application automatique sur le corpus produit, par conséquent, exactement les instances des hypothèses présentes dans corpus et répondant aux critères définis par le système de règles.

L'*analyse des résultats* va alors avoir pour objectif la mise au point du système de règles. Il s'agit tout d'abord d'une mise au point générale car l'implantation d'un système de règles est toujours très délicate. L'analyse passe par le calcul d'un écart entre les instances générées par la machine et les instances attendues. Ce calcul est réalisé automatiquement, s'il existe une version annotée du corpus, manuellement dans le cas contraire. Une mesure de la couverture du système pour le corpus de mise au point est alors obtenue.

L'analyse des résultats conduit alors à un *affinement du système de règles*, c'est-à-dire de la méthode de construction des instances des hypothèses, pour le corpus de mise au point. Cet affinement s'effectue par la modification du système de règles. On peut d'ailleurs parler de spécialisation car l'objectif est alors de pousser le modèle à son maximum, de tester sa souplesse, sa capacité à coller aux données, à traiter les cas particuliers. Cette spécialisation doit cependant rester maîtrisée car l'écueil de la «sur-spécialisation» n'est jamais évité d'emblée ; une telle spécialisation risque en effet de conduire à une dégradation importante des performances de l'analyseur lors du passage à un corpus différent. C'est pour cette raison que d'une part, nous choisissons des corpus de mise au point variés, permettant ainsi la vérification de la légitimité des spécialisations réalisées et que d'autre part, une prise de recul est toujours effectuée.

Prise de recul sur l'objet

La confrontation au corpus est un travail de longue haleine. Les hypothèses sont poussées à leur ultime limite. Une couverture maximale du corpus est recherchée. Sous l'effet de diverses pressions, il est souvent nécessaire de se détacher du corpus, de prendre du recul par rapport à ce travail parfois trop opératoire.

La prise de recul peut survenir lors de la constatation d'un blocage conceptuel du système, généralement dû à un concept trop rapidement considéré comme définitivement acquis, que nous utilisons en tant que loi, et sur lequel nous bâtissons nos hypothèses. Des hypothèses auxquelles nous tenons car elles ont requis des recherches et des développements importants et se sont révélées jusqu'à présent satisfaisantes. Un concept et des hypothèses pour lesquels certains sont prêts à refuser la réalité afin que le corpus se plie tout entier à une pseudo-loi... un corpus qui finalement résiste, s'impose, et pousse vers la révision de ce concept.

Elle se réalise également par la transmission du savoir et notamment par l'enseignement qui exige une synthèse et un exposé clair des recherches effectuées. Dans cette épreuve, toutes les zones laissées volontairement ou involontairement floues resurgissent et sont de nouvelles sources d'informations précieuses pour la poursuite des investigations.

La publication des résultats permet également d'obtenir un jugement scientifique de chercheurs du même domaine. La publication s'apparente à l'enseignement si l'on considère l'effort de synthèse et d'explicitation mais s'accompagne d'un retour critique qui peut parfois donner un nouvel élan aux travaux.

Enfin, des retours très importants sur la théorie peuvent naître de l'utilisation de l'analyseur par d'autres équipes de recherche et de développement. C'est pour cela que nous nous efforçons de mettre à disposition au plus tôt les ressources créées ainsi que les résultats que nous obtenons.

Évaluation et comparaison

L'évaluation d'un analyseur linguistique est une problématique en soi. Ce sujet est d'autant plus sensible qu'il comprend une phase de comparaison entre systèmes. Nous aborderons ultérieurement cette problématique et nous limiterons dans cette section à la démarche que nous nous sommes imposé pour réaliser nos propres évaluations.

La phase d'évaluation consiste à juger la qualité de l'analyseur par rapport à un étalon qualitatif appelé attendu. Ce jugement qualitatif est pondéré par plusieurs facteurs en fonction du principe général suivant : la conception d'un système de règles modélisant les hypothèses de structures et permettant le traitement des phénomènes linguistiques courants ne représente qu'un temps infime, comparé à la quantité d'efforts qu'il faut mettre en œuvre pour traiter les nombreux cas particuliers qui dérogent à ces principes généraux.

Ainsi, l'appréciation globale du système sera un compromis entre (1) la recherche de l'optimalité qualitative des résultats, (2) le maintien de la cohérence du système que l'on pourra se refuser à dégrader par la prise en compte des multiples cas particuliers qui vont complexifier la base de connaissances, tout en n'induisant qu'une très faible amélioration de la qualité des résultats, (3) et l'application dans laquelle sera utilisé l'analyseur linguistique qui ne requiert pas forcément le niveau d'excellence représenté par l'attendu.

Traditionnellement, l'évaluation s'effectue sur un corpus n'ayant pas servi à la mise au point de l'analyseur afin de garantir l'intégrité de la démarche. La démarche cohérente consiste donc à conserver une totale indépendance entre les corpus d'évaluation et les corpus de mise au point.

Cette démarche est confortée par l'argument suivant : la réalisation des attendus, qui se révèle des plus contraignantes. Ces attendus réclament en effet un travail manuel et intellectuel très important car ils doivent être à la fois d'une grande précision et d'une grande régularité, tout en étant de taille suffisamment conséquente pour couvrir un grand nombre de phénomènes linguistiques. Des ressources de qualité de ce type sont par conséquent très rares et leur réutilisabilité est essentielle.

La publication d'une évaluation nécessite la divulgation des paramètres déterminants concernant les sources de connaissances utilisées par l'analyseur ainsi que la qualification et la quantification du corpus d'évaluation. En y ajoutant l'analyse des résultats, principalement des échecs, la matière est présente afin que les résultats soient interprétables et que l'expérience puisse être éventuellement reproduite dans des conditions similaires.

Enfin, il faut noter que la qualité des résultats obtenus par un analyseur est fortement liée à l'objectif des recherches ou des développements pour lesquels il a été créé. Ainsi, une comparaison dans l'absolu de deux analyseurs de même type, n'ayant pas été conçus pour être intégrés dans une même chaîne de traitement, est peu pertinente. Cette comparaison sera également peu pertinente si les conditions d'évaluation ne sont pas strictement identiques.

2.2.3 Critique de la démarche

La réalisation informatique contribue à la validation ou à la réfutation des hypothèses faites sur les structures en nous offrant un nouveau regard sur notre objet : l'objet vu au travers de notre hypothèse. Il faut cependant être conscient qu'on ne peut parler que de validation empirique car même si toutes les instances de la structure conjecturée présentes dans le corpus sont correctement générées, l'exécution ne valide l'hypothèse que pour un corpus particulier, aussi vaste et représentatif soit-il.

Épistémologiquement, il faut en évaluer les conséquences : notre démarche nous permet, par l'ajout d'un nouveau texte au corpus, soit d'accroître la confiance en l'hypothèse, soit d'invalider l'hypothèse en permettant le diagnostic de contre-exemples. Changer de corpus ou bien ajouter un texte est une des méthodes que nous utilisons pour accroître la confiance en une hypothèse, ou bien invalider une hypothèse. Il en existe d'autres telles que soumettre la structure au regard d'un grammairien, ou bien utiliser l'unité définie en tant que concept pour créer une unité d'ordre supérieur.

2.2.4 De l'hypothèse au concept

L'exposé de la démarche scientifique nous a permis de comprendre la méthode utilisée pour la formation et la validation des hypothèses de structures. Pendant toute leur étude, les structures conjecturées ne sont toujours considérées qu'en tant qu'hypothèses. Nous ne parlons alors pas de concepts.

Ces hypothèses ne deviennent concepts que lorsqu'elles sont arrivées à un degré de maturité suffisant pour être utilisées lors de la définition d'unités d'ordre supérieur. Étant en cours de construction, ces nouvelles unités sont donc à leur tour appelées hypothèses.

Ainsi se construit progressivement la théorie de l'objet, une théorie hypothétique confrontée en permanence au corpus et dans laquelle l'ordinateur prend une place prépondérante.

2.2.5 Regards antagonistes sur l'objet

En abordant nos recherches en linguistique sur corpus avec un regard d'informaticien-linguiste, nous avons été confronté à des situations d'échanges difficiles avec des linguistes ; parfois, l'incompréhension réciproque était telle que certains échanges ont même pu être qualifiés d'échecs dialogiques. L'analyse de ces situations nous a permis de mettre en évidence un des facteurs à l'origine de ces impasses, à savoir les regards antagonistes que nous portions sur l'objet pour le théoriser.

Contrairement à la plupart des linguistes, nous sommes en permanence hanté par la problématique du calculable, pendant la construction de notre théorie. Cet aspect est crucial au regard de notre démarche car si nous ne savons pas comment calculer, c'est-à-dire comment établir le système de règles (cf., Élaboration d'un système de règles, page 13), alors nous ne pouvons valider. Cette obsession du calculable a rendu nos discussions parfois très difficiles au regard de certains points de vue que des linguistes avaient adopté et qui, vis-à-vis de notre démarche, étaient difficiles à prendre en compte car, d'une part, nous n'arrivions pas à établir des liens conduisant vers notre intérêt, le calculable, et que, d'autre part, eux-mêmes ne nous donnaient pas les moyens d'y parvenir puisque la question du calculable était en dehors de leurs préoccupations.

2.3 Utilisation de l'ordinateur

L'introduction de l'ordinateur apporte une nouvelle dimension à la recherche en linguistique classique. L'usage intensif qu'il en est fait dans l'approche orientée corpus est, de plus, tout à fait particulier. Il nous paraît donc intéressant de préciser les différents niveaux d'intervention de cet outil dans le cas particulier de notre recherche en syntaxe.

2.3.1 L'ordinateur, un instrument d'observation

Dans notre recherche, une des utilisations de l'ordinateur est l'observation. Bien sûr, sa puissance de calcul autorise avec des temps d'exécution relativement courts, une investigation sur une très grande quantité de données. Mais, le principal attrait ne se situe pas là.

L'intérêt de l'observation via l'ordinateur réside dans la systématicité de l'application automatique d'un processus à un corpus. Cette systématicité fournit une observation objective, régulière et reproductible à l'identique ; ce que l'homme ne peut pas faire.

On notera cependant que, manuelle ou automatique, une observation est toujours guidée par l'homme, sous l'influence de concepts intériorisés si bien que l'observation à l'œil, qui ne peut donc être considéré comme «nu», ou l'observation assistée par ordinateur ne feront toujours apparaître que ce que l'homme a délibérément permis de lui faire révéler. L'ordinateur n'est bien qu'un instrument d'observation : l'analyse des résultats produits par l'ordinateur reste à la charge de l'homme.

2.3.2 L'ordinateur, une aide à la théorisation

L'usage de l'outil informatique pour effectuer la confrontation au corpus est caractéristique de l'*approche orientée corpus* dans laquelle nous nous inscrivons. Dans cette perspective, l'ordinateur permet, au travers de la réalisation informatique d'un système de règles, une aide à la validation ou la réfutation d'hypothèses de structures, propos que nous avons abordés lors de la description de l'élaboration d'un système de règles, page 13.

Cet intérêt est majeur car une étude manuelle ne peut être parfaite car, outre sa capacité de traitement limitée, l'homme ne peut avoir la régularité et la systématicité de la machine. Il va donc pouvoir analyser dans les meilleures conditions la portée de ses hypothèses. Reste à sa charge le regard objectif sur les sorties produites par la machine, car l'homme peut tout simplement ne pas voir, ne pas vouloir voir, ou bien voir autre chose que ce qui lui est présenté.

2.3.3 L'ordinateur, de la recherche au développement

Tout au long de l'exposé de notre démarche, nous avons pu constater que la validation empirique des hypothèses conjecturées reposait sur la mise au point d'un analyseur linguistique qui génère les instances de cette structure dans le corpus. Il faut bien entendu ne pas être dupe et penser que l'analyseur linguistique final, intégrable dans un système de traitement des langues, puisse être ce même analyseur. Ce serait en effet mélanger les missions de la recherche et du développement.

Le chercheur travaille dans un atelier en perpétuelle évolution, contenant souvent de nombreux archaïsmes. Conscient des imperfections de notre chaîne d'études, nous n'avons souvent d'autre solution que de continuer à évoluer car nous ne pouvons redévelopper en permanence l'ensemble des outils à chaque détection d'imperfection. Nous en prenons simplement note et continuons à agir en en tenant compte. Parfois, l'inadéquation de l'architecture de notre atelier expérimental devient telle que nous sommes obligé de l'abandonner et d'en reconstruire un nouveau. Ce n'est qu'à ce moment là que nous corrigeons les imperfections préalablement rencontrées, et que nous intégrons certaines composantes manquantes.

De même, le chercheur conscient des problèmes de complexité cherchera à concevoir des processus efficaces, mais il n'aura pas le souci du développeur qui lui cherchera à optimiser toutes les procédures du logiciel.

2.3.4 Le paradoxe de l'approche orientée corpus

Dans les sections précédentes, nous avons brièvement évoqué la puissance de calcul de l'ordinateur et donc l'intérêt d'un tel instrument pour traiter des corpus de grande taille.

Il est cependant fondamental de noter que l'analyse de grosses quantités de données pose un nouveau problème : la vérification et l'évaluation des résultats produits par la machine. En effet, l'analyse des sorties reste encore bien souvent manuelle, semi-automatique dans le meilleur des cas, et s'effectue sur des masses de données proportionnelles à celles qui ont été données à analyser.

Nous aboutissons à une situation un peu paradoxale dans laquelle l'efficacité de l'ordinateur n'allège en rien la charge de travail du chercheur puisqu'il génère des sorties de tailles proportionnelles qu'il faudra vérifier et évaluer. La charge de travail en temps passera par exemple par la réalisation très délicate d'un attendu de corpus d'évaluation et dans l'analyse des causes de tous les écarts avec cet attendu.

Deuxième partie

Excursion dans le multilinguisme

Chapitre 1

Le diagnostic de langue sur documents multilingues

1.1 Contexte de l'étude

Dans cette partie, nous allons développer la manière selon laquelle s'est construite la prise de conscience du caractère multilingue de notre objet, et de la nécessité de le traiter comme tel. Nous allons commencer la présentation de nos travaux par l'exposé de notre expérience acquise sur le diagnostic de langue. Cette expérience fut enrichissante à plusieurs titres. En tant que première recherche sur corpus, elle nous a permis de nous familiariser avec leur manipulation et de prendre conscience de leurs propriétés élémentaires. En tant que première recherche dans un cadre multilingue, elle nous a permis de travailler sur l'émergence de propriétés communes à toute une famille de langues et sur l'identification de celles spécifiques à chacune d'elles. C'est en fait à partir de cette étude que furent érigées les fondations de notre méthodologie d'étude d'un corpus, méthodologie qui fut par la suite appliquée et bien sûr affinée dans nos recherches en syntaxe.

Le diagnostic de langue, ce n'est pas uniquement un premier contact avec la recherche en linguistique informatique. C'est également une expérience d'intégration verticale passionnante, couvrant en amont tous les aspects de la recherche, et débouchant sur toutes les exigences du développement d'un produit fini lors de son incorporation sous forme de module dans BrailleSurf, un logiciel de navigation sur internet développé au sein de l'Unité INSERM 483 et destiné à des utilisateurs handicapés visuels. Dans ce logiciel, le module réalise le diagnostic automatique de la langue d'un document, permettant alors une prise en compte automatique de celle-ci dans la synthèse vocale, processus de simulation d'une voix humaine à partir d'un document électronique écrit.

1.2 Présentation de notre problématique

Le diagnostic de langue est un prétraitement nécessaire à toute application relevant du traitement des langues et réclamant la langue d'un énoncé afin de l'analyser correctement. C'est le cas de la majorité d'entre elles car la plupart des algorithmes sur lesquels elles reposent, ne sont conçus que pour traiter des énoncés monolingues. De fait, ces algorithmes se trouvent soit dédiés au traitement d'une seule langue particulière, soit paramétrés par les ressources linguistiques (e.g., lexiques, grammaires) relatives à la langue de l'énoncé.

Une application doit, par conséquent, être capable de détecter les changements de langue au sein d'un document afin, d'une part, de ne soumettre à ces algorithmes que des énoncés monolingues pour lesquels on dispose des ressources linguistiques requises et, d'autre part, de gérer l'activation de ces ressources avant l'exécution de ces algorithmes. Le diagnostic de langue va participer à cette tâche en découpant un document en unités monolingues étiquetées par le nom de la langue dans laquelle elles sont écrites.

La nécessité d'un outil de diagnostic de langues apparaît au moment de l'explosion de la recherche documentaire et notamment de l'indexation automatique de documents. Le diagnostic a alors pour objectif la classification des documents en fonction de leur langue. Il constitue un prétraitement minimal essentiel à toute application ayant pour objectif l'analyse automatique de documents provenant du monde entier.

Les publications auxquelles nous avons eu accès et proposant une solution à ce problème considèrent le document comme monolingue ou majoritairement monolingue puisqu'ils cherchent à identifier *la* langue du document. Nous avons choisi d'ouvrir une nouvelle voie en menant une étude prenant en compte le caractère naturellement multilingue des documents. L'objectif de nos recherches est donc d'identifier les différents changements de langue survenant au sein d'un document et d'associer à chaque passage la langue associée. Il s'agit plus généralement de révéler la structure multilingue du document.

1.3 Diagnostic de langue et documents multilingues

1.3.1 De la nature multilingue du document

L'emploi de plusieurs langues dans un même document est, dans les faits, très fréquent : il se manifeste couramment par l'introduction de citations ou bien, par la présentation d'une même information traduite en différentes langues, la présence conjointe de la langue maternelle du rédacteur et de l'anglais constituant, pourrait-on dire, l'archétype même de cette situation. Il suffit pour s'en convaincre d'observer les résumés bilingues des articles scientifiques ou la page d'accueil bilingue de certains sites internet, l'ultime objectif visant à toucher un public toujours plus large.

Quant aux documents non plus bilingues, mais naturellement plurilingues dont il convient de ne pas négliger la quantité produite, nous nous reporterons aux manuels d'utilisation d'appareils électroniques ou aux documents officiels européens pour en trouver exemples. Avec une production certes moindre, cette approche est partagée par certains auteurs poursuivant un objectif cette fois-ci pédagogique, à savoir l'apprentissage des langues.

La structure de ces documents est comparable à celle des pages d'accueil des sites internet que nous mentionnions dans le premier paragraphe. Elle reflète en effet la mise en parallèle d'un même texte en plusieurs langues. Cette structure se concrétise d'ailleurs souvent par l'adoption d'une disposition en colonnes. Il est cependant nécessaire de la différencier de celle des documents dans lesquels sont introduits par exemple des citations, phénomène que nous qualifierons, *a distinguo*, d'alternance.

Au regard du caractère multilingue de cette quantité de texte sans cesse grandissante, le document, pour être analysé, ne peut plus être aveuglément fourni aux algorithmes n'acceptant que des énoncés monolingues. Pour les applications en faisant usage, le document doit alors être décomposé en unités monolingues. Le seul diagnostic de la langue du document n'est alors souvent plus suffisant pour arriver à cette fin. Le recours à des unités plus fines est incontournable pour le diagnostic de langue, unités qui mettront à jour les différents changements de langue intervenant dans le document et auxquels seront associés les noms des langues s'y relatant.

Dans ce cadre, le diagnostic de langue peut avoir l'ambition de révéler la structure multilingue des documents, au même titre que d'autres outils permettent de révéler leur structure logique, c'est-à-dire leur découpage en chapitres, sections, paragraphes, phrases... Ces deux structures peuvent d'ailleurs être considérées complémentairement, la structure multilingue paramétrant la structure logique du document en associant à chaque unité le nom de sa langue de rédaction.

1.3.2 Les langues du document

Dans cette section, nous avons souhaité faire un aparté autour de la notion de «langue du document», telle qu'elle fut entendue dans les approches du diagnostic de langue antérieures à nos travaux. Une telle discussion pourrait sembler peu substantielle au jugé du caractère intuitif de la réponse. Elle porte cependant à réflexion au regard de la démarche qui fut communément adoptée par la communauté s'étant penchée jusqu'alors sur le diagnostic.

Les différentes approches du diagnostic de langue se réclamant comme identificatrices de la langue d'un document concourent toutes à l'identification de la langue la plus employée dans le document. Concrètement, un document comportant majoritairement du texte anglais est identifié comme document de langue anglaise. Cette assimilation de «langue la plus employée» à «langue du document» est pernicieuse. En effet, il est tout à fait envisageable de rencontrer un texte français citant massivement des écrits anglais, le document n'en restant pas moins pour autant de langue française.

La langue du document ne peut donc être identifiée comme étant la langue la plus employée. En ce qui concerne les documents avec citations, ou bien les articles scientifiques avec un résumé bilingue, notre dernière illustration nous fournit implicitement une réponse : la langue du document est celle autour de laquelle s'articule la structure logique du document ; la quantité de lignes écrites dans cette langue en est indépendante. Notons cette corrélation entre les structures logique et multilingue du document, corrélation que nous avions préalablement évoquée et dont nous serons amené à reparler par la suite.

Les précédents paragraphes avaient pour objet la révision de la définition de la notion de «langue du document». Cependant, ce n'est plus la définition mais le concept lui-même qui se trouve fortement menacé au regard des documents plurilingues, composés de textes mis en parallèle. En effet, pour ces documents, il ne convient plus de parler de «langue du document» mais de «*langues* du document». On notera d'ailleurs que la structure logique n'est plus monolingue mais plurilingue.

Ce serait néanmoins un tort de déconsidérer la démarche visant à identifier la langue la plus utilisée d'un document, car même si la qualification d'outil de diagnostic de langue du document devient un peu abusive et qu'elle ne permet pas une couverture des documents plurilingues, il ne faut pas oublier que cette offre répond tout de même à une demande, certes peu exigeante sur la précision, mais bien réelle. Cette demande émane principalement de la recherche documentaire. En outre, dans la pratique, cette démarche s'avère être un bon compromis entre rapidité de développement de l'outil, efficacité de l'exécution, et nature des documents dans lesquels on constate que la langue de rédaction reste majoritaire.

1.3.3 Les nouvelles applications du diagnostic

Nous avons vu en section 1.2 page 26, qu'un prétraitement minimal des documents était essentiel lorsque l'application a pour objectif l'analyse automatique de documents provenant du monde entier. C'est le cas des applications liées à la recherche documentaire telles que l'indexation automatique de documents. Dans cette optique, le diagnostic est utilisé comme outil de classification et d'indexation, sa tâche étant restreinte à l'identification de la langue de rédaction du document.

Lorsque le diagnostic est entendu comme processus de construction de la structure multilingue du document, sa portée s'étend alors bien au delà du cadre de la recherche documentaire : il touche toute la communauté scientifique intéressée par le traitement de l'écrit. Son utilisation principale se trouve aujourd'hui centrée autour de la synthèse vocale et de l'analyse syntaxique. Cependant, étant nécessaire dès les traitements de l'écrit les plus primaires, il s'impose progressivement dans toutes les applications qui les utilisent, des plus simples au plus complexes.

Le diagnostic de langue est utilisable pour permettre à une application de sélectionner au moment approprié les connaissances linguistiques à appliquer à une partie d'énoncé. Il peut être également vu comme traitement préventif afin d'éviter à un analyseur de tenter l'analyse d'un énoncé, ou d'une partie d'énoncé, écrit dans une langue dont il ne possède pas les ressources.

Nous avons jusqu'à présent évoqué son utilisation dans des systèmes de traitement des langues mais les applications du diagnostic ne se limitent pas à ces seuls systèmes. L'utilisation du diagnostic est aussi pertinente dans une recherche en linguistique sur corpus, cela afin d'étudier des phénomènes relatifs à une langue particulière, tout en évitant les parasites générés par l'éventuelle présence d'autres langues.

Même s'il fut intégré par la suite dans une application en front d'un module de synthèse vocale afin d'y réaliser l'aiguillage de la sélection des connaissances linguistiques, c'est historiquement, d'abord en tant que système préventif pour que notre analyseur syntaxique du français cesse de générer des analyses incohérentes sur des énoncés en anglais, puis en tant qu'outil de recherche sur la syntaxe du français qu'il fut utilisé.

1.4 La structure multilingue du document

Le calcul de la structure multilingue du document n'est pas sans véhiculer un certain nombre de problèmes qui lui sont propres et intrinsèques. Nous avons évoqué dans les sections précédentes la nécessité de repérer des changements de langue. Cela implique à la fois la recherche d'un consensus sur la définition d'une unité monolingue minimale correspondant à un changement de langue pertinent, et la capacité à déterminer la langue d'une unité monolingue qui pourra être de taille réduite, probablement d'ailleurs toujours inférieure à celle du document.

Avant de définir ces unités ainsi que le processus permettant leur construction, il convient de présenter au préalable la structure multilingue à laquelle nous souhaitons aboutir.

1.4.1 Présentation de la structure

Le terme «structure multilingue» a été utilisé à maintes reprises sans qu'il nous fût incontournable de le définir précisément. En effet, la discussion restait jusqu'alors très générale et cette omission volontaire ne perturbait guère la compréhension de nos propos. Avant de poursuivre plus en avant le descriptif de nos recherches et d'aborder une partie plus technique, c'est-à-dire, le processus de mise en évidence de cette structure, il nous paraît désormais opportun de lever le voile sur ce qui constitue notre objectif calculatoire.

Nous appelons structure multilingue la structure qui caractérise les différents changements de langue intervenant au sein d'un document. Si l'on s'en tient à cette stricte définition et pour peu que l'on accepte de s'abstraire du découpage logique du document, la structure multilingue se matérialise par une contiguïté de segments monolingues étiquetés par le nom de leur langue, et telle que deux segments contigus ne portent pas la même étiquette.

Pour être complet, il nous reste à définir ce que nous entendons par segment monolingue, du point de vue de notre structure multilingue. Un segment monolingue est un segment dont la structure est organisée autour de mots structurants écrits dans une seule et même langue. Les mots structurants sont tout simplement les traditionnels mots grammaticaux, ou mots vides. On compte parmi eux, les prépositions, les conjonctions, les déterminants, les pronoms, les adverbes en liste close, les auxiliaires.

Par cette définition, nous pointons la nécessité de distinguer, à l'aide d'une bipartition, les mots d'une langue : les mots structurants d'un côté, les mots lexicaux de l'autre. En faisant assumer, pour le moment, la langue du segment aux seuls premiers, nous conférons à quelques centaines de mots une importance considérable, et désignons par là même ce qui constitue un des principaux écueils du diagnostic, à savoir, l'identification de la langue en l'absence de mots structurants.

Pour conclure cette section, il convient de présenter l'attitude que nous avons souhaité adopter quant au statut des mots d'emprunt. Cette attitude détermine en partie la structure multilingue visée. L'emprunt de mots à d'autres langues fait partie de l'évolution naturelle des langues. Vis-à-vis de notre structure multilingue se pose la question du statut de ces mots : les considère-t-on comme des mots étrangers, la structure devant alors les révéler, ou bien, les assimile-t-on a des mots de la langue d'accueil, la structure en faisant alors abstraction ? Privilégiant la dynamique de l'évolution des langues, et ne voulant surtout pas entrer dans un débat qui pourrait vite se transformer en une régression sans fin sur l'origine de chaque mot, nous avons délibérément écarté le problème de la langue du mot isolé au profit de l'étude de la langue du mot dans son contexte. Ainsi, c'est le contexte qui va déterminer la langue du mot et non le mot à lui seul.

1.4.2 Observation en contexte des segments monolingues

L'observation montre que le changement de langues entre deux segments monolingues contigus, tels que nous les avons définis dans la section précédente, est formellement marqué. Différentes marques coexistent, nous en avons repéré deux catégories :

1° celles qui modifient l'aspect visuel du texte en jouant sur ses attributs : il s'agit bien souvent de l'italique, parfois de la couleur ;

2° celles qui sont insérées dans le texte sous forme de ponctuations véhiculant l'idée d'une rupture : les points de fin de phrase, les parenthèses, les guillemets, les deux points, le tiret long sont autant de marques potentiellement introductrices d'un changement de langue. Dans certains textes multilingues, un caractère particulier fait office de séparateur de langues : la barre oblique, appelée aussi *slash* et notée «/».

Aucune de ces marques n'est spécifique au changement de langue mais leur présence est essentielle car elles segmentent naturellement le texte en unités. Ces unités sont, par construction, toutes monolingues et toutes de taille inférieure ou égale aux segments monolingues de notre structure multilingue. Nous les appelons «unités minimales». Elles nous permettent de proposer un principe de construction de la structure multilingue.

1.4.3 Principe de calcul de la structure

Le principe de calcul que nous proposons est le suivant. Il consiste à tronçonner les documents électroniques en unités minimales, en suivant séquentiellement toutes les marques formelles présentées dans la section précédente. Conformément à cette découpe, la structure obtenue est une liste de segments monolingues contigus. Cette structure temporaire est très proche de la structure multilingue attendue. Mais, contrairement à cette dernière, qui est telle que deux segments contigus sont toujours de langues différentes, notre structure d'unités minimales pourra, elle, contenir plusieurs unités contiguës de même langue car le tronçonnage s'effectue sur des marques *potentiellement* introductrices de changement de langues.

À chaque unité minimale nouvellement produite, il reste à vérifier que sa langue est identique à celle de la précédente unité. Si c'est le cas, nous les concaténons. S'il y a changement de langue, hésitation entre plusieurs langues, ou bien non-identification de la langue, un marqueur est introduit pour noter l'information. Le résultat final correspond alors à la structure multilingue escomptée.

La vérification de la langue des unités minimales passe par l'application d'une fonction d'identification sur le segment concerné. Cette fonction doit être efficace car appliquée à des unités de petite taille. Elle se doit d'intégrer les contraintes qui avaient été antérieurement définies dans le cadre du diagnostic de la langue du document. Ainsi, nous essaierons de tendre vers une certaine robustesse du processus. Par robustesse, il faut entendre une capacité à gérer des documents provenant de sources variées, parfois peu structurées, telles que le courrier électronique, voire même bruitées, le processus ne devant pas être alors trop sensible à la présence de fautes de frappe, d'orthographe, éventuellement de reconnaissance optique de caractères si le document électronique résulte d'une numérisation. La conception de cette fonction fera l'objet du chapitre suivant.

1.5 Conclusion

Notre recherche sur la structure multilingue des documents a commencé lorsque nous nous sommes aperçu que la plupart des documents étaient naturellement multilingues. Nous avons alors démarré des travaux qui permettraient un traitement automatique de tels documents. Nous sommes arrivé à la conclusion que seule la mise à jour d'une structure multilingue solutionnerait le problème. L'étude de l'état de l'art portant sur le diagnostic de langue a révélé que les recherches s'étaient jusqu'alors restreintes à l'identification de la langue d'énoncés monolingues, sans jamais étudier l'obtention de tels énoncés dans un cadre multilingue.

Après une étude sur documents, nous avons opté pour la définition d'une structure multilingue constituée d'une succession d'unités monolingues, telles que deux unités contiguës ne portent pas la même étiquette. Nous avons caractérisé ces unités monolingues en mettant en évidence des marques de changement de langues. Nous avons ensuite conçu un algorithme de production de la structure multilingue, basé sur la concaténation de segments monolingues minimaux sur lesquels nous procédons à une identification de la langue.

Le chapitre suivant est dédié à l'identification de la langue d'un énoncé monolingue. Il va valider à la fois la pertinence des marques de changement de langues et le bien-fondé de notre méthode de calcul de la structure multilingue. Avant d'y arriver, nous souhaitons revenir sur les relations entre la structure multilingue et la structure logique du document qui, bien qu'annoncées, n'ont pas été prises en compte.

Nous avons très tôt mentionné les relations étroites qu'entretenaient les structures logique et multilingue du document. Bien sûr, la tentation de prendre pour segments monolingues des unités logiques fut grande. Deux classes d'arguments le justifiaient :

- la première, de nature structurelle, tendrait à faire valoir une certaine prédominance de la structure logique et à exprimer le caractère multilingue comme simple paramètre de celle-ci ;

- la seconde, de nature opérationnelle, tendrait à défendre la position selon laquelle, à l'intérieur d'un même niveau logique, les unités du sous-niveau immédiat ont de grandes chances d'être de la même langue et que cette information pourrait être prise en compte pendant le diagnostic. On constate en effet que, dans un paragraphe, les phrases sont généralement écrites dans la même langue.

La tendance systématique à opter pour la vision hiérarchique d'une structure est courante en informatique mais pas forcément saine. Même s'il est très séduisant de les amalgamer, nos recherches sur le document n'étaient pas assez avancées pour prétendre fondre la structure logique et la structure multilingue. Il était donc préférable de perpétuer l'idée d'une séparation nette entre les deux structures, de considérer la structure multilingue comme une structure plane, et d'attendre d'ultérieurs développements des études sur la structure logique des documents multilingues pour s'y aventurer.

Ce point de vue reste d'ailleurs compatible avec une fusion *a posteriori* des deux structures, la structure multilingue paramétrant par exemple la structure logique. Par ailleurs, si l'on souhaitait réellement tirer parti de la structure logique pour construire la structure multilingue, le découpage que nous proposons ne l'empêcherait en rien. Cela pourrait se traduire par une séquentialisation des processus de construction des deux structures, ou bien par leur exécution simultanée et interactive, permettant ainsi à chaque processus de bénéficier de connaissances sur la structure en cours de construction par l'autre.

Nous souhaitons conclure par une hypothèse un peu audacieuse. Dans la première partie de notre étude, nous avons avancé que la (ou les) langue du document étaient celle (ou celles) de sa structure logique. Dans la seconde partie, nous avons montré que la langue d'une phrase, était définie par celle des mots et morphèmes grammaticaux, c'est-à-dire par les éléments qui articulent la structure syntaxique. Si l'on accepte ces résultats, il n'y alors qu'un pas à franchir pour considérer la structure logique comme une structure syntaxique de haut niveau.

Chapitre 2

Identification de la langue d'un énoncé monolingue

L'identification de la langue telle que nous allons l'aborder consiste à associer une langue à une unité textuelle supposée monolingue. Au cours d'une présentation critique de diverses approches, nous allons introduire les caractéristiques souhaitables d'une telle fonction ainsi que les objectifs scientifiques recherchés au travers de sa définition.

2.1 Étude des approches classiques

2.1.1 L'approche lexicale

L'identification de la langue peut se concevoir très simplement. Au travers d'une approche purement lexicale, par exemple. Si l'on définit la langue d'un énoncé par la langue des mots qui le composent, le simple recours à des lexiques de mots de chaque langue est suffisant pour identifier la langue : il suffit de reconnaître la langue du segment comme étant celle pour laquelle le lexique contient tous les mots.

Cette approche intuitive ignore quelques propriétés fondamentales des lexiques et notamment leur incomplétude. Plusieurs sources sont à l'origine de leur incomplétude : de nombreux termes techniques, internes à certaines spécialités, ne se trouvent jamais dans les lexiques généraux ; certains mots tels que les noms propres et les sigles ne sont pas exhaustivement recensables ; l'évolution permanente des langues mène à la création de nouveaux mots et à l'emprunt de mots à d'autres langues. Elle suppose par ailleurs l'absence des diverses fautes dans les énoncés, fautes d'orthographe, fautes de frappe, qui sont malheureusement trop courantes et qui perturbent la reconnaissance de mots.

Partant de ce constat, une approche lexicale évoluée est envisageable. Elle consiste non plus à identifier la langue du segment comme étant celle pour laquelle le lexique contient tous les mots, mais comme étant celle pour laquelle le lexique contient le plus de mots. L'approche devient alors un peu plus tolérante aux mots absents du lexique ainsi qu'aux diverses fautes des énoncés.

Même dans sa version évoluée, l'approche purement lexicale souffre de plusieurs problèmes intrinsèques. Aujourd'hui, obtenir des ressources lexicales électroniques n'est certes pas très difficile pour les langues qui font l'objet de traitements automatiques depuis bien des années. Entrent dans ce cadre les principales langues européennes. Le problème est bien différent pour les langues un peu plus exotiques. Par ailleurs, la faible tolérance à l'apparition de nouveaux mots et aux fautes n'encourage pas l'utilisation de cette approche pour l'identification de la langue d'énoncés courts.

2.1.2 Les approches statistique et probabiliste

Les approches statistique et probabiliste pallient le problème de l'acquisition des connaissances linguistiques auquel est confrontée l'approche lexicale, en utilisant des ressources construites automatiquement à partir d'un corpus textuel représentatif de la langue. Les connaissances linguistiques étant extraites à partir du texte, leur acquisition pour des langues peu courantes n'est pas un problème, elle requiert uniquement une quantité suffisante de matériau pour l'extraction.

Ces approches se font fort, au cours de l'extraction, de capturer au moyen de modèles statistiques ou probabilistes certaines régularités formelles des langues et de leur associer une fréquence ou probabilité d'apparition. Ce sont ces régularités, qui, une fois stockées, servent de connaissances linguistiques pour identifier la langue des textes proposés au diagnostiqueur. L'identification consiste à calculer, en fonction des régularités extraites, la probabilité d'un énoncé à appartenir aux différentes langues. Deux principaux types de régularités sont couramment exploités :

1° les mots les plus fréquents ;

2° les séquences de n caractères les plus fréquentes — ces séquences sont appelées n-grammes de caractères ou plus simplement n-grammes.

Les tableaux 2.1 page ci-contre et 2.2 page 42 donnent un aperçu des mots et des trigrammes les plus fréquents pour le français, l'anglais, l'espagnol et l'allemand. Une rapide analyse comparative permet de constater que le modèle trigramme capture la connaissance des mots les plus fréquents. On y reconnaît également quelques suffixes. De plus, il n'exige pas une segmentation préalable du texte en mots, caractéristique intéressante pour le

Français	Anglais	Espagnol	Allemand
de	the	de	und
la	of	y	der
et	and	la	die
les	to	que	ich
à	in	el	zu
le	a	los	den
des	that	en	in
que	is	del	das
en	as	por	des
du	with	las	mit

TAB. 2.1 - *Quelques mots les plus fréquents*

traitement de langues pour lesquelles les frontières entre mots ne sont pas fortement marquées.

Le modèle n-gramme se montre en cela plus puissant que le modèle des mots les plus fréquents et ses performances se dégradent naturellement moins vite sur les énoncés courts qui ne contiennent pas forcément de mots parmi les plus fréquents. Ce sont les seuls arguments vraiment substantiels des plaidoyers en sa faveur. Dans sa version bigramme ou dans sa version trigramme, ce modèle est à la base de la plupart des systèmes d'identification récents, par exemple ceux de CAVNAR et TRENKLE (1994), DUNNING (1994), GREFENSTETTE (1995), SIBUN et REYNAR (1996).

Note: Les modèles statistiques reposent sur le postulat suivant lequel les profils caractéristiques de deux textes d'une même langue (exprimés en liste de trigrammes les plus fréquents ou en liste de mots les plus fréquents) sont toujours proches. Il faudra se reporter aux travaux de ZIPF (1949) et notamment à la fameuse loi qui porte son nom, pour en trouver justification et pleine compréhension (c.f., annexe B page 167). Ses découvertes ont des applications multiples dans le traitement des langues sur corpus, comme l'expliquent de nombreux articles tels que celui de POWERS (1998).

Français	Anglais	Espagnol	Allemand
`es␣`	`␣th`	`␣de`	`en␣`
`␣de`	`the`	`de␣`	`er␣`
`de␣`	`he␣`	`␣y␣`	`ich`
`nt␣`	`␣of`	`os␣`	`␣de`
`ent`	`of␣`	`␣la`	`ch␣`
`␣le`	`␣an`	`el␣`	`ein`
`␣qu`	`ed␣`	`,␣y`	`der`
`le␣`	`nd␣`	`que`	`und`
`les`	`and`	`la␣`	`nd␣`
`que`	`␣in`	`␣a␣`	`␣un`
`e␣d`	`ing`	`␣el`	`ie␣`
`e␣l`	`␣to`	`␣en`	`nde`
`␣la`	`in␣`	`ue␣`	`␣di`
`et␣`	`ng␣`	`␣qu`	`cht`
`la␣`	`er␣`	`␣co`	`die`

le symbole ␣ représente le caractère espace

TAB. 2.2 - *Quelques trigrammes les plus fréquents*

2.1.3 Le principe de l'identification

À notre connaissance, l'idée sous-jacente à toutes les approches du diagnostic de la langue est toujours fondée sur le même principe, que ce diagnostic soit destiné à identifier la langue du document ou, comme dans nos travaux, destiné à calculer la structure multilingue du document. Elle consiste, dans une première phase que nous appelons phase d'acquisition des connaissances linguistiques, à établir pour chaque langue un profil caractéristique qui fera office de référence, puis, au cours d'une seconde phase que nous nommons phase de diagnostic, à rechercher pour chaque énoncé particulier, le profil de référence le plus proche.

Voyons comment s'effectue la mise en œuvre de ce principe dans l'approche proposée par CAVNAR et TRENKLE. Tout d'abord, examinons la phase d'acquisition des connaissances linguistiques. Chaque langue est caractérisée par un *profil trigramme* acquis automatiquement sur un corpus de référence. Un trigramme étant défini comme une séquence de trois caractères consécutifs, le profil trigramme consiste en la liste des contiguïtés de trois caractères les plus fréquentes, triée par ordre décroissant de leur fréquence d'apparition dans le corpus.

Au cours de la phase de diagnostic, lorsqu'un nouveau texte est présenté pour analyse, sa langue est identifiée en trois temps :

1° le profil trigramme du nouveau texte est calculé, comme s'il s'agissait d'un corpus de référence standard ;

2° pour toutes les langues, une distance est calculée entre son profil caractéristique et celui du nouveau texte. Cette distance repose sur la définition d'une mesure de similarité : elle correspond à la somme des écarts de rangs entre chaque trigramme du nouveau profil et ce même trigramme dans le profil de référence, s'il y est présent (un écart maximal est attribué si le trigramme est absent du profil de référence) ;

3° la langue est diagnostiquée. Elle correspond à celle pour laquelle la distance est la plus petite.

La description de ce processus met bien en évidence les deux phases traditionnelles, celle d'acquisition des connaissances linguistiques, une acquisition basée sur un modèle statistique et réalisée automatiquement, et la phase de diagnostic qui consiste à rechercher le profil trigramme de référence le plus proche du profil trigramme de l'énoncé.

2.1.4 Vers une approche plus linguistique

Le modèle trigramme s'est rapidement imposé par sa facilité de mise en œuvre, sa robustesse, sa souplesse quant à l'ajout de nouvelles langues, et bien entendu, par ses performances. Pourtant, ce n'est pas vers cette approche que nous nous sommes orienté, et cela pour deux principes majeurs inhérents à notre démarche scientifique.

> 1° Pour analyser correctement un énoncé, il faut *dominer les connaissances linguistiques* à utiliser.

La connaissance linguistique extraite automatiquement de corpus au travers du modèle trigramme n'est pas faite pour être étudiée, elle est faite pour être appliquée systématiquement. En cela, elle ne se prête pas facilement à la réalisation de contrôle de qualité sérieux et elle ne peut servir de base à une étude linguistique rigoureuse : les profils trigrammes capturent différentes propriétés linguistiques, mais ces propriétés ne sont pas toutes immédiatement identifiables et leur contribution pour l'identification de la langue n'est pas toujours claire.

Illustrons ce constat par quelques cas concrets tirés du profil trigramme français générés à partir de nos propres corpus. On trouve à des rangs discontigus les trigrammes `␣le`, `les`, et `es␣` sans qu'il soit possible de conclure immédiatement que le mot «*les*» est très fréquent et qu'il contribue par conséquent fortement à la haute fréquence de ces trigrammes. Si l'on veut expliquer la présence de `tio`, parmi les trigrammes les plus fréquents, il faut avoir connaissance de la terminaison fréquente «tion» qui n'est pas modélisable par un trigramme. À l'inverse, on note dans le profil les trigrammes `née` et `rée`, qui ne sont en fait qu'issus de la fréquence de la terminaison «ée». On constate qu'un n-gramme plus long, un quadrigramme par exemple, ou bien tout naturellement un mot, aurait parfois été préférable pour modéliser un certain phénomène. Dans d'autres situations, au contraire, nous avons l'intime conviction qu'un simple bigramme aurait été suffisant.

> 2° Dominer les connaissances linguistiques utilisées n'est pas tout. Encore faut-il *en maîtriser l'application*.

Le critère positionnel est en effet essentiel en linguistique. Or, dans l'approche statistique précédemment exposée, le mode d'application est aveugle, c'est-à-dire insensible à la position. Ainsi, dans la méthode d'identification de GREFENSTETTE (1995), la fréquence d'apparition d'un trigramme est utilisée dans des contextes différents de ceux ayant permis son émergence. Il en résulte que la fréquence du trigramme français `ion` va contribuer identiquement à l'identification de la langue des mots «*incarnation*» et «*ioniser*». Sur le premier, l'application de la fréquence associée à `ion` est pertinente car cette fréquence est totalement issue des mots terminés par «tion» : le contexte d'application est identique au contexte ayant permis l'émergence. L'application de la même fréquence en début du second mot n'est par contre pas justifiée : la chaîne de caractères est identique mais elle ne correspond pas au même élément linguistique. Nous venons de décrire le cas d'une terminaison d'un mot lexical appliquée sur le début d'un autre mot lexical. Nous aurions pu établir la même démonstration pour des fréquences de trigrammes issues de mots grammaticaux fréquents (e.g., `des`) et appliquées aveuglément sur des mots lexicaux (e.g., «*dessin*»).

Dans cette méthode et dans plusieurs autres, celle proposée par CAVNAR et TRENKLE (1994) notamment, le poids associé à un trigramme ne permet pas de comprendre à lui-seul le succès de l'identification. L'explication doit être cherchée dans la contribution mutuelle des poids de tous les trigrammes. Cette contribution mutuelle est un complexe jeu d'influence des poids. Elle permet de pallier dans bien des cas les faiblesses du modèle, mais ne s'avère malheureusement pas suffisante pour toutes les compenser, surtout lorsqu'il s'agit d'analyser des énoncés courts. C'est alors que ressortent le manque de précision des connaissances linguistiques et le manque de maîtrise de leur application.

Outre ces problèmes, nous aurions pu évoquer celui de la constitution des corpus représentatifs des différentes langues servant de base à la création des profils de référence. Nous ne l'abordons cependant pas car il s'agit là d'un problème non spécifique à cette approche mais commun à toutes les expériences orientées corpus (PÉRY-WOODLEY, 1995). Le phénomène est cependant amplifié ici puisqu'il s'agit de constitution manuelle de textes en langues étrangères et il est difficile d'estimer comment les facteurs qualitatifs (niveau de langue, nature, ...) et quantitatif de ces corpus d'apprentissage modifient les profils et donc affectent les performances du système ; c'est en l'occurrence un des paramètres de l'évaluation de SIBUN et REYNAR (1996).

2.2 Notre approche de l'identification de la langue

C'est donc par l'identification, la compréhension et la modélisation des phénomènes linguistiques contribuant à l'identification de la langue, puis par la maîtrise de leur application que nous avons souhaité définir notre analyseur, l'efficacité ne devant être qu'une simple résultante. Dès lors, une approche lexicale aussi bien qu'une approche statistique ou probabiliste ne pouvaient constituer le fondement même de notre système, seule une approche linguistique s'y prêtait. Nous allons maintenant présenter les résultats de nos recherches sur les différentes propriétés linguistiques concourant à l'identification de la langue d'un énoncé.

Méthodologiquement, une double alternative s'offrait à nous quant à la formulation de ces propriétés. Nous pouvions chercher à qualifier les langues soit par leurs propriétés intrinsèques, soit par opposition aux autres langues. Ces propriétés pouvaient par ailleurs être définies soient positivement, c'est-à-dire par ce que l'on trouve obligatoirement ou très souvent dans une langue, soit négativement, c'est-à-dire par ce que l'on ne trouve jamais ou alors très rarement.

Nous avons opté pour une qualification des langues par leurs propriétés intrinsèques, plutôt que pour une qualification par opposition à d'autres langues. Cette position est préférable puisqu'elle rend les connaissances linguistiques d'une langue indépendantes de celles des autres langues, ce qui en facilite l'ajout ou le retrait. Quant à la formulation de propriétés positives ou négatives, notre choix s'est pour le moment arrêté sur la première des possibilités, car elle nous permet une validation empirique sur corpus des propriétés linguistiques des langues que nous ne connaissons pas.

2.2.1 Les mots grammaticaux comme discriminants

Dans cette section, nous allons motiver les raisons qui nous ont conduit à choisir les mots grammaticaux comme discriminants pour l'identification de la langue. Nous avons parlé de ces mots (prépositions, conjonctions, déterminants, pronoms, adverbes en liste close, auxiliaires) en section 1.4.1 page 32, les présentant comme porteurs de la langue de l'énoncé.

L'utilisation des mots grammaticaux pour l'identification de la langue se justifie par le fait qu'ils permettent un diagnostic sûr, même lors de l'utilisation de mots d'emprunt ou de néologismes. Il serait en effet préjudiciable d'assigner la langue anglaise à un énoncé français sous prétexte qu'il contient plusieurs mots anglais ou d'origine anglaise. Les mots grammaticaux nous gardent à l'abri de ce risque car les deux phénomènes, mots d'emprunt et néologismes, affectent les mots en catégorie non finie, principalement les noms, les adjectifs, les verbes, jamais les mots grammaticaux.

Les mots grammaticaux sont utilisés pour structurer un énoncé dans toutes les langues. Ils sont propres à chaque langue et relativement différents d'une langue à l'autre. Ils représentent en moyenne 50% des mots d'une phrase dans la plupart des langues et leur présence est incontournable, sous peine de rendre l'énoncé inintelligible. Étant amené à catégoriser des énoncés courts afin de calculer la structure multilingue, disposer d'indices incontournables au sein de cette unité est un atout majeur.

En ce qui concerne la constitution des ressources linguistiques, les mots grammaticaux ont la propriété d'être courts et en nombre limité, ce qui nous permet d'envisager la construction de listes exhaustives. En cela, ils se distinguent des «mots les plus fréquents» qui, eux, s'acquièrent sur corpus, entraînant la problématique de la constitution du corpus d'apprentissage, ainsi que la subjectivité du nombre de mots à considérer comme faisant partie des plus fréquents.

Nous verrons lors de l'évaluation de notre système que l'utilisation des mots grammaticaux permet une catégorisation parfaite des phrases de plus de huit mots. En effet, plus les phrases sont courtes, et plus le nombre de mots grammaticaux diminue. Le nombre d'erreurs à l'identification est négligeable, quasiment nul, nous disposons donc d'une base très solide qu'il convient cependant de compléter pour améliorer la catégorisation des phrases courtes.

2.2.2 L'alphabet

Une vérification très simple à effectuer pour tenter d'améliorer l'identification consiste à tester l'appartenance des lettres d'un mot aux différents alphabets. Sous-entendu, si le mot ne peut s'écrire dans l'alphabet d'une langue alors il n'en fait pas partie.

L'utilisation d'une connaissance si peu élaborée peut être déconcertante. Cependant, il faut reconnaître en elle l'intérêt d'être similaire à celle des mots grammaticaux dans la mesure où elle est bien maîtrisée, listable exhaustivement, et indépendante d'un quelconque apprentissage. Bien sûr, beaucoup d'alphabets ont une base commune importante, dérivée du latin. Mais des signes diacritiques (tels que les accents) les particularisent. C'est en utilisant cette information que nous espérons obtenir une meilleure catégorisation.

En pratique, la vérification d'appartenance des lettres d'un mot à un alphabet n'est pas très convaincante : leur utilisation n'est pas aussi fréquente que nous l'espérions et beaucoup de lettres accentuées s'avèrent présentes dans plusieurs langues, le facteur discriminant s'en trouvant amoindri. Néanmoins, les erreurs d'identification générées par l'utilisation de cet indice sont, comme précédemment, extrêmement limitées. La raison tient au fait que peu de mots contenant des diacritiques sont empruntés d'une langue à l'autre. Nous en concluons que l'alphabet peut être utilisé pour écarter certaines langues, mais rarement pour isoler la bonne. Cette technique est donc à utiliser en combinaison avec d'autres connaissances plutôt que seule.

2.2.3 Les affixes fréquents

L'utilisation de l'alphabet s'étant révélée peu fructueuse, nous avons cherché d'autres moyens d'améliorer l'identification des segments courts. La première méthode consistant à examiner les mots grammaticaux, nous nous sommes concentré sur la caractérisation des mots lexicaux.

Différents contrôles linguistiques, notamment morphologiques, étaient envisageables. La syllabation fut une des pistes explorées, l'idée étant de vérifier la bonne constitution d'un mot en explicitant les transitions acceptables entre syllabes. Cette piste fut abandonnée car elle demandait une acquisition non triviale de connaissances sur toutes les langues. Dans le même ordre d'idées, nous avons étudié la possibilité d'incorporer les affixes fréquents (préfixes et suffixes).

Dans cette optique, nous avons utilisé les profils bi- et trigramme obtenus sur nos corpus afin d'étudier les propriétés des mots lexicaux qui y étaient capturées. Nous savions d'après nos précédentes études que nous trouverions, en tête, des bigrammes et trigrammes de suffixes fréquents, même si ceux-ci n'étaient pas bien constitués. Les profils se sont en fait révélés inutilisables car la présence majoritaire des mots grammaticaux perturbait sérieusement l'émergence de connaissances sur les mots lexicaux. Nous avons donc réitéré le calcul des profils mais en éliminant cette fois-ci tous les mots grammaticaux contenus. Le résultat fut alors beaucoup plus proche de nos attentes. Des morceaux de suffixes ont émergé ainsi que diverses séquences de lettres internes aux mots. Nous avons choisi de nous restreindre à la prise en compte des suffixes (proprement construits) car il fut difficile de distinguer parmi les séquences de lettres celles qui étaient caractéristiques de la langue, de celles qui n'étaient qu'induites par les thématiques des documents contenu dans le corpus d'apprentissage.

La construction des suffixes fut tout d'abord semi-automatique, pour des raisons de rapidité de prototypage. Nous avons par la suite inclus dans nos travaux l'algorithme d'extraction d'affixes mis au point par Hervé DÉJEAN (1998b). Cet algorithme sera présenté dans la section suivante.

Utilisée de manière complémentaire aux autres connaissances, l'inclusion des suffixes a induit une hausse très sensible de la qualité de l'identification des énoncés de cinq à huit mots. Pour les segments de taille moindre, la capacité de désambiguïsation s'améliore également mais au prix d'une légère progression du nombre d'erreurs.

2.3 Mise en œuvre informatique

Un développement informatique complet a été réalisé afin de tester la validité de nos hypothèses. Il poursuit un double objectif : la validation de notre méthode de calcul de la structure multilingue par une décomposition en segments minimaux monolingues (c.f., section 1.4 page 31) et la validation de notre vision de l'identification de la langue (c.f., section 2.2 page 45).

2.3.1 Construction d'un espace expérimental

Nous avons souhaité que la réalisation informatique puisse constituer un véritable laboratoire de simulations, un lieu d'expériences diverses et variées.

Le laboratoire que nous avons bâti doit être perçu comme un espace expérimental. À l'intérieur de cet espace, nous disposons d'un ensemble de substances de base, les connaissances linguistiques et nous sommes doté d'une série d'instruments permettant leur manipulation, les modèles de compatibilité. Chaque expérience se déroule en trois phases :

1° *la préparation* : nous sélectionnons des substances parmi celles qui sont à notre disposition, nous précisons l'instrument permettant leur manipulation, et décrivons le protocole de réalisation, c'est-à-dire le mode et les conditions d'incorporation des substances ;

2° *la réalisation* : nous effectuons l'expérience suivant le protocole sur un matériau ciblé, le corpus de test ;

3° *l'évaluation* : nous étudions les effets de l'application des substances sur le matériau.

2.3.2 Constitution des ressources linguistiques

Dans le cadre de nos travaux, les ressources linguistiques ont été constituées pour quatre langues : il s'agit du français, de l'anglais, de l'espagnol et de l'allemand. Bien entendu, la réalisation informatique permet d'étendre le nombre de langues couvertes à volonté, mais il fut restreint afin de privilégier la démarche et sachant que les langues sélectionnées possédaient une relative proximité, ce qui exige une bonne qualité du système d'identification.

Lors de la construction des ressources et afin de satisfaire les critères d'évolutivité, de souplesse d'utilisation et surtout de qualité, tous les éléments communs à plusieurs langues ont été conservés, tels que le mot «*de*» et la voyelle «*é*» qui se rencontrent à la fois dans les ressources française et espagnole, ou bien le suffixe fréquent «*-tion*» qui est à la fois français et anglais. La taille des données collectées est présentée tableau 2.3.

	mots grammaticaux	lettres alphabétiques	suffixes de mots lexicaux
français	345	84	168
anglais	206	54	57
espagnol	277	65	60
allemand	175	60	68

TAB. 2.3 - *Taille des données collectées*

Constitution des ressources de mots grammaticaux Les ressources de mots grammaticaux furent constituées semi-automatiquement, à partir de sources variées : corpus et lexiques. L'utilisation de lexiques pourrait sembler en contradiction avec les reproches formulés à l'approche lexicale, à savoir la disponibilité des lexiques. Cette option fut choisie pour des raisons de rapidité de développement. Par ailleurs, Hervé DÉJEAN (1998b) propose aujourd'hui une méthode multilingue d'extraction automatique de ces mots sur corpus non annotés.

Une ressource annexe contenant les graphies des numéraux non composés est également disponible. Ces graphies sont séparées de la ressource des mots grammaticaux car nous hésitions sur leur statut. Elles y sont restées avec le temps.

Enfin, deux autres ressources sont issues d'une étude sur la segmentation en mots d'énoncés de langue inconnue. Cette étude fait l'objet du chapitre 3 page 70 qui porte plus généralement sur la problématique de la segmentation dans un cadre multilingue. Les conclusions de cette étude nous ont amené à choisir de ne séparer ni les mots composés, ni les mots élidés car leur segmentation est dépendante de la langue. Cependant, nous souhaitions tout de même profiter des mots grammaticaux commençant ou terminant un «mot» (e.g., «*l'avion*», «*dit-elle*», «*John's*»). C'est ce que nous avons fait en constituant une ressource de mots grammaticaux préfixes et une ressource de mots grammaticaux suffixes.

Constitution des ressources des alphabets La constitution des ressources des lettres de l'alphabet fut manuelle.

Constitution des ressources de suffixes de mots lexicaux Nous allons dès maintenant nous pencher sur la méthode de calcul des suffixes de mots lexicaux, et donc sur l'algorithme de découverte des morphèmes d'une langue, mis au point par Hervé DÉJEAN (1998b) et dérivé de celui de Zellig HARRIS (1955). Cet algorithme travaille sur la liste des mots d'un corpus et fonctionne en deux étapes principales :

La première exploite la diversité des lettres précédant une séquence suffixe donnée. Si le nombre de lettres distinctes dépasse un certain seuil, en pratique la moitié des lettres de l'alphabet, alors nous sommes en présence d'une frontière morphémique. Concrètement, dans notre corpus français, seules sept lettres précèdent la lettre terminale «*r*» : «*a*», «*e*», «*i*», «*o*», «*u*», «*ï*» et «*û*». La séquence «*r*» n'est donc pas un morphème. Par contre, vingt lettres précèdent la séquence «*er*» : elle est donc déclarée morphème de la langue.

Afin de ne pas extraire de morceaux de morphèmes, l'algorithme étudie la fréquence d'apparition de chacune des lettres précédant la frontière : dans le corpus français, vingt et une lettres précèdent la séquence «*on*», ce qui devrait satisfaire le critère d'acceptabilité. Cependant la lettre «*i*» a une fréquence nettement dominante, ce qui fera préférer la séquence «*ion*» à la séquence «*on*». Par ailleurs, le processus continue sa recherche de morphèmes après une identification réussie, ce qui permet par exemple de générer le morphème «-*ique*» une fois le morphème «-*e*» découvert. Cette première étape ne génère qu'un nombre restreint de morphèmes ; une seconde étape permet d'étendre très fiablement la liste.

La seconde étape dérive de nouveaux morphèmes à partir de ceux générés dans la première étape. L'algorithme cherche alors des séquences préfixes pouvant être complétées par plusieurs morphèmes suffixes déjà calculés : par exemple «*rest*», qui peut être complétée par les suffixes «-*ent*», «-*e*» et «-*er*». On extrait alors du corpus les autres mots commençant par la même séquence préfixe, e.g., les mots «*restons*» et «*restèrent*». Après segmentation, des morphèmes potentiels sont obtenus : «ons» et «èrent». Ils seront déclarés morphèmes si un certain nombre de séquences préfixes différentes confirment leur validité, par exemple «*ajout*» et «*invent*». Cette seconde étape est itérée afin d'incorporer dynamiquement les nouveaux morphèmes et d'étendre ainsi progressivement la liste. L'opération converge en fait très rapidement avant de se stabiliser.

Cet algorithme peut être utilisé en sens inverse pour calculer les préfixes. Il présente deux autres caractéristiques importantes : il ne génère qu'un nombre négligeable de «mauvais» morphèmes (pas plus d'un ou deux par langue) et surtout, il ne génère pas de morphème si la langue n'en possède pas naturellement, le vietnamien par exemple.

2.3.3 Représentation des connaissances

motGram.flx anglais	*suffixLex.flx* allemand	*suffixGram.flx* anglais	*prefixGram.flx* français	*numeraux.flx* espagnol
10 a	6 schaft	10 'll	10 c'	10 billn
10 about	6 liches	10 're	10 d'	10 catorce
10 above	6 licher	10 n't	10 j'	10 cien
10 across	6 lichen	10 'd	10 l'	10 ciento
10 after	6 keiten	10 'm	10 m'	10 cinco
10 against	6 ischen	10 's	10 n'	10 cincuenta
10 all	5 ungen		10 qu'	10 cuarenta

alphabet.flx espagnol
10 [a-zA-ZéÉ]+

TAB. 2.4 - *Extraits de fichiers de ressource linguistique*

Nous avons adopté un format de description des connaissances homogènes pour toutes nos ressources. Cette représentation tient compte du fait que nos connaissances ne décrivent que des caractéristiques internes aux mots.

Les mots grammaticaux, les suffixes de mots lexicaux, les numéraux non composés, les mots grammaticaux préfixes et les mots grammaticaux suffixes ont tous simplement été représentés par leur graphie. Les lettres alphabétiques sont, quant à elles, utilisées au sein d'expressions régulières à appliquer sur un mot pour vérifier que toutes ses lettres appartiennent à l'alphabet.

Les ressources peuvent être vues comme des listes d'associations entre un élément de connaissance (mot, expression régulière, préfixe, suffixe) et un degré de confiance. Le degré de confiance est utilisé dans un cumul géré lors de l'identification de la langue lorsque l'élément de connaissance est compatible avec un mot donné. Ce nombre modélise la confiance que le mot donné puisse appartenir à la langue, relativement à la nature de la connaissance.

Dans notre système, nous avons fixé les degrés de confiance de manière à donner plus d'importance aux ressources exhaustives qu'aux ressources acquises statistiquement, car l'application de ces dernières est beaucoup moins maîtrisable, comme nous le verrons par la suite.

Au travers des extraits réels de fichiers donnés par le tableau 2.4, on remarque que toutes les ressources linguistiques sont déclaratives, ce qui facilite leur gestion.

2.3.4 Les modèles de compatibilité

À chaque ressource doit être associé un modèle de compatibilité nécessaire à son utilisation. Le modèle indique la manière d'utiliser la connaissance contenue dans la ressource afin de vérifier sa compatibilité avec un mot particulier.

Les modèles de compatibilité génériques implémentés entre un mot et une ressource sont :

- `WORD` : recherche dans la ressource d'un élément identique au mot courant ;
- `SUFFIX` : recherche dans la ressource du plus long élément suffixe du mot courant ;
- `PREFIX` : recherche dans la ressource du plus long élément préfixe du mot courant ;
- `REGEX` : recherche dans la ressource de la première expression régulière vérifiant le mot courant.

Aux ressources de mots grammaticaux et de numéraux non composés est associé le modèle `WORD`, à la ressource «alphabet» le modèle `REGEX`, à la ressource de mots grammaticaux préfixes le modèle `PREFIX`, enfin, aux ressources contenant les suffixes de mots lexicaux et contenant les mots grammaticaux suffixes, le modèle `SUFFIX`. En figure 2.1 page suivante, nous montrons à titre d'exemple, la déclaration d'associations entre six ressources représentées par leur nom de fichier, et leur modèle de compatibilité.

2.3.5 Segmentation en unités monolingues

La segmentation en unités monolingues doit nous permettre de valider notre approche du calcul de la structure multilingue. Elle exploite par conséquent la majeure partie des marques visuelles potentiellement introductrices d'un changement de langue observées en section 1.4.2 page 33.

DEF	MotGram	**WORD**	*motGram.flx*
DEF	Nombre	**WORD**	*numeraux.flx*
DEF	Alphabet	**REGEX**	*alphabet.flx*
DEF	PrefixGram	**PREFIX**	*prefixGram.flx*
DEF	SuffixGram	**SUFFIX**	*suffixGram.flx*
DEF	SuffixFreq	**SUFFIX**	*suffixLex.flx*

FIG. 2.1 - *Définition d'associations ressource/modèle de compatibilité*

Ces marques, que constituent les attributs du texte et certaines ponctuations, servent de points d'ancrage aux algorithmes de segmentation : le principe consiste à réaliser une coupe à chaque fois qu'une de ces marques est rencontrée et que le contexte en confirme la validité. Dans notre analyseur, nous n'avons géré que les coupes sur signes de ponctuation car nous souhaitions conserver toute liberté quant au choix du format des textes électroniques traités. Les attributs du texte, tels que la couleur et l'italique, nécessitent en effet un encodage spécifique des documents, ce qui n'aurait fait que compliquer la constitution des corpus.

Ces attributs peuvent cependant être pris en compte très simplement par le segmenteur si les documents électroniques en conservent l'information. Dans un document au format HTML par exemple, la mise en italique et la coloration d'une portion de texte se traduisent par un encadrement de la zone ciblée au moyen d'une balise de début et d'une balise de fin (`<I>` et `</I>` pour l'italique, `<FONT COLOR=...>` et `</FONT>` pour la coloration). Il est alors possible pour le segmenteur de s'appuyer sur ces marques et de réaliser des coupes comme s'il s'agissait de ponctuations.

La segmentation en phrases (ou en segments dans le cas présent) est bien connue en tant qu'opération délicate. Nous ne parlerons donc pas des écueils induits par la présence des traditionnels acronymes, abréviations, noms propres et autres entités qui font qu'entre autres une fin de phrase n'est pas un simple point suivi d'un espace et d'une majuscule. La littérature est abondante sur ce sujet et nous vous invitons à vous y reporter pour saisir la complexité du problème dans toute sa généralité. GREFENSTETTE et TAPANAINEN (1994) exposent clairement les cas non triviaux.

Acceptant un certain nombre de compromis qualitatifs, nous avons produit un segmenteur assez robuste pour traiter du texte tout-venant, le découper proprement, et ainsi nous procurer une base solide pour l'identification de la langue. Ce segmenteur est écrit sous la forme d'un automate dans lequel un mécanisme récursif *ad-hoc* permet le masquage des segments inclus (parenthèses, entre-guillemets, incises) lors de l'identification de la langue du segment qui les contient.

2.3.6 Identification de la langue

Le module d'identification de la langue a été conçu de manière à autoriser différentes simulations. Le contrôle de ce module est composé d'une partie procédurale et d'une partie déclarative :

- la boucle de contrôle consiste à itérer sur les mots d'un segment et à appliquer, sur chacun d'eux, et pour chaque langue, une fonction d'évaluation paramétrée par les ressources linguistiques adéquates. Elle cumule simultanément pour chaque langue le résultat de l'application de la fonction d'évaluation sur chaque mot. La ou les langues assignées au segment sont celles qui obtiennent le meilleur score ;

- la partie déclarative définit la fonction d'évaluation. Cette fonction détermine un poids à associer à chaque mot du segment, selon la combinaison souhaitée des ressources linguistiques qui s'y appliquent. L'application d'une ressource consiste en la vérification de la compatibilité du mot courant et de la connaissance contenue dans la ressource, selon le modèle de compatibilité attaché à la ressource (`WORD`, `PREFIX`, `SUFFIX`, `REGEX`). Lorsqu'un élément de la ressource activée est compatible avec le mot courant, le degré de confiance qui lui est associé est retourné.

Le caractère déclaratif de la fonction d'évaluation permet de définir la structure d'enchaînement des ressources linguistiques à appliquer. Nous pouvons ainsi tester souplement différentes stratégies d'activation de ces ressources et étudier finement la contribution de chaque type de connaissances. Le langage de configuration autorise ou bien l'activation inconditionnelle d'une ressource, ou bien son application conditionnelle en fonction du résultat de l'application d'autres. Il se compose de trois instructions dont la

DEF	MotGram	**WORD**	*motGram.flx*
DEF	Alphabet	**REGEX**	*alphabet.flx*
DEF	SuffixFreq	**SUFFIX**	*suffixLex.flx*

(a) Définition des associations ressource/modèle de compatibilité

```
IF Alphabet {
    IF NOT MotGram
        SuffixFreq
    }
```

(b) Définition de la structure d'enchaînement des ressources

FIG. 2.2 - *Exemple de définition d'une expérience complète*

sémantique est la suivante :

1° instruction → Ressource : vérifier si le mot courant est compatible avec la connaissance de la ressource ;

2° instruction → `IF` [`NOT`] Ressource `ALORS` instruction$_1$: si le mot courant est (ou n'est pas) compatible avec la connaissance de la ressource alors exécuter l'instruction 1;

3° instruction → { instruction$_1$ ... instruction$_n$ } : Exécuter successivement les instructions 1 à n.

L'incrémentation des scores est implicite et s'effectue à chaque fois que le mot courant est compatible avec la connaissance d'une ressource. Le poids alors retourné est celui associé à l'élément de la ressource ayant déterminé la compatibilité avec le mot courant.

Toujours à titre d'exemple, le sens de la structure d'enchaînement 2.2(b) est la suivante : si l'alphabet permet d'écrire le mot courant, alors chercher si un mot grammatical est identique à ce mot. Si non, chercher si un suffixe existe pour ce mot. Ce type de contrôle ne déclenche la recherche des suffixes que sur des mots lexicaux. Il interdit de plus l'application des ressources «MotGram» et «SuffixFreq» si l'alphabet ne permet pas d'écrire le mot.

L'analyse d'un énoncé s'effectue en temps linéaire par rapport à son nombre de mots puisqu'un mot est examiné une fois et une seule. Pour réduire cette complexité, il n'existe que deux voies : soit consulter moins de mots (c'est une modification de la boucle de calcul), soit diminuer le temps de traitement de chaque mot (c'est une modification de la fonction d'évaluation).

2.3.7 Évaluation

Les études sur corpus requièrent un travail de dépouillement toujours très fastidieux, voire laborieux, s'il est effectué manuellement. Afin d'alléger au maximum cette charge, le calcul de statistiques résumant l'activité a été intégré au programme. Par ailleurs, une interface graphique a été développée et permet une visualisation graphique rapide des résultats. Ayant été conçue sous la forme d'un formulaire HTML et d'un programme CGI, cette interface (annexe A.2 page 165) est accessible par internet :

http://www.info.unicaen.fr/~giguet/diagnostic-fr.html

2.4 Expérimentations

2.4.1 Objectifs

Une série d'expérimentations a été conduite de manière à évaluer les capacités de notre système à identifier le français, l'anglais, l'espagnol et l'allemand. Pour chaque expérimentation, nous avons sélectionné deux critères d'évaluation :

1° la capacité des connaissances linguistiques à converger vers l'identification d'une seule langue ;

2° la capacité de ces connaissances à identifier la bonne langue.

2.4.2 Deux expériences

Nous confrontons ici deux expérimentations. La première engage les ressources acquises sans apprentissage : l'alphabet et toutes les ressources de mots grammaticaux. La seconde implique en plus les suffixes de mots lexicaux. Le paramétrage déclaratif des expériences est présenté figure 2.3 page suivante.

2.4.3 Protocole d'évaluation

Afin de simplifier le dépouillement manuel des résultats sur des données de taille importante, quatre sous-évaluations indépendantes ont été réalisées : une par langue. Pour cela, nous avons constitué quatre corpus monolingues rassemblant des textes de nature variée. Ces textes sont bien entendu différents de ceux ayant permis l'extraction des suffixes lexicaux, cela pour n'introduire aucun biais. La métrique des corpus est fournie par le tableau 2.5.

	caractères	segments monolingues
français	324 959	2 968
anglais	286 678	3 034
espagnol	252 666	3 334
allemand	289 506	2 247

TAB. 2.5 - *Caractéristiques des corpus*

Pour chaque sous-évaluation, correspondant au traitement d'un corpus monolingue, nous n'avons procédé à une vérification manuelle que pour les segments qui ont été diagnostiqués comme écrits dans une des trois autres langues. Ces vérifications manuelles se sont avérées fort utiles d'une part pour comprendre l'origine des erreurs, d'autre part pour constater que les corpus supposés monolingues ne l'étaient en fait pas totalement, à cause de citations.

```
 DEF   MotGram     WORD     motGram.flx
 DEF   Nombre      WORD     numeraux.flx
 DEF   PrefixGram  PREFIX   prefixGram.flx
 DEF   SuffixGram  SUFFIX   suffixGram.flx
 DEF   Alphabet    REGEX    alphabet.flx

IF Alphabet
   IF NOT MotGram
      IF NOT Nombre {
         PrefixGram
         SuffixGram
      }
```

(a) Suffixes désactivés

```
 DEF   MotGram     WORD     motGram.flx
 DEF   Nombre      WORD     numeraux.flx
 DEF   PrefixGram  PREFIX   prefixGram.flx
 DEF   SuffixGram  SUFFIX   suffixGram.flx
 DEF   Alphabet    REGEX    alphabet.flx
 DEF   SuffixFreq  SUFFIX   suffixLex.flx

IF Alphabet
   IF NOT MotGram
      IF NOT Nombre {
         PrefixGram
         IF NOT SuffixGram
            SuffixFreq
      }
```

(b) Suffixes activés

FIG. 2.3 - *Paramétrage déclaratif des expériences*

2.4.4 Analyse des résultats

Capacité à converger vers une seule langue

La capacité à converger vers l'identification d'une seule langue est un paramètre d'évaluation que nous avons jugé pertinent car un segment monolingue peut se voir assigner plusieurs langues lorsque le système ne possède pas assez de connaissances pour en déterminer une seule. Pour mesurer cette capacité à converger, nous avons classé les segments monolingues dans trois classes selon leur niveau de désambiguïsation :

- la classe L qui contient les segments pour lesquels le système a identifié une seule langue ;
- la classe H qui contient les segments pour lesquels une langue au moins a été écartée mais pour lesquels il y a toujours hésitation ;
- la classe I qui contient les segments pour lesquels aucune langue n'a été écartée.

		Expérience n° 1 : suffixes désactivés			Expérience n° 2 : suffixes activés		
nb de mots	nb de segments monolingues	%L	%H	%I	%L	%H	%I
1	531	9,42	16,57	74,01	33,90	21,28	44,82
2	281	46,98	12,81	40,21	75,09	13,17	11,74
3	294	74,49	17,35	8,16	92,18	6,80	1,02
4	316	87,03	11,39	1,58	96,20	3,80	0
5	427	93,91	5,15	0,94	99,06	0,94	0
6	405	97,78	2,22	0	99,75	0,25	0
7	438	98,86	1,14	0	99,77	0,23	0
8	476	98,95	1,05	0	100	0	0
9	446	99,78	0	0,22	100	0	0
10	488	99,80	0,20	0	100	0	0
...	...	100	0	0	100	0	0

TAB. 2.6 - *Capacité à converger vers une seule langue*

Les résultats concernant la capacité des connaissances à converger vers l'identification d'une seule langue sont récapitulés au tableau 2.6 page précédente. Une première analyse superficielle montre que l'utilisation des suffixes fréquents augmente très sensiblement la capacité à converger. Une analyse comparative plus approfondie des deux expériences permet d'identifier quatre types de segments en fonction de leur longueur et de faire apparaître la contribution des sources de connaissances :

- De 1 à 2 mots, il n'y a généralement pas de mots grammaticaux. L'analyseur traite des segments inclus (entre-guillemets, entre-parenthèses), parfois des répliques brèves de dialogue. Dans ce genre de situation, mieux vaut ne pas utiliser des méthodes qui catégorisent systématiquement mais plutôt chercher à réduire l'ambiguïté et se réserver ainsi la possibilité d'activer par la suite des connaissances d'une autre nature telles que le contexte d'apparition ;

- De 3 à 5 mots, le paysage est assez varié. Il s'agit aussi bien de titres et de phrases courtes que d'énumérations. Les mots grammaticaux commencent à percer mais de manière trop sporadique pour espérer une catégorisation de haute qualité. C'est sur les lexicaux qu'il faut alors compter pour effectuer l'identification ;

- Entre 6 et 8 mots, les mots grammaticaux apparaissent désormais en bon nombre. Il reste certaines ambiguïtés provenant du fait que certains mots peuvent être grammaticaux simultanément dans plusieurs langues. Un simple suffixe lexical peut alors faire la différence car après l'application des mots grammaticaux, il ne reste généralement plus que deux langues en compétition ;

- Au dessus de 8 mots, nous affrontons des phrases relativement bien construites, souvent même complexes. Les mots grammaticaux sont alors nombreux et suffisent à catégoriser ces énoncés. Rares sont les cas où les suffixes viennent en soutien.

Capacité à identifier la bonne langue

Les matrices d'analyse des confusions 2.7(a) et 2.7(b) permettent de contrôler la capacité à identifier la bonne langue, à chaque fois qu'une seule langue a été isolée. Pour conserver une homogénéité des chiffres affichés, les doublons n'ont pas été supprimés. Ainsi, parmi les 25 phrases allemandes catégorisées à tort françaises, on dénombre 24 occurrences de la phrase constituée de l'unique mot «*Artikel*».

langue du corpus	nb de segments monolingues	bien identifiés	identifiés à tort fra.	ang.	esp.	all.
français	2 968	(2 618)	?	0	0	0
anglais	3 034	(2 728)	0	?	0	0
espagnol	3 334	(3 282)	0	1	?	0
allemand	2 247	(2 155)	0	0	0	?

(a) Expérience n° 1 : suffixes désactivés

langue du corpus	nb de segments monolingues	bien identifiés	identifiés à tort fra.	ang.	esp.	all.
français	2 968	(2 802)	?	2	5	5
anglais	3 034	(2 769)	4	?	5	10
espagnol	3 334	(3 296)	9	2	?	0
allemand	2 247	(2 186)	25	0	0	?

(b) Expérience n° 2 : suffixes activés

Les corpus étant supposés monolingues, nous n'avons pas vérifié les segments étiquetés avec la langue du corpus. Ceci explique la présence des points d'interrogations dans les matrices de confusions et la mise entre-parenthèses des segments bien étiquetés.

TAB. 2.7 - *Matrices des confusions inter-langues*

La première observation de ces résultats dévoile sans grande surprise le fait que la quasi-totalité des étiquettes erronées sont à imputer à l'utilisation des suffixes fréquents. L'étude au cas par cas nous montre que ces erreurs concernent majoritairement des segments de longueur comprise entre 1 et 5 mots, ce qui corrobore les conclusions de l'analyse précédente. Nous notons cependant quelques absences dans nos bases de mots grammaticaux.

La coopération des trois méthodes s'avère qualitativement très satisfaisante :

- Les mots grammaticaux font un travail quasiment parfait : on constate au regard du tableau 2.7(a) page précédente qu'une seule erreur leur est attribuable : elle est causée par une absence dans la base des mots grammaticaux espagnols ;

- Les suffixes fréquents améliorent nettement la catégorisation des segments courts, d'une part en agissant seuls sur ceux de un à cinq mots, d'autre part en intervenant en soutien des mots grammaticaux sur la tranche des six/huit mots. La catégorisation qu'achèvent ensemble les deux sources est bonne, les mots grammaticaux écartant de manière sûre les langues improbables. Mais, sur les segments courts, l'action des suffixes fréquents est vraiment massive et cela induit des erreurs d'analyse qu'il convient de mettre en relation avec le manque de précision de l'application de la connaissance sur un mot cible : dans notre approche, le suffixe fréquent français *-ant* marquant le participe présent vérifiera aussi bien «*ajoutant*» que «*éléphant*», le suffixe fréquent français *-el* vérifiera aussi bien le mot français «industriel» que le mot allemand «*Artikel*». Ces erreurs restent cependant en nombre restreint comparé au gain obtenu ;

- L'action de l'alphabet est mineure. Cette connaissance n'est vraiment intéressante que par le fait qu'elle limite l'application des suffixes lexicaux fréquents aux mots lexicaux dont toutes les lettres appartiennent à l'alphabet. En cela, elle permet d'éviter quelques erreurs. Objectivement très peu car dans les langues que nous avons sélectionnées, les diacritiques ne sont pas vraiment discriminants et, dans les corpus ils apparaissent en nombre limité, surtout dans les segments courts.

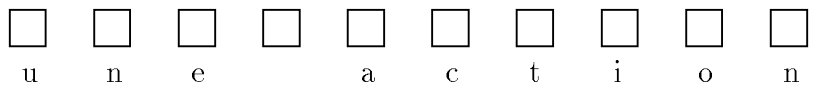
Les évaluations détaillées pour les quatre langues sont présentées en annexe A.1 page 158.

2.5 Regard sur la méthode

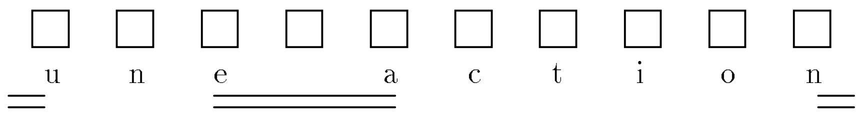
Dans ces travaux, nous avons été amené à développer une méthode permettant une application pertinente de connaissances formelles caractérisant différentes parties des mots d'une langue (e.g., les préfixes, les suffixes). Nous avons cherché à ce que ces connaissances soient utilisées dans un contexte similaire à celui dont elles ont été extraites. Sachant que l'unité d'entrée est une chaîne de caractères et que les connaissances portent sur les mots, nous partitionnons la chaîne de caractères en mots de manière à faire apparaître les domaines au sein desquels ont été extraites ces connaissances. L'obtention de cette partition passe par l'identification de marques de début et de fin de mots. Une fois cette partition établie, nous utilisons la position à l'intérieur du mot pour appliquer correctement la connaissance (e.g., un suffixe est appliqué en fin de mot). Cette méthode est schématisée figure 2.4 page suivante. Notons que les trois opérations (b), (c) et (d) sont exécutées simultanément, en une seule passe, de la gauche vers la droite, le texte étant considéré comme un flux.

Après analyse des sorties, nous concluons que la partition en mots et que l'usage de la position dans cette unité ne suffisent pas pour obtenir une application pertinente des connaissances formelles. Nous avons remarqué en effet que le suffixe «*-ant*», introduit pour caractériser le participe présent, pouvait sous ces conditions être appliqué aussi bien sur *ajoutant* que sur *éléphant*, ce qui n'est pas pertinent. Il manque en fait à la méthode la prise en compte de la catégorie du mot en contexte pour permettre une application sélective des règles : le suffixe «*-ant*» caractérise les verbes au participe présent, il faut chercher à ne l'appliquer que sur ces verbes. Le problème de la catégorie du mot se pose. Elle peut être déterminée de deux manières : d'une part à l'aide d'indices internes (i.e., morphologiques), en utilisant la forme des éléments qui le compose (e.g., le suffixe «*-aient*» caractérise un verbe), d'autre part à l'aide d'indices externes (i.e., morpho-syntaxiques), en utilisant la manière dont il est intégré dans son environnement (l'unité supérieure).

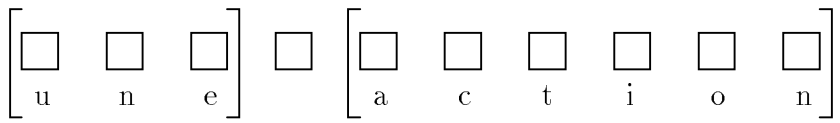
Si nous souhaitions diminuer le nombre d'erreurs causé par l'application incertaine des suffixes fréquents, il faudrait alors restreindre la liste aux suffixes catégorisant de manière relativement sûre un mot. Il s'agit là de l'utilisation d'indices internes. Le recours à des indices externes complémentaires améliorerait les résultats. Le principe consiste en l'extraction de contextes distributionnels sûrs à partir de marques formelles individuellement plus ou moins discriminantes portées par des mots contigus. Par exemple, en utilisant le contexte distributionnel formé des mots grammaticaux «*en lui*» et du suffixe peu discriminant «*-ant*», il est possible de catégoriser avec certitude le dernier mot : un verbe au participe présent. Hervé DÉJEAN (1998a) propose un algorithme d'extraction de tels contextes.



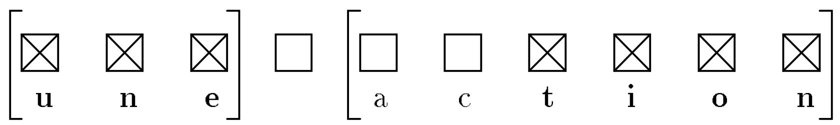
(a) Une suite de caractères



(b) Identification de marques de début et de fin de mots



(c) Délimitation des mots



(d) Utilisation du critère positionnel dans le mot

FIG. 2.4 - *Partition en mots pour une application pertinente des connaissances sur le mot*

2.6 Conclusion

Nos travaux sur l'identification de la langue d'un énoncé monolingue montrent qu'il est possible d'obtenir un taux d'identification de la langue très élevé en utilisant des connaissances linguistiques à la fois bien construites et en en maîtrisant l'application. Les ressources que nous utilisons sont les mots grammaticaux, l'alphabet et les suffixes fréquents des langues.

Nous avons réalisé un programme informatique nous permettant de simuler aisément diverses stratégies d'application des ressources linguistiques. Ces simulations ont été rendues possibles par l'utilisation simultanée de ressources déclaratives et d'un paramétrage déclaratif de la structure de leur enchaînement.

Testé sur la discrimination entre le français, l'anglais, l'espagnol et l'allemand, le système réussit une identification parfaite des énoncés de plus de cinq mots et génère un nombre d'erreurs très faible sur les segments de taille inférieure, ces erreurs étant dues au manque de maîtrise du critère positionnel lors de l'application de la ressource des suffixes fréquents. Nous en concluons que l'utilisation de cette ressource sur les petits segments n'est pas à prescrire systématiquement et dépend du compromis entre décision et précision que l'on s'accorde.

Nous insistons donc particulièrement sur l'importance du critère positionnel lors de la constitution d'une base de connaissances extraites d'un corpus. La base doit englober non seulement la connaissance mais également le contexte précis duquel elle a été extraite afin que lors de son utilisation, le système puisse s'assurer de la similarité des contextes d'application et d'extraction. Le critère positionnel permet à la fois une caractérisation linguistique pertinente et un traitement informatique bien intégré.

La langue de certains segments demeure ambiguë lorsque le système n'a pas suffisamment d'information pour réaliser le diagnostic ; quatre pour cent des segments sont concernés. Cela réserve la possibilité d'incorporer de nouvelles connaissances pour traiter ces reliquats. Dans l'analyse des sorties de l'analyseur, nous avons par exemple constaté que nous pouvions incorporer les ponctuations qui sont discriminantes pour l'espagnol, nous avons également noté que la majorité des ambiguïtés seraient correctement levées si nous faisions usage de la langue des segments environnants, dans de telles situations.

Concernant le calcul de la structure multilingue d'un document, l'implémentation que nous avons réalisée permet de valider la pertinence des marques de changement de langues identifiées dans le chapitre précédent ; les segments délimités sont toujours monolingues. L'implémentation met également en évidence le bien-fondé de notre méthode de calcul basée sur la concaténation d'unités monolingues minimales car nous sommes parvenu à catégoriser des segments courts, ce qui *a priori* n'était pas évident. Le segmenteur en unités monolingues est encore un peu primaire et il génère parfois des unités plus minimales qu'attendues ; il nécessiterait d'être amélioré, ce qui faciliterait alors l'identification de la langue. Enfin, nous avons été agréablement surpris par le fait que notre logiciel révélait des portions multilingues dans des corpus de tests que nous pensions exclusivement monolingues.

Concernant l'intégration de cet outil en front d'une synthèse vocale, notons que savoir décider la langue sur un nombre très restreint de mots est un atout réel. En effet, dans ce type de traitements temps-réel où le texte est traité comme un flux (e.g., la synthèse vocale, le filtrage d'informations), il faut non seulement savoir traiter correctement l'information mais également la traiter rapidement, la synthèse vocale devant être au moins aussi rapide que la vitesse d'élocution d'un humain.

Les travaux sur l'identification de la langue ont fait l'objet de deux publications se trouvant sous les références (GIGUET, 1995a) et (GIGUET, 1995b). Une démonstration est accessible par internet à l'URL

http://www.info.unicaen.fr/~giguet/diagnostic-fr.html

et l'intégration de notre système en tant que module dans différentes applications industrielles est actuellement en cours.

Chapitre 3

Segmentation dans un cadre multilingue

3.1 Introduction

L'étude dans un cadre multilingue entraîne la mise en évidence de propriétés linguistiques partagées par plusieurs langues et l'émergence de caractéristiques spécifiques à chacune d'entres elles. Le multilinguisme permet une meilleure organisation de la connaissance. Il élargit les connaissances sur un ensemble de langues tout en enrichissant celles que nous possédons sur chaque langue particulière.

Le travail dans un univers monolingue après un détour par le multilinguisme est à tout jamais empreint d'une certaine maturité car les propriétés multilingues, du fait de leur généralité, constituent des points d'ancrage solides pour une réflexion monolingue de qualité. Sur le plan conceptuel, une application monolingue reflétant ces distinctions s'avère être de structuration bien meilleure. L'application est naturellement tournée vers le multilinguisme et son extension vers d'autres langues nécessite des modifications bien moins profondes qu'une application pensée uniquement dans un cadre monolingue.

Nous allons illustrer ces propos au travers de notre expérience acquise sur la segmentation en mots des énoncés monolingues servant d'entrée au module d'identification de la langue présenté dans le chapitre précédent. Le problème que nous souhaitions résoudre était le suivant : la segmentation en mots d'un énoncé dépendant de sa langue, comment réussir une segmentation propre d'un énoncé alors que l'on ne connaît pas encore sa langue ?

Trois méthodes de segmentation ont été expérimentées. Nous allons les présenter successivement.

3.2 Utilisation d'un segmenteur monolingue

La première expérience consiste en l'utilisation d'un segmenteur traditionnel du français, situé en amont de la fonction d'identification de la langue. Dans cette configuration, les résultats de la fonction d'identification sont globalement bons pour les quatre langues discriminées (i.e., le français, l'anglais, l'espagnol et l'allemand). Ceci s'explique simplement par le fait que beaucoup de points de segmentation sont identiques entre le français et les autres langues. Nous constatons cependant beaucoup d'unités mal segmentées car les règles de segmentation du français ne sont pas multilingues. À titre d'illustration, étudions le cas de l'élision qui marque très souvent la disparition d'une voyelle. Selon la langue, il existe différentes manières de réaliser la segmentation au niveau de l'élision. Voici quelques exemples :

- en français, l'élision d'un pronom, d'un déterminant ou d'une conjonction est courante si le mot suivant débute par une voyelle. La voyelle élidée est dans ce cas celle qui termine le mot grammatical (e.g., *le avion* ⇒ *l'avion* ⇒ *l'* + *avion*) ; dans la langue populaire, l'élision reflète par contre la disparition d'une voyelle au sein d'un seul mot et ne donne pas matière à segmentation (e.g., *petit* ⇒ *p'tit* ⇒ *p'tit*) ;

- en anglais, l'élision derrière le pronom sujet se traduit par une disparition de la voyelle du verbe (e.g., *they are* ⇒ *they're* ⇒ *they* + *'re*) ; en cas de contraction de la négation, l'élision est effectuée au sein de la négation (e.g., *does not* ⇒ *doesn't* ⇒ *does* + *n't*) ;

- en italien, l'élision du mot grammatical devant le nom est similaire à celle du français (e.g *della arte* ⇒ *dell'arte* ⇒ *dell'* + *arte*).

Si l'on s'en tient à une segmentation «à la française», il est alors clair que beaucoup de mots grammaticaux non français ne peuvent pas être identifiés. C'est un inconvénient pour la fonction d'identification de la langue qui se base précisément sur ces mots pour effectuer son diagnostic.

Bien sûr, il reste alors la possibilité de modifier les lexiques de mots grammaticaux de manière à y faire apparaître les unités produites par le segmenteur français mais la solution n'est vraiment pas acceptable. Sur le plan linguistique, on imagine assez mal en effet l'ajout d'unités telles que *doesn'* ou *they'* dans le lexique anglais. Par ailleurs, même si nous concédions de tels ajouts, nous perdrions une information car des segments tels que *doesn't* et *they're* sont la contraction de deux mots grammaticaux et on ne peut vraiment tolérer l'ajout de *t* et *re* dans le lexique anglais.

3.3 Conception d'un segmenteur multilingue

La conception d'un segmenteur multilingue paraît être la seule voie acceptable pour obtenir un découpage correct de toutes les unités. Enrichir le segmenteur français avec des règles propres aux autres langues est une solution à court terme car nous obtiendrions un segmenteur, certes «multilingue», mais avec un ensemble de règles multilingues grossissant difficilement gérable et redondant avec celui des segmenteurs monolingues. En effet, les règles du français et de l'anglais que contiendrait le segmenteur multilingue seraient également contenues dans les segmenteurs monolingues français et anglais. Nous aurions donc une kyrielle de segmenteurs à maintenir avec qui plus est, un nombre important de règles redondantes à gérer.

Nous avons donc préféré reprendre l'étude de la segmentation dans un cadre multilingue afin d'en dériver une organisation conceptuellement élégante des connaissances. C'est à partir de cette organisation que nous avons défini notre segmenteur multilingue.

3.3.1 Organisation des connaissances

En étudiant les problèmes de segmentation dans cinq langues ouest-européennes[1], nous avons constaté que la plupart des marques de séparation des mots d'un énoncé leur étaient communes. Parmi ces marques, nous trouvons par exemple les espaces et la ponctuation.

D'autres marques jouent le rôle de délimiteur mais ont un comportement spécifique à chaque langue. Nous avons précédemment noté le cas des contractions en anglais (e.g., *that's*, *couldn't*), nous aurions pu ajouter la marque du génitif (e.g., *John's*). En français nous avons vu des cas de contraction par élision (e.g., *l'envie*, *j'aime*, *qu'elle*), nous pourrions citer l'inversion des pronoms (e.g., *donne-le*, *veux-tu*).

D'un point de vue conceptuel, cette analyse multilingue nous amène à considérer une organisation simple et cohérente des connaissances sur la segmentation. Cette structure nécessite :

– une base de connaissances partagée par toutes les langues ;

1. le français, l'anglais, l'espagnol, l'allemand et l'italien

– une base de connaissances spécifique pour chaque langue.

Du point de vue monolingue, cette organisation clarifie le principe de la segmentation car elle distingue les règles multilingues relativement solides, des règles plus spécifiques. À la place d'une base de connaissances non structurées, les connaissances sont désormais organisées dans deux bases différentes dont le contenu est bien différencié.

Du point de vue gestion, cette organisation facilite la maintenance car elle supprime la redondance des règles de segmentation. Les règles communes à chaque langue sont regroupées au sein d'une même base au lieu d'être dupliquées dans chacun des segmenteurs monolingues.

D'un point de vue opératoire, cette organisation évite la création de plusieurs segmenteurs : un seul segmenteur suffit pour réaliser les différentes segmentations. Pour segmenter une langue particulière, le segmenteur doit appliquer sur un énoncé la base de règles de segmentation partagée et la base de règles de segmentation spécifique à la langue ciblée. Pour segmenter dans un contexte multilingue, il doit appliquer toutes les bases de règles.

3.3.2 Application des bases de connaissances

La section précédente présente la segmentation monolingue et multilingue comme l'application, sur un énoncé, de plusieurs bases de règles de segmentation: une base de règles partagée et une ou plusieurs bases de règles spécifiques. La faisabilité de l'application des bases de règles de segmentation n'est pas immédiate car certaines règles peuvent s'avérer conflictuelles :

– dans le cas de la segmentation monolingue, le problème ne se pose pas vraiment car les règles partagées et les règles spécifiques agissent sur des contextes différents : la base de règles partagée s'appuie sur des marques non ambiguës sur le plan multilingue alors que la base spécifique s'appuie sur des zones ambiguës sur ce même plan. L'intersection des contextes d'application est donc vide ;

– dans le cas de la segmentation multilingue l'application simultanée de toutes les bases spécifiques est susceptible de révéler des points de segmentation contradictoires puisqu'elles agissent dans des contextes d'application identiques.

Qu'en est-il en pratique? Jusqu'à présent, la mise en œuvre informatique que nous avons effectuée dans le cadre de l'identification de la langue n'a pas permis de détecter de conflits sur les cinq langues que nous avons étudiées. Nous pouvons avancer une explication de ce constat.

La plupart des règles spécifiques écrites mettent en évidence des mots grammaticaux qui resteraient agglutinés à d'autre mots si l'on n'utilisait que la base de règles partagée. Comme nous l'avons noté lors de l'identification de la langue, les mots grammaticaux sont peu nombreux et relativement différents d'une langue à l'autre. Les cas d'agglutination n'impliquent qu'une petite proportion de ces mots. La probabilité de rencontrer un conflit est donc faible.

Un conflit apparaîtrait par exemple si deux règles de bases différentes impliquant un mot grammatical dans un phénomène d'élision ou d'inversion prescrivaient des segmentations différentes. Nous n'avons pas noté de tels cas au niveau multilingue mais pouvons illustrer cette situation au niveau monolingue : le segment *rendez-vous* peut constituer un nom au quel cas il s'agit d'une seule unité ou bien le verbe *rendre* conjugué suivi du pronom inversé *vous*.

3.3.3 Mise en œuvre informatique

Une mise en œuvre informatique à base de règles de segmentation a été réalisée. Chaque règle de segmentation caractérise non pas le mot mais, d'une manière duale, la frontière entre deux mots, le mot étant défini implicitement par son début et sa fin.

La segmentation par caractérisation des frontières est intéressante à plusieurs titres. D'une part, l'écriture des règles de segmentation devient très agréable car la caractérisation des frontières permet un certain détachement quant à la structure interne du mot. D'autre part, l'organisation interne des bases de règles n'est soumise à aucune contrainte d'ordonnancement car les contextes d'application des règles sont toujours différents, ce qui facilite leur constitution et leur maintenance. Dans notre implémentation, le caractère déclaratif des bases élimine par ailleurs toute phase de compilation ou précompilation.

Concrètement, une règle de segmentation se traduit par une expression régulière qui précise le contexte d'application et le point de segmentation. Le contexte d'application est composé de trois parties dont 2 optionnelles : la frontière, son contexte gauche et son contexte droit.

Nous avons proposé un algorithme effectuant la segmentation dans un cadre aussi bien monolingue que multilingue en une seule passe de la gauche vers la droite sur l'ensemble des caractères de l'énoncé. Cet algorithme est donc de complexité linéaire. Ces travaux sont détaillés et ont fait l'objet d'une publication se trouvant sous la référence GIGUET (1996).

En utilisant un tel système, les mots grammaticaux agglutinés sont proprement segmentés. Du point de vue identification de la langue, il suffit alors d'ajouter leur graphie dans les bases de mots grammaticaux pour que ceux-ci soit pris en compte. Il faut par exemple ajouter *n't*, *'s* et *'re* dans le lexique anglais, *l'*, *qu'* et *-vous* dans le lexique français.

3.4 Révision du segmenteur multilingue

Le segmenteur précédent propose une manière très élégante de résoudre la segmentation dans un cadre aussi bien monolingue que multilingue. L'organisation conceptuelle des bases de connaissances, distinguant les connaissances de segmentation générales multilingues des connaissances spécifiques monolingues paraît intéressante. Ce segmenteur apporte également une nouvelle vision de la segmentation monolingue.

Nous avons cependant noté un défaut quant à l'application des règles de segmentation monolingues. Ces dernières peuvent en effet se révéler théoriquement contradictoires. Même si en pratique nous n'avons pour le moment pas encore été confronté à la situation, l'ajout de nouvelles langues au module d'identification de langue (i.e., appelé identificateur de langue ou plus simplement identificateur) pourrait en faire apparaître.

Cette situation nous a amené à reconsidérer la répartition des tâches entre le segmenteur et l'identificateur de langue qui lui succède. Il convient de noter que le segmenteur et l'identificateur sont toujours cohérents : le segmenteur ne produit que des unités que l'identificateur sait manipuler, ces unités étant définies dans les ressources lexicales. Dans notre cas, nous avons choisi de concevoir un segmenteur produisant des unités minimales en désagglutinant les mots liés par une apostrophe ou un tiret. Lorsque l'identificateur recherche une unité produite par le segmenteur dans le lexique, il n'effectue donc que des comparaisons à l'identique de chaînes de caractères. Le système trouve ainsi sa cohérence.

La production d'unités minimales est une tradition dans le traitement des langues. Libre à nous d'imaginer une répartition des tâches plus adaptée entre le segmenteur et l'identificateur. Une solution consiste à étendre dans l'identificateur la puissance de manipulation des unités produites par le segmenteur. En l'occurrence, si l'identificateur peut manipuler une séquence préfixe ou suffixe d'une unité produite par le segmenteur, l'action de ce dernier se simplifie puisqu'il n'a plus à séparer les mots agglutinés : c'est à l'identificateur que revient désormais la charge de rechercher les agglutinations. Au niveau du module de segmentation, l'application de règles monolingues devient caduque, ce qui a pour effet de supprimer naturellement le problème des règles contradictoires entre les bases de segmentation. En effet, le segmenteur n'utilise plus qu'une seule base : la base de règles de segmentation multilingue.

Nous avons effectivement enrichi la puissance de manipulation de l'identificateur de manière à ce qu'il puisse effectuer des comparaisons de séquences préfixes et suffixes sur les unités produites par le segmenteur. La segmentation s'est donc réduite à une segmentation purement multilingue très simple et il a suffi de créer trois bases de connaissances sur les mots grammaticaux pour assurer la cohérence du système : une base de mots grammaticaux standards, une base de mots grammaticaux préfixes et une base de mots grammaticaux suffixes.

Concrètement, alors que pour l'énoncé «*s'imaginent-ils*», le premier segmenteur multilingue cherchait à produire trois unités minimales monolingues «*s'*» «*imaginent*» «*-ils*», dans la nouvelle répartition des tâches, le segmenteur se limite à la production d'une seule unité dite multilingue : «*s'imaginent-ils*». Pour vérifier l'existence de mots grammaticaux dans une unité multilingue, l'identificateur effectue une recherche de préfixes ou de suffixes sur cette unité à partir de deux bases : l'une contenant des préfixes grammaticaux (tels que «*s'*»), l'autre des suffixes grammaticaux (tels que «*-ils*»).

3.5 Conclusion

Dans la plupart des applications du traitement des langues, la segmentation entraîne toujours la mise en place de solutions *ad-hoc* jamais vraiment satisfaisantes. En réalisant cette étude multilingue, nous avons apporté un regard nouveau sur la segmentation monolingue. Nous avons montré qu'il était possible et profitable, sur le plan conceptuel, de séparer des règles de segmentation générales multilingues et des règles spécifiques monolingues, la segmentation monolingue consistant alors en l'application d'une base de règles commune à toutes les langues et d'une base de règles spécifiques à une langue particulière.

Par tradition, la segmentation d'un énoncé monolingue se résume toujours en un découpage en unités minimales monolingues. C'est dans cet état d'esprit que nous avons conçu un premier modèle de segmentation multilingue. Nous avons cependant rappelé que la taille des unités segmentées n'était en fait qu'une question de répartition de tâches entre un segmenteur et l'analyseur qui lui succède ; il est possible de modifier cette répartition tout en conservant un équilibre du système.

Dans le cadre de l'identification de langue, le nouvel équilibre vers lequel nous avons évolué consiste en la simplification de la mission du segmenteur afin ne plus lui faire appliquer que des règles de segmentation multilingues. Cette simplification a nécessité, au niveau de l'identificateur de langue, une extension de la puissance de manipulation des unités produites par le segmenteur car la sortie de ce dernier n'est plus une suite d'unités minimales monolingues mais une suite d'unités multilingues.

Les règles de segmentation multilingues que nous avons mises à jour paraissent très intéressantes car elles fournissent implicitement une définition multilingue du «token». Elles permettent donc à notre identificateur de langue, et par extension à toute application multilingue, de travailler sur une unité très stable.

Par ailleurs, cette définition multilingue du token peut être réinvestie dans la conception d'applications monolingues. Si elle est intégrée dans de telles applications, leur ouverture vers une ou plusieurs autres langues n'en sera que facilitée. En effet, la prise en compte d'une nouvelle langue ne remet pas en cause la segmentation multilingue et l'analyseur monolingue ayant nécessité dès sa création le développement d'outils adaptés au traitement d'unités multilingues n'aura que peu de modifications à recevoir. Les seules vraies modifications à apporter se trouvent dans la constitution des ressources linguistiques nécessaires au traitement de la nouvelle langue.

En suivant ce chemin, nous nous orientons vers une unification du «token» dans les différentes applications du traitement des langues. Par exemple, si le «token» d'un analyseur syntaxique monolingue est identique à celui de l'identificateur de langue, les deux modules ne font plus appel qu'à un segmenteur unique et la mise en place d'un identificateur de langue en amont d'un analyseur syntaxique s'effectue en toute harmonie. Le niveau syntaxique doit être cependant conçu de manière à gérer des unités multilingues : il peut appliquer en interne sur chaque token des règles spécifiques de segmentation monolingue pour retrouver une segmentation classique ou bien gérer naturellement des tokens multilingues (e.g., gestion de séquences préfixes et suffixes de tokens, création de catégories syntaxiques reflétant le caractère amalgamé de certaines unités).

Troisième partie

Voyage dans l'analyse syntaxique automatique

Chapitre 1

Vers un nouveau processus d'analyse syntaxique

1.1 L'analyse syntaxique traditionnelle

La chaîne de traitement de l'analyse syntaxique traditionnelle (voir figure 1.1) comprend au minimum deux modules : (1) l'analyseur morpho-lexical, qui, à partir de ressources lexicales considérées comme exhaustives, produit un graphe des différents découpages de la phrase en unités lexicales[1](ou «*token*») munies chacune d'une ou plusieurs descriptions morpho-syntaxiques, (2) suivi de l'analyseur syntaxique proprement dit, qui, pour chacune de ces propositions, produit zéro, une ou plusieurs analyses.

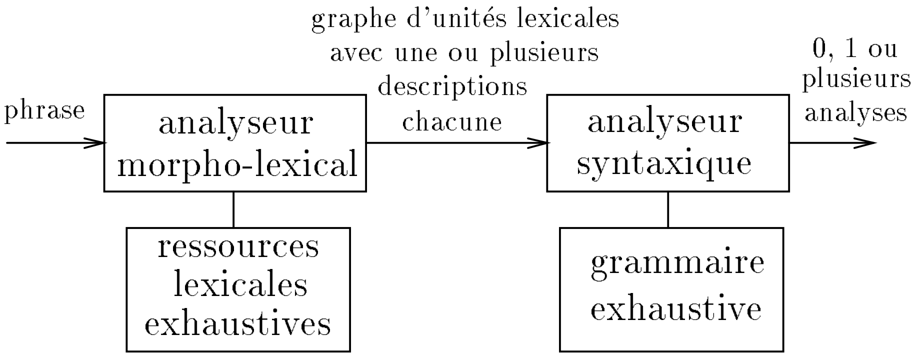


FIG. 1.1 - *Chaîne de traitement de l'analyse syntaxique traditionnelle*

Le problème central de l'analyse syntaxique traditionnelle est son aspect combinatoire, une combinatoire que l'on rencontre à tous les niveaux : au niveau structural mais également au niveau lexical. Au niveau structural, c'est bien entendu à la richesse des constructions syntaxiques de la langue que le processus d'analyse se heurte. Au niveau lexical, c'est à l'ambiguïté du découpage en unités lexicales et surtout à l'homographie polycatégorielle qu'il est confronté.

Il faut bien réaliser que dans ce type d'approche les sources de combinatoire ne s'additionnent pas : elles se multiplient. Le processus d'analyse doit produire *toutes* les structures syntaxiques associées à *toutes* les suites de descriptions morpho-syntaxiques obtenues par substitutions lexicales, et cela

1. Une unité lexicale ou «*token*» est un mot générique et désigne aussi bien un mot qu'une ponctuation, un nombre qu'un acronyme.

pour *chacune* des possibilités de découpage de la phrase en unités lexicales.

Maîtriser la combinatoire est une quête permanente en analyse syntaxique traditionnelle. Une technique récemment introduite et dans laquelle beaucoup d'efforts ont été investis est l'étiquetage morpho-syntaxique. L'étiquetage s'attaque au problème de l'homographie polycatégorielle en utilisant le contexte d'apparition d'un mot dans un énoncé pour en déduire sa catégorie. Toute déduction est réinvestie pour en effectuer de nouvelles : l'étiquetage fonctionne par propagation de contraintes contextuelles.

1.2 L'étiquetage morpho-syntaxique

L'étiquetage morpho-syntaxique, couramment appelé «*tagging*», consiste à catégoriser les unités lexicales d'un énoncé selon des critères morpho-syntaxiques. Pour cela, le système d'étiquetage dispose d'un jeu de classes morpho-syntaxiques communément appelées «étiquettes» ou «*tags*» et le processus cherche à associer à chaque unité lexicale sa classe morpho-syntaxique (ou étiquette) en contexte.

je		le		bois
pronom sujet				
pronom sujet	⇒	pronom objet		
		pronom objet	⇒	verbe

FIG. 1.2 - *Propagation de déductions contextuelles dans l'étiquetage*

Le processus d'étiquetage morpho-syntaxique réalise une propagation de déductions contextuelles sur les différents tokens d'un énoncé comme le montre l'exemple de la figure 1.2. Il est conduit par les fréquences des contiguïtés des étiquettes dans un étiqueteur statistique, ou par des règles de déductions contextuelles symboliques dans un étiqueteur symbolique.

Le processus n'a pas toujours pour ambition d'assigner une étiquette et une seule à chaque unité ; il peut en cas de trop grande pauvreté du contexte d'apparition ou bien en cas de manque de connaissances ne pas lever certaines «ambiguïtés» et produire plusieurs étiquettes pour un seul token.

L'étiquetage morpho-syntaxique a trouvé de multiples applications dans le traitement des langues mais c'est principalement dans la recherche documentaire que les retombées ont été les plus grandes et dans l'analyse syntaxique qu'il a soulevé les plus grands espoirs.

1.3 L'étiquetage et l'analyse traditionnelle

L'étiquetage dans l'analyse syntaxique traditionnelle est vite adopté par une partie de la communauté : l'étiqueteur vient se substituer à l'analyseur morpho-lexical (voir figure 1.3) pour fournir à l'analyseur syntaxique des tokens munis chacun d'une seule étiquette (ou le moins possible), et ainsi annuler, ou réduire le plus possible, la combinatoire engendrée par l'homographie polycatégorielle.

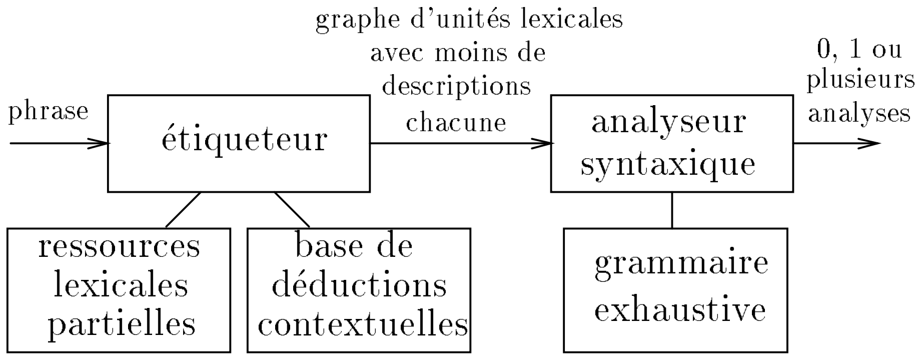


FIG. 1.3 - *Nouvelle chaîne de traitement de l'analyse syntaxique*

Le problème que soulignent SEGOND et COPPERMAN (1997) est que les étiqueteurs commettent des erreurs sur environ cinq pour cent des mots étiquetés, c'est-à-dire un mot sur vingt en moyenne. Sachant que la moyenne des longueurs de phrases dans un corpus tout-venant est supérieure à 20 mots, il manque régulièrement la description morpho-syntaxique attendue sur un des mots et la bonne structure syntaxique n'a alors plus aucune chance d'être détectée.

Alors que l'étiquetage a fait naître beaucoup d'espoirs et continue aujourd'hui d'en nourrir, on ne peut pas vraiment dire qu'il s'agisse d'un véritable renouveau pour l'analyse syntaxique automatique. La qualité des étiqueteurs actuels ne permet pas de résoudre de manière satisfaisante le problème de l'homographie polycatégorielle pour envisager de lui faire succéder un analyseur syntaxique traditionnel.

Par ailleurs, l'étiquetage n'a pour vocation de s'attaquer qu'à une des sources de combinatoire, l'homographie polycatégorielle, il n'apporte pas de solution aux autres sources de la combinatoire constituées d'une part par l'ambiguïté de la segmentation d'un énoncé en unités lexicales, et d'autre part par la richesse des constructions grammaticales utilisées qu'une grammaire a bien du mal à refléter.

1.4 Genèse d'un nouveau processus d'analyse

Dans cette section, nous allons présenter une approche originale de l'analyse syntaxique automatique sur laquelle nous travaillons. Cette approche a pour ambition de calculer la structure syntaxique de phrases de corpus tout-venant. L'analyse, dans son état actuel, consiste à identifier des constituants et à les relier les uns avec les autres.

À l'origine du nouveau processus d'analyse que nous avons conçu se trouve un segment intermédiaire entre les mots et la phrase, un segment qui permet de ne plus considérer la phrase comme une suite continue, indifférenciée de mots, et qui nous donne donc la possibilité de hiérarchiser les difficultés.

1.4.1 Un segment entre les mots et la phrase

Les caractéristiques du matériau linguistique qui ont guidé nos choix dans la conception d'un nouveau processus d'analyse syntaxique résident dans l'existence de deux degrés de liberté très différents coexistant à deux niveaux distincts. Dans des extraits bien connus du Bourgeois Gentilhomme, Mo-

LIÈRE nous offre une illustration parfaite de ce phénomène :

«[Belle marquise], [vos beaux yeux] [me font] [mourir] [d'amour]»
«[D'amour] [mourir] [me font], [belle marquise], [vos beaux yeux]»
«[Vos beaux yeux] [d'amour] [me font], [belle marquise], [mourir]»

L'ordre de petits groupes de mots (présentés ici entre-crochets) paraît assez libre, comparé à l'ordre des mots à l'intérieur de ces groupes qui lui, est relativement fixe. Outre cet ordre contraint des mots à l'intérieur des groupes, on note des contraintes d'accords également fortes entre leurs mots. Nous appelons ces courts segments *syntagmes minimaux* et les appellerons plus simplement syntagmes par la suite. Le syntagme minimal permet de définir une hiérarchie de segments à trois niveaux : les mots, les syntagmes minimaux, les phrases, hiérarchie où le segment d'un niveau est constitué de segments de niveau inférieur : un segment a un type différent de celui de ses parties.

1.4.2 Définition du syntagme minimal

Au travers des extraits du Bourgeois Gentilhomme de MOLIÈRE, nous avons donné une définition implicite du syntagme minimal : il s'agit d'un segment permutable dans lequel l'ordre des mots est relativement fixe et au sein duquel nous constatons des contraintes d'accord assez fortes. Caractérisons plus précisément cette structure.

Le syntagme minimal est une unité constituée d'une série de mots tous contigus les uns aux autres et regroupés autour d'une tête lexicale, le plus souvent un nom ou un verbe, plus rarement un adverbe, un adjectif ou un pronom. Toutes les relations syntaxiques entre mots internes au syntagme ont la particularité d'être calculables de manière positionnelle et non ambiguë. On retrouve l'équivalent de ce segment sous le nom de «chunk» dans la littérature anglaise (ABNEY, 1991, 1995). Il faut également noter que ce segment est stable entre les langues, Hervé DÉJEAN (1998b) a montré sa validité multilingue. Par ailleurs, il trouve approximativement son équivalent à l'oral sous la forme du groupe accentuel. C'est donc sur des bases solides que nous introduisons ce segment pivot dans le calcul de la structure syntaxique. La structure de cet agrégat très contraint est explicitable exhaustivement mais nous ne nous y risquerons pas pour deux raisons :

La première est que la grande variété des formes de notre objet ne nous permet pas de lister toutes les constructions possibles sans risquer d'en oublier quelques unes. Nombreuses sont en effet les catégories syntaxiques intervenant dans la définition du syntagme minimal. On dénote en effet pas moins de neuf catégories syntaxiques dans le syntagme nominal (la préposition, le déterminant, l'adjectif épithète antéposé et postposé, l'adverbe antéposé à l'adjectif épithète, la coordination de prépositions, de déterminants et d'adjectifs épithètes, le nom) et pas moins de treize dans le syntagme verbal (la préposition, les pronoms atones antéposés et postposés, la négation, l'auxiliaire, la copule être, l'adjectif attribut, l'adverbe antéposé à l'adjectif attribut, la coordination d'adjectifs attributs, le verbe conjugué, infinitif, participe passé ou présent).

La seconde raison est qu'il n'est nullement nécessaire de la détailler exhaustivement pour la calculer. Le processus de délimitation consiste en effet simplement à repérer le début d'un syntagme (en se basant soit sur un mot grammatical débutant un syntagme si son début est marquée, soit sur la fin du syntagme précédent si la fin de ce syntagme est marquée, soit sur la présence d'une catégorie syntaxique externe aux syntagmes, soit sur une rupture d'accord), puis à étendre progressivement le syntagme tant qu'il y a compatibilité positionnelle des traits morpho-syntaxiques des tokens parcourus avec ceux précédemment incorporés dans le syntagme en construction.

Il faut noter que la définition d'une unité adaptée au traitement automatique des langues est délicate car deux objectifs sont en concurrence : l'objectif linguistique et l'objectif opératoire. L'objectif linguistique nous oblige à tendre vers une représentation toujours uniforme d'un même phénomène alors que l'objectif opératoire nous pousse vers une définition d'unités facilement manipulables. Dans la conception de nos processus, nous privilégions toujours le côté opératoire qui constitue le centre de notre problématique sachant que nous garantissons toujours la possibilité de retrouver *a posteriori* une représentation uniforme car toutes les relations syntaxiques entre mots internes au syntagme ont la particularité d'être calculables de manière positionnelle et non ambiguë.

À titre d'illustration, lors de la construction d'un syntagme nominal «opératoire», les épithètes postposés contigus au nom se trouvent inclus dans le syntagme lorsque cela n'introduit aucune ambiguïté de rattachement, et se trouvent détachés dans le cas contraire. Le processus de segmentation est donc susceptible de produire deux représentations différentes d'un même phénomène linguistique. Cependant, séparer ou fusionner systématiquement les épithètes *a posteriori* ne pose pas de problèmes et dépend de la théorie linguistique adoptée.

1.5 Conception d'un nouveau processus

Le syntagme minimal fait émerger deux environnements que tout oppose : un environnement très contraint, sa structure intérieure, et un environnement plus souple, les syntagmes dans la phrase. Ces deux environnements si différents nous ont mené à l'idée qu'un processus d'analyse bien construit, c'est-à-dire, respectant les propriétés de son objet d'analyse, se devait de tenir compte de cette observation. Nous en avons tout naturellement déduit que la conception de deux sous-processus utilisant chacun sa propre représentation et exploitant chacun les propriétés de son niveau s'imposait :

1° le premier processus travaille sur une suite de tokens et utilise une technique proche de l'étiquetage morpho-syntaxique pour construire une structure de constituants : une liste composée de syntagmes minimaux et de mots extra-syntagmatiques[2].

2° le deuxième processus accepte cette liste et utilise une technique originale basée sur une propagation de contraintes relationnelles pour construire une structure relationnelle entre syntagmes et mots extra-syntagmatiques.

Ces deux sous-processus dont le dénominateur commun est la propagation de contraintes feront chacun l'objet d'une description détaillée dans les chapitres 2 page 95 et 3 page 117.

2. parmi les mots extra-syntagmatiques nous trouvons les conjonctions, les locutions conjonctives, les coordinations de syntagmes et de propositions, les pronoms relatifs et les ponctuations (la virgule, le point-virgule, le point et les parenthèses).

Dans notre approche, les deux structures sont construites simultanément par l'interaction des deux processus. La figure 1.4 présente le déroulement de l'analyse : deux processus libellés 1 et 2 contrôlent respectivement la construction des syntagmes et leur mise en relation. Le premier processus assigne des informations morpho-syntaxiques sous forme de traits à chaque unité lexicale et délimite les syntagmes à l'aide de marques de début et de fin de syntagme. Le deuxième processus définit des relations entre les syntagmes. L'interaction entre les deux niveaux s'effectue via deux interactions libellées α et β. L'interaction α permet d'informer le processus n° 2 de la détection d'une nouvelle unité syntagmatique et de transmettre les traits morpho-syntaxiques de cette unité. L'interaction β permet au processus n° 1 d'affiner ou de modifier les traits morpho-syntaxiques des unités lexicales internes à un syntagme.

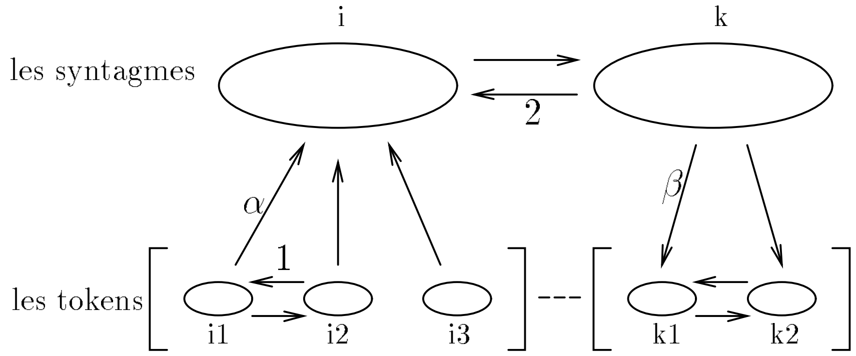


FIG. 1.4 - *Interaction des processus d'analyse*

L'exécution de l'analyse est cyclique. Un cycle d'exécution complet commence par une déduction du processus 1, qui va enrichir à l'aide d'une interaction α le processus 2. Le processus 2 peut alors tenter de construire de nouvelles relations et préciser via une interaction β la structure construite par le processus 1.

1.6 Méthodologie de la conception

Dans cette section, nous abordons un point méthodologique quant à la conception du processus d'analyse. Il s'agit du fait que nous avons opté pour la réalisation d'un processus séquentiel, c'est-à-dire, un processus qui, après un parcours unique des unités de la phrase, de la gauche vers la droite, produit une analyse unique. Trois arguments appuient ce choix.

Premièrement, nous menons une recherche en syntaxe dans laquelle l'ordinateur est un outil d'investigation, un travail qui par définition est centré sur l'étude de la forme des énoncés. La séquentialité, choix délibérément contraignant, nous oblige à exploiter tous les critères formels détectables permettant de diriger le calcul de la structure syntaxique. Aujourd'hui, seule une infime partie de ces critères a été étudiée car notre unité de travail se limite à la phrase. Beaucoup d'indices internes et tous les indices externes à ce domaine restent encore à découvrir afin de caractériser précisément les phénomènes linguistiques que la séquentialité nous fait perdre. Nous plaidons en fait pour une utilisation maximale des propriétés du matériau.

Deuxièmement, nous gardons à l'esprit que l'analyseur que nous construisons devra être intégré à un système de traitement des langues car le calcul de la structure syntaxique ne peut se restreindre à l'utilisation de connaissances purement formelles. Les différentes structures construites par les processus engagés dans un tel système ne peuvent l'être par la sérialisation de processus autonomes mais par leur fonctionnement coopératif. La construction simultanée de toutes les structures linguistiques (e.g., syntaxique, sémantique) va s'effectuer prioritairement de la gauche vers la droite. Dans cette optique et pour chaque processus, le rôle que joue un élément dans une structure est déterminé, non seulement par le rôle des éléments le précédant relativement à ce processus, mais également par l'état de toutes les autres structures en cours de construction par les autres processus. C'est uniquement dans le cadre d'une telle architecture et sur détection d'impossibilité d'intégration d'un élément que l'on envisagera la restructuration d'une ou de plusieurs structures.

Troisièmement, nous travaillons sur la mise au point d'un processus d'analyse syntaxique de corpus tout-venant. Dans cette voie, on ne peut se permettre de proposer plusieurs analyses par phrase car les masses de données traitées sont relativement importantes et produire plusieurs analyses pour chaque phrase rendrait alors les sorties inexploitables car trop abondantes.

Ce sont ces trois arguments qui tout naturellement, nous font accepter de produire des analyses partielles. Ces sorties témoignent à la fois d'un processus en cours d'élaboration pour lequel il reste encore beaucoup d'indices formels à exploiter, et d'un processus encore isolé mais que nous concevons de manière à pouvoir interagir avec d'autres processus faisant intervenir des connaissances non formelles pour construire et enrichir simultanément différentes structures linguistiques.

Chapitre 2

La construction des syntagmes minimaux

2.1 Construction des syntagmes

Au travers du processus de calcul des syntagmes minimaux esquissé dans le chapitre précédent (c.f., section 1.4.2 page 87), on constate que la délimitation des syntagmes s'appuie sur la détection de mots grammaticaux caractéristiques du début ou de la fin des syntagmes, et sur la rupture de l'accord entre deux mots lexicaux contigus. Pour pouvoir détecter ces situations, le processus doit connaître les descriptions morpho-syntaxiques des différents tokens de l'énoncé en contexte. C'est donc vers un processus proche de l'étiquetage morpho-syntaxique préalablement présenté (c.f., section 1.2 page 83) que nous nous tournons.

2.2 Révision des objectifs de l'étiquetage

L'esprit dans lequel nous envisageons l'étiquetage est tout à fait différent de celui avec lequel il a été pensé dans son approche classique. Contrairement à l'usage qu'il en a été fait dans l'analyse syntaxique traditionnelle, l'étiquetage n'est ici pas entendu comme un prétraitement censé réduire la polycatégorie lexicale à son minimum pour que l'analyse syntaxique puisse être calculée en un temps plus raisonnable. Dans notre processus d'analyse syntaxique, l'étiquetage des tokens se déroule tout au long de la construction de la structure syntaxique : il débute, certes, l'analyse syntaxique, mais nous ne le considérons définitif qu'au moment où l'analyse syntaxique complète est elle-même terminée. Cette position se justifie tout simplement par le fait que certains traits morpho-syntaxiques ne peuvent être déduits que de la structure syntaxique. Nous pensons notamment au calcul de la catégorie syntaxique de «*de*» partitif ou préposition et à celui de «*que*» pronom relatif, conjonction ou adverbe ou bien au genre et nombre des pronoms relatifs.

L'étiquetage parfait est un objectif inaccessible pour un module d'étiquetage isolé, mais un objectif que peut atteindre le système d'analyse syntaxique dans son ensemble. Pour cela, il doit y avoir interaction permanente entre les différents traitements de l'analyse syntaxique afin que ceux-ci s'affinent mutuellement ; c'est d'ailleurs l'objectif du processus d'analyse que nous proposons. Dans notre système, l'objectif de l'étiqueteur n'est donc pas tant de poursuivre une quête obsessionnelle de l'étiquetage optimal que de fournir un bon découpage en syntagmes et un bon typage de ces syntagmes afin que leur mise en relation puisse débuter rapidement et que des échanges fructueux entre les modules apparaissent au plus tôt.

Le découpage en syntagmes et leur typage ne nécessitent en fait qu'un étiquetage minimal que peut assumer l'étiqueteur seul, étiquetage amorcé par les mots grammaticaux de début de syntagme, de début de proposition, les ponctuations et aidé par les ressources de mots lexicaux. Il n'est par exemple pas toujours nécessaire de lever les homographies nomino-adjectivales puisque ces catégories sont internes au même syntagme minimal. Bien entendu, si le but des recherches est la construction de l'analyse syntaxique la plus précise alors un étiquetage de qualité supérieure devra être recherché.

2.3 Les traits morpho-syntaxiques

2.3.1 Les traits morpho-syntaxiques pertinents

Nous avons vu que le découpage en syntagmes minimaux nécessitait un étiquetage morpho-syntaxique des tokens de l'énoncé : à chaque token doit être associée une description morpho-syntaxique en contexte. Le terme description morpho-syntaxique est très large et il nous faut le préciser en présentant les traits morpho-syntaxiques pertinents pour l'étiquetage.

Nous savons que la recherche des limites de syntagmes se fonde principalement sur la catégorie syntaxique des mots (notamment la catégorie des mots grammaticaux), et que, dans une moindre mesure, elle utilise les marques morphologiques des mots lexicaux (plus précisément les marques d'accord).

Les connaissances morpho-syntaxiques devant être manipulées au cours de l'étiquetage sont donc tout naturellement la catégorie syntaxique, le genre, le nombre et la personne. Les attributs morphologiques n'étant pertinents que pour certaines catégories syntaxiques, il convient que leur définition soit subordonnée à celle des catégories.

Les traits morphologiques ont une définition qui ne pose pas vraiment de problème, par contre la constitution d'un jeu de catégories syntaxiques adapté à l'étiquetage n'est pas triviale. Nous devons donc rappeler les caractéristiques souhaitées d'un tel jeu.

2.3.2 Jeu de catégories syntaxiques et distribution

Dans cette section, nous souhaitons insister sur l'importance de la notion de «distribution» lors de la définition du jeu de catégories syntaxiques. La distribution d'un élément est l'ensemble des environnements (i.e., des contextes) dans lesquels l'élément apparaît (BLOOMFIELD, 1933).

La question à laquelle nous devons répondre avant d'entamer la définition de ces catégories est la suivante : quelles sont les caractéristiques d'un jeu de catégories adéquat aux déductions contextuelles ? Pour répondre à cette question, nous devons rappeler que l'étiquetage grammatical traditionnel est basé sur l'étude des régularités des contiguïtés de catégories, c'est-à-dire sur des régularités *distributionnelles*. Sachant que les régularités sont obligatoirement captées au travers du filtre que constitue le jeu de catégories, ce dernier doit être conçu de manière à en permettre la perception. Deux conditions doivent être remplies pour satisfaire cet objectif :

1° des catégories différentes doivent regrouper des mots ayant des environnements différents, aux homographies polycatégorielles près ;

2° une catégorie doit regrouper des mots ayant des environnements identiques, aux homographies polycatégorielles près.

Quand elles sont vérifiées, ces deux conditions assurent la distributionnalité du jeu de catégories. Nous allons concrétiser les deux conditions précédentes en montrant qu'elles mettent à défaut le caractère distributionnel des «parties du discours» (GREVISSE, 1969).

On considère «*un*» et «*une*» en tant qu'articles, «*cette*» et «*ces*» en tant qu'adjectifs démonstratifs, «*sa*» et «*ses*» en tant qu'adjectifs possessifs. Ces choix violent la condition n° 1 car leur distribution est identique : ils apparaissent tous en tête de syntagme nominal. Ces mots sont maintenant plus souvent regroupés sous une catégorie unique : le déterminant, catégorie fondée sur la régularité de la classe distributionnelle des déterminants, différenciée de celle des adjectifs.

Les parties du discours regroupent par contre les mots «*il*», «*elle*», «*nous*», «*y*», «*qui*», «*dont*» sous une seule classe : celle des pronoms. Bien qu'on puisse les trouver sous-catégorisés en tant que pronoms personnels, pronoms relatifs et pronoms adverbiaux, les environnements de ces mots sont si différents que chaque mot pris indépendamment devrait être considéré comme l'unique élément de sa propre classe distributionnelle comme le suggère la condition n° 2.

Cette illustration est choisie à des fins pédagogiques car il est bien évident que les parties du discours constituent une taxinomie traditionnelle et empirique surtout consacrée à l'enseignement. Cette taxinomie est justifiée par des propriétés qui ne sont pas uniquement positionnelles : c'est la différence entre une catégorisation syntagmatique (critère positionnel dominant) et une catégorisation paradigmatique (critère morphologique dominant). Nous avons toute liberté d'en créer une autre, si celle-ci est plus adéquate à l'étiquetage et si elle reste linguistiquement fondée.

En conclusion, pour capter (puis utiliser) des régularités distributionnelles, chaque catégorie doit définir une classe distributionnelle de mots. Par exemple, parmi les adjectifs, on devra différencier les épithètes antéposés, les épithètes postposés, et les attributs, car leurs distributions sont différentes.

Un jeu de catégories distributionnel, associé au concept de syntagme minimal, implique que ce jeu soit partitionné en autant de sous-ensembles qu'il y a de types de syntagme minimal (leur intersection est vide). Par exemple, la catégorie de l'adjectif épithète, dans le syntagme nominal, est différente de celle de l'adjectif attribut, dans le syntagme verbal.

2.3.3 Granularité du jeu de catégories syntaxiques

Dans la section précédente, nous avons montré l'importance de la notion de distribution dans la définition du jeu de catégories syntaxiques. Nous allons maintenant aborder l'influence de la granularité du jeu de catégories sur la qualité de l'étiquetage. La granularité correspond au nombre de catégories du jeu.

Il est encore aujourd'hui commun d'entendre que la granularité influence directement la qualité de l'étiquetage de sorte que plus la granularité est fine (i.e., plus il y a de classes dans le jeu) et plus la qualité de l'étiquetage se dégrade et qu'inversement, plus elle est grosse et plus la tâche est aisée.

Cette vision est un peu trop primaire. Luttant contre cet *a priori*, ELWORTHY (1995) montre empiriquement que la variation en taille d'un jeu peut avoir des répercussions différentes sur la qualité de l'étiquetage selon la langue.

La réalité s'avère en fait différente. En créant une relation de causalité directe entre la granularité du jeu de catégories et la qualité de l'étiquetage, on oublie que les régularités distributionnelles n'apparaissent qu'au travers du filtre du jeu de catégories et que ces régularités ne peuvent être précisément captées si le filtre n'est pas lui-même conçu pour en permettre l'émergence.

En l'occurrence, moins le jeu comporte de catégories syntaxiques et plus il a de chance de regrouper des contextes distributionnels différents à l'intérieur d'une même catégorie. Et, plus les contextes distributionnels sont variés à l'intérieur d'une même catégorie et moins les régularités peuvent être captées.

En conclusion, même si la granularité des jeux de catégories peut varier d'un système à l'autre selon les objectifs poursuivis, l'important est qu'ils restent discriminants car c'est ce facteur qui détermine la qualité de l'étiquetage.

À titre d'illustration, nous donnons en annexe C page 178 le jeu de catégories distributionnel que nous utilisons.

2.4 Le token

2.4.1 Du mot au token

Le mot, qui exprime une segmentation conventionnelle et instable entre des langues différentes, et évolutive à l'intérieur d'une langue, n'est pas le bon candidat pour être l'unité minimale d'analyse d'un traitement de langue. On lui préfère le «token», une unité qui vient se substituer systématiquement à chaque portion de texte pour en permettre la manipulation.

Le token prend naissance lors de la transformation du texte original, vu comme une suite de caractères, en une suite d'unités minimales manipulables par le système : les tokens, précisément. En cela, ni le segment de texte substitué par un token, ni le token lui-même ne peuvent être définis dans l'absolu. La définition du token est toujours relative à la représentation des connaissances qui a été adoptée lors de la conception du système, en rapport avec les objectifs du traitement visé.

Le token est une unité générique qui permet de représenter aussi bien des mots que des ponctuations, des acronymes que des symboles d'unités monétaires. Par ailleurs, lors de la tokenisation (i.e., l'opération de production des tokens), un même mot peut donner naissance à plusieurs tokens (cas des amalgames, des mots composés), et plusieurs mots consécutifs peuvent être représentés par un seul token : locutions diverses, noms propres composés, numéraux composés, mots composés.

2.4.2 Principe de constitution des tokens

Il faut toujours avoir à l'esprit que la définition du token n'est pas étrangère à la qualité de l'étiquetage car elle contribue à l'identification de contextes distributionnels. En substituant un token particulier à un segment de l'énoncé, le token peut faciliter ou bien compliquer les traitements distributionnels.

Lorsque l'on décide de substituer, par exemple, à des amalgames tels que «*du*» et «*au*», les séquences de tokens *de* + *le* et *à* + *le*, le contexte distributionnel naturel se trouve perturbé, et l'étiquetage peut s'en trouver affecté. En effet, en français, «*du*» et «*au*» marquent toujours le début d'un syntagme nominal et «*de le*» et «*à le*» toujours le début d'un syntagme verbal. Une fois la substitution des amalgames effectuée, les tokens *de* + *le* et *à* + *le* peuvent aussi bien introduire un syntagme nominal qu'un syntagme verbal.

Il en est de même pour les pronoms clitiques postposés. Si l'on tient à détacher du verbe un pronom clitique postposé, il faut le faire de manière à ne pas perdre le contexte distributionnel : pour cela, il faut préserver la marque du trait d'union qui fait partie intégrante du contexte distributionnel. On remplacera donc «*dit-il*» ni par la séquence de tokens *dit* + *il*, ni même par la séquence *dit* + - + *il* (un tiret entre deux mots n'est pas forcément un trait d'union) qui fausse le contexte distributionnel.

Une solution consiste à effectuer une segmentation qui préserve le contexte distributionnel en gardant trace des opérations réalisées. Cela se traduit couramment par la production d'une séquence de tokens dont certains, plus ou moins artificiels, n'introduisent aucune ambiguïté artificielle : par exemple *dit* + *-il* où le token unique *-il* est considéré pronom clitique postposé ou bien *dit* + #-# + *il* où le token unique #-# symbolise le trait d'union.

La mauvaise définition d'un token peut bien entendu générer de faux contextes distributionnels. Par exemple, le regroupement systématique de «*avant de*» pourrait induire à tort que la locution peut avoir un contexte droit ou bien nominal («*l'avant de la voiture*») ou bien verbal («*avant de partir*»). À l'inverse, en regroupant des expressions composées telles que les locutions sûres ou bien les numéraux composés, le regroupement facilite le repérage de contextes distributionnels.

2.5 Les déductions contextuelles

2.5.1 Champ d'action des déductions contextuelles

Nous avons rappelé à maintes reprises que l'ordre des syntagmes minimaux dans la phrase était soumis à des contraintes relativement relâchées, comparées à celles agissant sur les mots internes aux syntagmes minimaux. De ce fait, on ne peut considérer la phrase comme une suite continue, indifférenciée de mots, dans laquelle toutes les contiguïtés seraient équivalentes. On se doit de distinguer une contiguïté de deux mots appartenant à un même syntagme minimal, et une contiguïté de deux mots situés de part et d'autre de la frontière entre deux syntagmes minimaux contigus.

Puisque le principe de l'étiquetage est d'utiliser les régularités de contiguïtés de catégories pour réaliser des déductions, les seules déductions fiables sont celles qui ont un contexte restreint à un seul syntagme minimal, et qui n'utilisent que ses contraintes internes très fortes (i.e., ordre des mots, accord entre les mots). Invoquer des informations en dehors de ce cadre n'est pas fiable et revient à violer les propriétés des langues. Pour poursuivre cette analyse :

> Lorsque le contexte local défini par un syntagme minimal contient trop peu d'informations sur lesquelles s'appuyer, certaines ambi-

guïtés ne doivent alors pas être levées par des techniques relevant de l'étiquetage car elles sortent de son champ d'action théorique.

Le défaut de la plupart des étiqueteurs actuels est de ne pas inclure assez de connaissances linguistiques et surtout de sous-estimer l'importance du critère positionnel lors de l'extraction et de l'utilisation des connaissances. La phrase est souvent perçue comme une suite continue de mots, dans laquelle on ignore que des mots contigus situés dans des syntagmes minimaux différents ne sont pas supposés avoir une relation positionnelle directe. Essayer alors de capturer des régularités dans ce type de situation est risqué et bien souvent sans réel fondement. Étudions au travers de cas concrets les conséquences sur le champ d'action des déductions contextuelles.

2.5.2 Classification des déductions contextuelles

Dans la section précédente, nous avons vu que les seules déductions fiables étaient celles qui avaient un contexte restreint à un seul syntagme minimal. Comme le syntagme minimal est constitué d'un élément central lexical (généralement un nom ou un verbe), entouré d'éléments périphériques grammaticaux (prépositions, déterminants, clitiques, adverbes, ...), nous pouvons dégager un certain nombre de conséquences sur la nature des déductions contextuelles.

Les déductions fiables : première caractérisation

Pour le français, nous notons trois principaux types de déductions contextuelles fiables, toutes internes au syntagme minimal :

1° d'un mot grammatical vers un autre mot grammatical :

$je_{pp}\ le_{det|po}\ bois \longrightarrow je_{pp}\ le_{po}\ bois$

2° d'un mot grammatical vers un mot lexical (on peut considérer cette déduction comme étant la déduction contextuelle canonique) :

$je\ le_{po}\ bois_{n|vb} \longrightarrow je\ le_{po}\ bois_{vb}$.

3° d'un mot lexical vers un autre mot lexical :

$petits_{p}\ radis_{s|p} \longrightarrow petits_{p}\ radis_{p}$

Avec pour notation : det = déterminant, po = pronom objet direct, pp = pronom personnel, vb = verbe, n = nom, s = singulier, p = pluriel.

Il est intéressant de noter que les deux premiers types de déduction utilisent l'ordre des mots relativement fixe à l'intérieur du syntagme minimal et que le troisième type de déduction utilise la propagation de l'accord, c'est-à-dire les deux caractéristiques qui nous ont amené au concept de syntagme minimal.

Les déductions fiables : seconde caractérisation

Dans notre première caractérisation, les déductions fiables sont exprimées en fonction de mots-source et de mots-cible. En restant au niveau des mots, ces déductions s'inscrivent complètement dans l'esprit de l'étiquetage. Nous allons à présent montrer que l'on peut envisager d'impliquer beaucoup plus le syntagme minimal dans cette déduction.

Auparavant, nous avons insisté sur la nécessité de contraindre les mots-source et les mots-cible d'une déduction à appartenir au même syntagme minimal. Si l'on tient compte du fait qu'un mot grammatical marque le plus souvent (en français) le début d'un syntagme minimal et qu'il détermine son type, alors la déduction peut être exprimée de manière indirecte en passant par le typage du syntagme minimal. Ces déductions, d'une autre nature que les premières, s'effectuent en deux temps :

1° faire hériter le syntagme de propriétés de certains mots-source internes ;

2° utiliser les caractéristiques du syntagme pour contraindre les mots qui le constituent.

Concrètement :
dans «*une ferme*» : si une_{det} ⇒ [*une ferme*] nominal ⇒ $ferme_{nom}$
dans «*ne ferme*» : si ne_{neg} ⇒ [*ne ferme*] verbal ⇒ $ferme_{vb}$
dans «*je le bois*» : si je_{ps} ⇒ [*je le bois*] verbal ⇒ le_{po} $bois_{vb}$

Les déductions non fiables

Une fois les déductions fiables caractérisées, nous pouvons maintenant nous intéresser aux cas pour lesquels il est impossible de faire une déduction locale et interne au syntagme minimal : c'est par exemple le cas d'un syntagme minimal constitué d'un seul mot lexical.

Cette déduction locale impossible ne porte pas à conséquence si ce mot n'a qu'une seule catégorie possible. Dans l'exemple 1, «*remporte*» constitue à lui seul un syntagme minimal, et ne peut être que verbe :

> « *Comme prévu, M. Museveni remporte la quasi-totalité des votes dans l'Ouest, ...* » (1)

Mais, si ce mot a plusieurs catégories possibles, alors la décision ne peut pas être prise par déduction contextuelle locale : la décision ne pourra être prise que par la mise en relation du syntagme minimal de type multiple avec un syntagme minimal de type connu. Dans l'exemple 2, «*montre*» constitue à lui seul un syntagme minimal, et peut être verbe ou nom ; la décision ne peut pas être prise localement, mais par le calcul de la relation sujet-verbe qu'il entretient avec *La présence* :

> « *La présence de Florence Arthaud au milieu d'un plateau de spécialistes montre que cette Transat a été la course la plus disputée de ces dix dernières années.* » (2)

Une autre situation pour laquelle la déduction contextuelle au niveau des mots est également malvenue touche les syntagmes comportant plusieurs mots dont la polycatégorie lexicale ne permet de lever aucune ambiguïté. Dans l'exemple 3, «*la porte*» constitue un syntagme minimal mais la polycatégorie de «*la*» (déterminant/pronom objet) et de «*porte*» (nom/verbe) ne permet pas de savoir par une déduction contextuelle fiable au niveau des mots si le syntagme est nominal ou verbal.

> « *Bérégovoy a ouvert la porte tout en agitant la menace d'une grève générale.* » (3)

En conclusion, seuls les trois premiers types de déductions exposés au début de cette section sont fiables car ils entrent dans le champ d'action de l'étiquetage grammatical. Les autres cas sortent de ce champ d'action car ils sont fondés sur une relation entre deux syntagmes. Chercher à les résoudre par des déductions contextuelles au niveau des mots peut conduire à des erreurs : il est alors préférable de marquer l'ambiguïté sur le syntagme minimal et d'envisager d'autres techniques de résolution. C'est vers un processus exploitant les propriétés des syntagmes minimaux qu'il faut se tourner.

2.6 Les ressources lexicales

2.6.1 De l'incomplétude des lexiques

Les ressources lexicales sont par nature incomplètes. Doublement incomplètes, devrait-on dire. Une première source d'incomplétude est due aux nombreuses occurrences, dans les textes, de mots inconnus des lexiques classiques : les mots d'emprunt, les mots spécifiques à un domaine, les néologismes, les mots erronés (CHANOD et TAPANAINEN, 1995a). La seconde source d'incomplétude est à attribuer au fait qu'une ou plusieurs descriptions morpho-syntaxiques peuvent être manquantes pour chaque mot du lexique. Un étiqueteur doit donc savoir gérer correctement des ressources lexicales partielles.

Le recours à des analyseurs morphologiques (i.e., «*guesser*») est aujourd'hui courant pour prendre en charge les mots inconnus (e.g., un mot terminé par *tion* est un nom féminin). Il faut cependant noter que ce type d'outil ne pallie que partiellement la première source d'incomplétude car l'outil n'est utile que si le mot est absent du lexique ; il peut par ailleurs introduire des descriptions morpho-syntaxiques non pertinentes et n'est vraiment efficace que sur les langues à forte morphologie.

Pour étiqueter correctement, il doit y avoir synergie entre ressources lexicales et déductions contextuelles, ces dernières devant intégrer le fait que les ressources lexicales sont des sources de connaissances partielles, le monde analysé n'étant pas un monde clos et le corpus réservant toujours des surprises. Étudions différentes articulations entre ces connaissances.

2.6.2 Articulations entre lexiques et déductions contextuelles

La désambiguïsation

Voyons comment l'étiquetage est habituellement posé : pour chaque token, toutes ses descriptions morpho-syntaxiques sont tout d'abord exhaustivement énumérées à partir de sources d'informations lexicales (lexique de lemmes ou de formes, «guesser»), puis des contraintes contextuelles éliminent des choix en fonction de contiguïtés impossibles, c'est le cas des «grammaires de contraintes» (Karlsson, 1990; Karlsson et al., 1994; Voutilainen et al., 1992; Voutilainen, 1994; Chanod et Tapanainen, 1994). Exemple : la catégorie «verbe» *ne suit jamais* la catégorie «déterminant». L'homographie polycatégorielle étant couramment appelée «ambiguïté», ce processus d'étiquetage est souvent appelé «désambiguïsation», vue comme une annulation ou une diminution de l'«ambiguïté».

Le principal défaut d'une déduction négative est que la description morpho-syntaxique attendue du token doit appartenir à la liste des descriptions extraites des ressources lexicales, ce qui revient à faire l'hypothèse que tout token appartient au lexique et est muni de la liste exhaustive de ses descriptions morpho-syntaxiques possibles, hypothèse manifestement fausse, même en présence d'un «guesser».

L'assignation

Une deuxième articulation entre ressources lexicales et déduction contextuelles consiste à faire usage de déductions contextuelles affirmatives. Selon cette approche, les déductions assignent aux tokens des descriptions morpho-syntaxiques qui sont sûres dans leur contexte local (Constant, 1991; Vergne, 1994; Giguet et Vergne, 1997a). Exemple : si le mot qui suit un pronom personnel sujet n'est pas un clitique *alors c'est* un verbe.

Les approches par déductions négatives et par déductions affirmatives peuvent *a priori* sembler équivalentes car elles tentent de converger toutes les deux vers une même solution, la première de manière destructive et la seconde de manière constructive. L'équivalence serait garantie si l'univers dans lequel est effectué l'étiquetage était clos, ce qu'invalide la double incomplétude du lexique.

Alors que l'écriture d'une déduction négative part du principe que la description attendue du token doit appartenir à la liste des descriptions exhaustivement énumérées, la déduction affirmative a, elle, l'avantage de ne pas présupposer la complétude du lexique. Elle peut ainsi s'attaquer aux deux sources d'incomplétude en établissant une supériorité de la connaissance contextuelle sur la connaissance lexicale. Exemple : une déduction telle que «si le mot qui suit un pronom personnel sujet n'est pas un clitique alors c'est un verbe» peut gérer des phrases telles que «*je positive*», même si *positive* n'est inscrit qu'en tant qu'adjectif dans le lexique (le verbe *positiver* est un néologisme et est absent de la plupart des lexiques). On laisse ainsi sa place à la créativité dans la langue.

Nous pouvons voir cette articulation de manière plus générale en la faisant reposer sur le concept de syntagme minimal. Nous savons que l'ordre des éléments dans un syntagme est relativement fixe. Il est donc possible, une fois le syntagme construit et typé de déduire la catégorie d'un mot par simple utilisation du critère positionnel. L'usage de contraintes affirmatives n'est qu'une mise en œuvre particulière et partielle de ce principe général.

L'attribution de valeurs par défaut

On remarque que les différentes catégories possibles d'un token sont loin d'être équiprobables : en général une catégorie est de beaucoup la plus fréquente (VERGNE et GIGUET, 1998). Prenons l'exemple caricatural de *le*, *l'*, *la* et *les* : dans un corpus[1] de 10 687 mots du journal Le Monde : 1054 occurrences se répartissent en 1029 déterminants (97,6%), et 25 pronoms clitiques objet (2,4%).

Au lieu de poser au départ des déductions que ces graphies peuvent être déterminants ou pronoms, posons qu'elles sont déterminants *par défaut* (ces informations par défaut sont codées dans le lexique), et qu'elles seront pronoms dans un contexte particulier (ces informations liées au contexte sont codées dans des règles de déduction contextuelle) :

pronom clitique sujet suivi de *le*, *l'*, *la* et *les*
$\Rightarrow$ *le*, *l'*, *la* et *les* pronoms clitiques objet
négation ne suivie de *le*, *l'*, *la* et *les*
$\Rightarrow$ *le*, *l'*, *la* et *les* pronoms clitiques objet
le, *l'*, *la* et *les* suivi d'un verbe sûr
$\Rightarrow$ *le*, *l'*, *la* et *les* pronoms clitiques objet

1. Il s'agit du corpus des essais de l'action GRACE, étiqueté à la main par Josette LECOMTE du comité GRACE.

On trouvera une démarche analogue dans (Chanod et Tapanainen, 1995b, page 151, §4.2.2), mais restreinte à certains mots grammaticaux qui ont une homographie très rare avec un mot lexical (*est*, *cela*, *avions*), homographie détectée à l'aide du contexte.

Donner une catégorie par défaut dans le lexique, et la modifier éventuellement par le contexte constitue en quelque sorte une implémentation du concept de «translation du premier degré» de Tesnière (1959, à partir de la page 361). Un exemple généralisé est donné dans SYLEX, l'analyseur de Patrick Constant (1991).

2.7 Le processus en tant que système

Les sections précédentes suggèrent que les traits morpho-syntaxiques, les tokens, les déductions contextuelles et les ressources lexicales, sont tous interdépendants. Ils forment *un système*, au sens propre du terme.

Pour en prendre pleinement conscience, étudions comment se traduit la gestion d'un phénomène tel que la postposition du pronom clitique. Tout commence par l'identification de son contexte distributionnel : dans le corpus, un verbe précédant un pronom clitique est toujours lié à ce pronom par un trait d'union. Sachant que les homographies verbo-nominales et verbo-adjectivales sont nombreuses et que leur résolution en contexte est déterminante, une déduction contextuelle pertinente est donc : «lorsque le système détecte un pronom clitique postposé alors le mot qui le précède est un verbe». Pour être en mesure de réaliser une telle déduction, le système doit (1) intégrer le concept «pronom clitique postposé», (2) définir les éléments appartenant à cette classe distributionnelle et (3) permettre aux déductions contextuelles d'exploiter le contexte distributionnel.

L'intégration du concept «pronom clitique postposé» passe par la définition d'une nouvelle classe distributionnelle dans le jeu des catégories syntaxiques : celle des pronoms clitiques postposés. Les éléments appartenant à cette classe sont introduits dans le lexique où leur est associée la catégorie «pronom clitique postposé». La définition de ces nouvelles unités lexicales dépend de la manière dont la segmentation révèle le contexte distributionnel : si le segmenteur produit une séquence telle que *dit* + *-il* alors toutes les séquences du type *-il*, *-elle* doivent être ajoutées au lexique, si le segmenteur choisit par contre de produire la séquence *dit* #-# *il* alors il faut ajouter à tous les pronoms clitiques postposables du lexique (e.g., *il*, *elle*) leur appartenance à la classe des pronoms clitiques postposés.

Cette illustration montre clairement que toutes les ressources linguistiques sont à concevoir simultanément, dans un même esprit et suivant un même principe : le respect du critère distributionnel. Le concept de système est donc tout à fait approprié. En filigrane des principes de conception du jeu de catégories syntaxiques, des tokens, des déductions contextuelles et des ressources lexicales, il faut noter la présence récurrente du syntagme minimal qui définit en quelque sorte le cadre distributionnel de l'étiquetage.

2.8 Ouverture sur l'étiquetage

Nous avons préalablement vu que la délimitation des syntagmes ne nécessitait pas un étiquetage très raffiné et qu'un simple étiquetage des mots grammaticaux était bien souvent suffisant. Le typage d'un syntagme quant à lui se déduit naturellement des étiquettes des mots qui composent le syntagme. Dans notre implémentation actuelle, une quinzaine de déductions contextuelles suffisent à obtenir une délimitation et un typage des syntagmes de bonne qualité.

Notre objectif est cependant l'obtention d'une analyse syntaxique précise. Un étiquetage superficiel n'est donc pas suffisant et nous sommes amené à rechercher un étiquetage de qualité supérieure. Exposons comment s'est organisé le système en fonction de cet objectif.

La possibilité de délimiter des syntagmes par un étiquetage vraiment minimal est un atout majeur qui ouvre les portes d'un étiquetage de précision. L'étude du champ d'action des déductions contextuelles nous a en effet permis de caractériser les déductions fiables : ce sont celles qui sont réalisées à l'intérieur d'un syntagme. De cette analyse, nous concluons qu'il ne faut surtout pas chercher à affiner l'étiquetage avant d'avoir posé les limites de syntagmes. Par contre, une fois ces limites posées, l'unité est délimitée et nous pouvons alors nous occuper de ses parties en toute confiance puisque les déductions contextuelles que nous y effectuerons seront fiables.

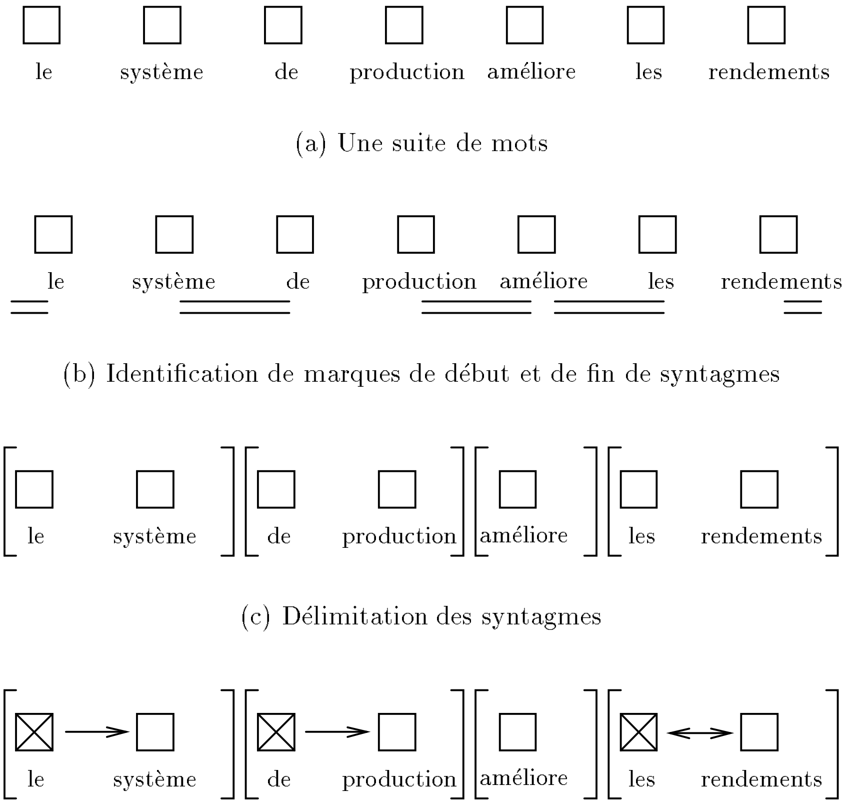
Notons qu'à ce stade, ce n'est pas l'étiquetage optimal qui est visé. Nous avons montré que l'étiquetage parfait était un objectif qui ne pouvait être atteint que par un système d'analyse syntaxique dans tout son ensemble. L'étiquetage que nous effectuons se restreint à l'utilisation des informations contextuelles qu'offre le syntagme. Nous laisserons au processus de mise en relation des syntagmes le soin d'affiner cet étiquetage.

Il est important à ce stade de noter d'une part l'importance de savoir délimiter une unité pour s'occuper correctement de ses parties, et d'autre part de savoir accepter que l'imperfection de la structure interne d'une unité n'est pas une entrave à sa manipulation à un niveau supérieur : bien au contraire, c'est en manipulant l'unité et en cherchant comment elle s'intègre dans son environnement que l'on obtiendra de nouvelles connaissances sur cette unité et que l'on pourra par conséquent affiner les connaissances sur les éléments qui la constituent.

2.9 Regard sur la méthode

Sachant que l'unité d'entrée est une suite de mots et que le champ d'action théorique des déductions contextuelles est le syntagme, nous partitionnons la suite de mots en syntagmes de manière à faire apparaître les domaines syntagmatiques au sein desquels les déductions contextuelles sont fiables. L'obtention de cette partition passe par l'identification de marques de début et de fin de syntagmes. Une fois cette partition établie, nous utilisons la position à l'intérieur du syntagme pour y appliquer les déductions contextuelles pertinentes et ainsi étiqueter les mots. Cette méthode est schématisée figure 2.1 page ci-contre. Notons que les trois opérations (b), (c) et (d) sont exécutées simultanément, en une seule passe, de la gauche vers la droite, le texte étant considéré comme un flux.

Après étude du champ d'action théorique des déductions contextuelles, nous remarquons que l'étiquetage de certains mots n'est pas possible par ce type de technique. Cette situation se présente lorsque le contexte défini par le syntagme n'est pas assez informatif, c'est-à-dire lorsque les éléments internes au syntagme ne suffisent pas à le catégoriser. La catégorie du syntagme peut en effet être définie de deux manières. D'une part à l'aide d'indices internes, en utilisant des catégories syntaxiques discriminantes de mots internes au syntagme (e.g., «*dans*» caractérise un syntagme nominal) et des relations intra-syntagmatiques discriminantes («*de le*» caractérise un syntagme verbal). D'autre part à l'aide d'indices externes, en utilisant la manière dont le syntagme est intégré dans son environnement (e.g., ci-dessous, «*[améliore]*» est catégorisé verbal par sa position). L'utilisation d'indices externes permettant la catégorisation des syntagmes et nécessitant la reconnaissance de l'unité supérieure au syntagme sera traitée dans le chapitre suivant.



(a) Une suite de mots

(b) Identification de marques de début et de fin de syntagmes

(c) Délimitation des syntagmes

(d) Utilisation du critère positionnel dans le syntagme

FIG. 2.1 - *Partition en syntagmes pour une application pertinente des déductions*

2.10 Conclusion

L'étiquetage morpho-syntaxique est un processus qui permet l'organisation de la structure interne d'une unité : le syntagme. Pour organiser une unité, il faut à la fois reconnaître, accepter, l'existence de l'unité et savoir la délimiter. En ces deux points, nous nous plaçons en marge des recherches classiques sur l'étiquetage qui utilisent certes, un processus d'étiquetage, mais sur une suite de mots indifférenciés, c'est-à-dire sans utiliser le concept de syntagme, ou en sous-estimant son importance.

Reconnaître l'existence du syntagme, c'est s'ouvrir à sa délimitation, une opération à la fois simple et fondamentale. Simple car basée sur des indices formels facilement identifiables : le repérage de mots marquant les débuts et les fins de syntagmes, le repérage de ruptures d'accord. Mais fondamentale car une fois l'unité délimitée, il est alors possible de se repérer en son sein et d'utiliser le critère positionnel pour l'organiser correctement. En fait, la délimitation des syntagmes permet de vérifier la similarité du contexte dans lequel une connaissance formelle a été acquise et du contexte dans lequel cette connaissance est appliquée.

La propagation de contraintes contextuelles à l'intérieur d'un syntagme est un processus tout à fait pertinent car l'ordre des mots dans cette unité est relativement fixe. Le processus d'étiquetage fournit les bases de la structuration interne des syntagmes : la catégorisation morpho-syntaxique en contexte des mots qui les composent. L'étape suivante de la structuration consiste à faire apparaître les relations syntaxiques entre les mots d'un même syntagme. Ces relations sont déductibles de la seule position de chacun des éléments dans la structure. Nous ne les calculons pas pour le moment.

La délimitation des syntagmes ne suffit en fait pas pour permettre leur structuration précise. Un syntagme ne prend en effet réellement existence que lorsqu'il est plongé dans son environnement : la phrase. Chercher à structurer un syntagme en l'observant uniquement de l'intérieur est insuffisant. Il faut savoir l'étudier de l'extérieur, dans son contexte, pour affiner les connaissances sur sa structure. La catégorisation des mots pouvant être amenée à s'affiner, le calcul des relations internes au syntagme peut être retardé et le système dans son ensemble tirera profit de cette remise en contexte. L'étude du syntagme dans l'énoncé est précisément l'objet du chapitre suivant.

Ces travaux ont fait l'objet de deux publications : (Giguet et Vergne, 1997a) et (Vergne et Giguet, 1998). Enfin, tous les principes décrits dans ce chapitre sont empiriquement validés par la première place qu'occupe notre système dans l'action GRACE (1997) d'évaluation des étiqueteurs morpho-syntaxiques du français.

Chapitre 3

La mise en relation des syntagmes

3.1 Cadre linguistique

Le cadre linguistique de la mise en relation des syntagmes minimaux s'inspire des travaux de Lucien TESNIÈRE (1959). À partir de sa définition perceptive et calculatoire de la connexion (page 11) :

> «*Tout mot qui fait partie d'une phrase cesse par lui-même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des* ***connexions****, dont l'ensemble forme la charpente de la phrase.*»

nous proposons de ne plus considérer la relation de manière statique, restreinte à sa dimension spatiale, mais de l'envisager de manière dynamique en introduisant une dimension supplémentaire : la dimension chronologique. Nous proposons par ailleurs de concevoir un processus ne se restreignant pas au calcul de la seule relation de dépendance mais de l'ouvrir à d'autres types de relations telles que la coordination, l'antécédence.

Alors que TESNIÈRE définissait la connexion entre les mots, nous avons revu cette notion de sorte qu'elle relie non plus des mots mais des syntagmes minimaux. Cette caractéristique différencie notre approche de celles des autres travaux en analyse syntaxique automatique fondée sur le modèle de la dépendance (COVINGTON, 1990; SLEATOR et TEMPERLEY, 1993; TAPANAINEN et JÄRVINEN, 1997). Comme le note Steven ABNEY (1996) :

> «*By reducing the sentence to chunks* [i.e., syntagmes minimaux], *there are fewer units whose associations must be considered, and we can have more confidence that the pairs being considered actually stand in the syntactic relation of interest, rather than being random pairs of words that happen to appear near each other.*»

Comme le rappellent GIGUET (1997) et VERGNE (1998) :

> *Dans notre démarche, constituance et dépendance ne sont déclarées ni équivalentes, ni opposées, mais utilisées ensemble à deux niveaux différents : constituance à l'intérieur des SR,* [i.e., syntagmes minimaux] *dépendance entre SR.*

3.2 Genèse du processus d'analyse

3.2.1 Des contraintes sur les unités à relier...

Les formalismes basés sur les grammaires de dépendances autorisent généralement la spécification des relations en définissant :

1° des contraintes sur la catégorie des deux unités intervenant dans la relation considérée (e.g., une unité nominale et une unité verbale pour la relation sujet-verbe) ;

2° des contraintes morphologiques entre les deux unités (e.g., accord en personne, en genre, en nombre).

Ces connaissances formelles ne suffisent cependant pas à la description des relations linguistiques : beaucoup de couples d'unités peuvent dans une phrase satisfaire ces contraintes sans qu'ils soient pour autant liés par une relation fonctionnelle (une unité nominale et une unité verbale accordées en personne et en nombre ne sont pas obligatoirement liées par une relation sujet-verbe).

Tenter de construire un analyseur n'utilisant que ces connaissances est rédhibitoire car soit les contraintes entre les deux unités à relier s'avèrent trop lâches et le bruit est insupportable (le nombre de candidats potentiels trop grand pour chaque relation entraîne une surgénération de solutions incohérentes), soit les contraintes sont trop strictes et là, inversement, le silence le plus total se fait entendre. Si l'on ajoute à ce phénomène le fait que la polycatégorisation morpho-syntaxique des différentes unités syntaxiques à relier est massive, la situation devient alors inextricable.

Ces contraintes sur la structure des unités à relier étant incontournables, un processus d'analyse fondé sur la dépendance doit avoir recours à d'autres connaissances pour devenir maîtrisable. Plusieurs solutions sont envisageables pour enrichir la description linguistique des relations :

- laisser la possibilité d'introduire des obligations ou interdictions d'occurrences d'unités entre les éléments reliés ;

- permettre la définition de la position relative des unités reliées ;

– offrir même le paramétrage d'une distance maximale entre les éléments reliés.

Une kyrielle de propositions que l'on peut toujours prouver inadéquates au traitement de corpus.

3.2.2 ... aux contraintes sur les relations

Il faut bien comprendre que ce ne sont pas les éléments en soi qui définissent et déterminent la relation : une relation n'a une existence réelle qu'au travers des autres relations ; l'existence d'une relation est liée à celle des autres relations (HAGÈGE, 1982).

Ce n'est donc évidemment pas en n'agissant que sur les unités à relier que l'on peut assurer la cohérence globale des analyses proposées. C'est en étudiant ce qu'une relation apporte à la structure relationnelle et ce que les composantes de la structure relationnelle lui apportent. Nous postulons que telle est la voie à suivre pour garantir la cohérence globale des analyses proposées et la calculabilité de la structure relationnelle.

La cohérence globale est garantie si toutes les relations prises individuellement sont compatibles avec toutes les autres relations calculées. Pour atteindre cet objectif, il faut préserver systématiquement et à chaque nouvelle instanciation de relations, la cohérence du graphe de relations construit. C'est donc tout naturellement vers l'étude des contraintes relationnelles que nous nous dirigeons.

Dans les travaux en linguistique de MEL'ČUK (1988), nous trouvons au travers du rappel de la définition de *projectivité* les prémisses des concepts de contraintes relationnelles. La projectivité caractérise la structure relationnelle de tout un ensemble de phrases dans le cadre de la théorie des grammaires de dépendance. La projectivité est définie de la sorte : l'arbre de dépendance d'une phrase est projectif s'il est *planaire* dans le demi-plan délimité par les mots de la phrase (i.e., les relations ne se croisent pas lorsqu'elles sont dessinées au dessus des mots) et si aucune relation ne couvre la racine de l'arbre.

La projectivité peut se concevoir comme le résultat de l'application de contraintes relationnelles entre les unités, dans le cadre des grammaires de dépendance. Il est d'ailleurs très tentant d'en dériver un modèle de traitement suivant la voie que nous indiquions : toute instanciation de relation est contrainte par les relations précédemment créées car elle ne doit pas les croiser, et une fois créée, une nouvelle relation va contraindre celles à venir. La projectivité n'a cependant pas pour ambition de rendre compte de toutes les phrases attestées et ne peut en aucun cas servir de modèle unique à un traitement approprié de la syntaxe, comme le suggèrent les articles de Tapanainen et Järvinen (1997) et de Giguet et Vergne (1997b). La projectivité permet simplement d'établir une classification. Plusieurs s'y sont pourtant mépris, par exemple Sleator et Temperley (1993).

3.2.3 Prise en compte des contraintes relationnelles

Alors que les contraintes sur la structure des unités à relier sont aujourd'hui bien caractérisées, les contraintes relationnelles restent, elles, peu étudiées. La raison en est simple : l'analyse de la structure des éléments à relier peut s'effectuer de manière statique et ne fait intervenir que la dimension spatiale de l'objet ; l'analyse des contraintes relationnelles nécessite quant à elle une étude de la dynamique du processus de construction de la structure relationnelle. Elle requiert donc la prise en compte d'une dimension supplémentaire : la dimension chronologique.

Nous postulons que les contraintes relationnelles sont en partie liées à l'activité dynamique de l'interprétant sur l'objet, lors de la construction incrémentale de sa structure :

- le processus d'analyse est guidé par les attentes (entre autres syntaxiques) suscitées par le caractère non définitif de la structure qu'il crée ;

- l'arrivée d'une nouvelle unité va entraîner son raccrochement à la structure en satisfaisant une de ces attentes, ou en déclenchant préalablement sa restructuration ;

- la nouvelle unité, une fois introduite, va supprimer certaines attentes et en susciter de nouvelles.

Pour mieux comprendre ce mécanisme, nous présentons l'explicitation de l'interaction entre les unités syntaxiques d'un énoncé et la structure relationnelle qui les intègre. L'étude considère les deux temps de l'interaction suivants : (1) la manière dont le rattachement d'une unité est conditionné par la structure relationnelle en construction et (2) la manière dont ce rattachement va conditionner via la structure relationnelle tous les rattachements à venir.

Côté psycholinguistique, les concepts de création d'attentes et de suppression d'attentes sont analogues aux concepts d'attentes et d'oublis (c.f., GRUNIG, 1993, § Suspension). Notre approche rejoint d'ailleurs à plusieurs titres les travaux de Blanche-Noëlle GRUNIG : on se reportera aux relations entre la charge mémorielle et les prédictions syntaxiques (GRUNIG, 1993) puis à sa conception dynamique du contexte (GRUNIG, 1995).

Côté traitement des langues, nous trouvons des approches très similaires à la nôtre dans des domaines parallèles. Nous commencerons par mentionner les travaux de Bernard VICTORRI (1998) qui propose sur des bases identiques une construction dynamique de la sémantique. Nous citons également les travaux de Michel DUPONT (1997) sur le traitement automatique de la référence à l'aide d'un «Modèle des Attentes du Lecteur». Ces travaux émergeant simultanément et dans des univers variés du traitement des langues annoncent très certainement l'aube d'un nouveau courant.

3.3 Description du processus d'analyse

Afin de valider notre hypothèse, nous avons modélisé ce processus de mise en relation. Nous allons ici en exposer les concepts et principes.

Le principe général est de traiter tous les syntagmes successivement et de procéder à leur intégration dans un graphe de relations qui s'étoffe ainsi progressivement. À chaque intégration de syntagme vont correspondre des créations de relations, des attentes, des oublis qui vont permettre le maintien de la cohérence globale du graphe.

3.3.1 Intégration d'un syntagme

L'intégration d'un syntagme dans le graphe des relations correspond au moment de sa prise en charge par la boucle de calcul. Le rattachement d'un syntagme à d'autres syntagmes préalablement traités ou futurs suit le processus général suivant :

- le processus construit tout d'abord et, dans la mesure du possible, la ou les relations du syntagme courant au graphe de relations en cours de construction. Cette opération dépend :

  1° de ce que le graphe autorise comme extensions cohérentes ;

  2° de la compatibilité du syntagme courant avec les extensions accessibles.

- le processus complète ensuite éventuellement le graphe en instanciant de nouvelles relations déduites des relations qu'il vient de créer entre le syntagme courant et les autres nœuds du graphe.

- il définit ensuite les nouvelles possibilités d'extensions du graphe de relations :

  1° en ajoutant des contraintes relationnelles qui assureront la cohérence future du graphe : certaines extensions encore valides avant l'ajout des nouvelles relations sont désormais supprimées car incompatibles au vu de la nouvelle structure du graphe ;

  2° en précisant les nouvelles extensions possibles à partir du nœud courant et nées de l'introduction de ce syntagme dans le graphe.

3.3.2 Instanciation d'une relation

L'instanciation d'une relation est donc le fruit des événements chronologiques suivants :

1° un syntagme est ajouté au graphe et se propose comme point d'extension pour une relation d'un type particulier ;

2° les modifications successives du graphe de relations nées de l'intégration des syntagmes suivants ne génèrent aucune contrainte venant invalider l'extension proposée ;

3° un syntagme apparaît alors et peut satisfaire l'extension proposée. L'instanciation de la relation est alors réalisée entre les deux syntagmes.

Étudions concrètement comment se traduit l'instanciation d'une relation par le processus d'analyse. Soit la phrase prédécoupée en syntagmes :

« *[Le système] [de production] [améliore] [les rendements].* » (4)

1° Le processus prend en charge le syntagme «*[Le système]*». Le syntagme étant nominal et ne pouvant être attaché à un quelconque élément, il est considéré comme sujet potentiel et devient un point d'extension du graphe pour la relation sujet-verbe.

2° L'intégration de «*[de production]*» à la structure en tant que subordonné à «*[Le système]*» n'empêche en aucun cas ce dernier de rester point d'extension de la relation sujet-verbe.

3° Le syntagme «*[améliore]*» apparaît. Le processus recherche alors un sujet compatible avec les propriétés morphologiques du syntagme verbal dans le graphe. Il trouve «*[Le système]*» et instancie la relation sujet-verbe entre les deux syntagmes.

Lorsqu'une relation est instanciée entre un nouveau syntagme et un syntagme défini comme point d'extension du graphe, l'attente est satisfaite. Reste alors à appliquer les contraintes relationnelles nées de cette nouvelle relation et à définir les nouveaux points d'extension.

3.3.3 Contraintes relationnelles

La génération de contraintes relationnelles résulte toujours de l'instanciation d'une nouvelle relation entre deux syntagmes, l'objectif étant de garantir que les futures relations resteront compatibles avec le graphe de relations actuel. À chaque instanciation, ces contraintes garantissent la cohérence globale du graphe de relations construit.

L'application d'une contrainte relationnelle consiste en la suppression d'une des extensions possibles du graphe de relations. Lorsqu'une extension d'un type particulier est supprimée, plus aucune relation de ce type n'est possible à partir du syntagme auquel était attachée l'extension. L'application d'une contrainte relationnelle correspond donc à la mort d'une relation qui fut pendant un temps envisageable mais que l'évolution de la structure du graphe a fini par invalider.

Dans la section précédente, nous avons implicitement vu l'application d'une contrainte relationnelle : lorsque nous avons réalisé l'instanciation d'une relation sujet-verbe entre «*[Le système]*» et «*[améliore]*», le point d'extension défini sur «*[Le système]*» est supprimé car l'attente s'est trouvée satisfaite. Si ce point d'extension n'est pas reconduit, plus aucune relation sujet-verbe ne pourra être créée avec «*[Le système]*» comme sujet.

Nous venons d'étudier un cas particulier d'application de contraintes. Voyons à présent comment ce cas particulier s'inscrit dans un processus général d'application des contraintes relationnelles et permettant une extension contrôlée du graphe.

3.3.4 Extension contrôlée du graphe

Le processus général d'extension contrôlée du graphe modélise les hypothèses psycholinguistiques d'attentes et d'oublis que nous avons esquissées section 3.2.3 page 120.

Le processus général repose sur deux principes que l'on tente d'appliquer systématiquement à chaque intégration de syntagme :

1° (a) satisfaire une attente en instanciant une relation avec le syntagme, (b) oublier toutes les attentes situées entre le syntagme courant et le syntagme dont on vient de satisfaire l'attente, (c) définir des oublis sur le graphe ;

2° (a) inscrire le syntagme courant en attente de relations et (b) définir des attentes sur le graphe.

Pour comprendre parfaitement ce processus général, étudions-le en action en reprenant l'exemple 4 page précédente :

« *[Le système] [de production] [améliore] [les rendements].* »

– traitement de «*Le système*» :

1° le principe n° 1 n'est pas applicable car aucune attente n'existe ;

2° le principe n° 2(a) permet l'inscription du syntagme en tant que sujet potentiel, régissant potentiel et tête potentielle de coordonnée de syntagmes.

– traitement de «*de production*» :

1° le principe n° 1(a) permet de relier le syntagme à «*Le système*» qui est régissant potentiel. Le principe n° 1(b) est inapplicable car les syntagmes reliés sont contigus.

2° le principe n° 2(a) permet l'inscription du syntagme en tant que régissant potentiel et coordonné potentiel. (b) L'attente de «*Le système*» est satisfaite mais est naturellement renouvelée car «*Le système*» peut toujours régir.

– traitement de «*améliore*» :

1° le principe n° 1 permet (a) de relier le syntagme à «*Le système*» qui est sujet potentiel, (b) d'oublier toutes les attentes situées entre les deux syntagmes reliés, tous les points d'extension sont supprimés sur «*de production*», (c) «*Le système*» ne peut plus être tête potentielle de coordonnée de syntagmes.

2° le principe n° 2 (a) permet l'inscription du syntagme «*améliore*» en tant que régissant potentiel, coordonné potentiel, en attente d'objet potentiel. (b) d'inscrire «*Le système*» comme tête potentielle de coordonnée de propositions.

La construction du graphe se fait donc par l'interaction permanente des différentes relations au travers de la création d'attente et de la suppression d'attente par oubli ou satisfaction. Il est important de noter que les attentes sont toujours potentielles, c'est le traitement des syntagmes futurs qui décidera si elles seront ou non satisfaites. À la fin de la phrase, toutes les attentes non satisfaites sont effacées. Cela vient d'une hypothèse que toute phrase est complète (l'observation prouve que ce n'est pas toujours le cas).

3.3.5 Instanciation et extensions concurrentes

Deux syntagmes sont nécessaires à l'instanciation d'une relation. Le premier syntagme est le syntagme à relier, c'est normalement le syntagme courant. Le second syntagme est un syntagme défini comme point d'extension du graphe pour la relation considérée.

Bien entendu, dans le graphe de relations, plusieurs syntagmes peuvent être simultanément définis comme point d'extension pour une même relation. Par exemple, toujours dans la phrase 4 page 123, avant le traitement de «*améliore*», les syntagmes «*Le système*» et «*de production*» sont tous les deux régissants potentiels.

Au cas où une relation doit être établie et que plusieurs points d'extension sont concurrents, nous devons sélectionner celui qui convient le mieux en fonction du type de la relation instanciée, du graphe de relations courant et des propriétés du syntagme à relier.

Sur le plan bibliographique, il serait difficile de faire une liste exhaustive de tous les critères syntaxiques, sémantiques, psycholinguistiques qui ont été suggérés et bien entendu très discutés sur ce sujet au cours des temps. On peut se reporter à l'article critique de Wilks, Huang, et Fass (1985) pour en avoir un aperçu et constater combien dans ce domaine l'imagination n'est pas en manque.

N'ayant pour le moment pas étudié toutes les situations de concurrence, nous proposons d'une part que le choix des connaissances permettant la sélection des extensions concurrentes reste ouvert, c'est-à-dire que le système de sélection soit souple et incrémental, et d'autre part que les connaissances à utiliser puissent varier en fonction du type de la relation à instancier.

À titre d'illustration, la relation sujet-verbe nécessite des connaissances morpho-syntaxiques (i.e., personne, nombre et genre) pour sélectionner le point d'extension sujet à un syntagme conjoint verbal, la gestion de la coordination, quant à elle, exploite la similarité des structures à relier. Dans un cadre plus général, il est possible d'utiliser des critères d'ordre psycholinguistique pour restreindre la distance entre les deux éléments à relier ou bien pour favoriser l'attachement de dépendants dans le contexte gauche. Bien que nous n'en ayons pas étudié la faisabilité, certains auteurs allient des connaissances sémantiques pour effectuer le rattachement des groupes prépositionnels (WERMTER, 1989).

3.3.6 Délimitation des segments propositionnels

Les contraintes relationnelles ne constituent pas un arsenal suffisant pour garantir la cohérence globale du graphe de relations. Certaines relations n'existent en effet qu'au travers du concept de proposition (e.g., la relation sujet-verbe) alors que d'autres sont trans-propositionnelles (e.g., la relation d'antécédence).

À ce stade de notre présentation, nous ne pouvons garantir que les relations restreintes au domaine propositionnel soient effectivement et exclusivement construites dans ce domaine.

« *La base que la SSII développe accélérera les traitements.* » (5)

Dans la phrase 5, au moment du traitement du syntagme «*développe*», deux syntagmes nominaux «*La base*» et «*la SSII*» sont des points d'extension possible de relations sujet-verbe. Alors que les points d'extension devraient toujours garantir des extensions valides, on constate que l'extension définie sur «*La base*» bien que légitime n'est pas une extension valide dans le domaine défini par «*que la SSII développe*».

Pour résoudre proprement ce problème, la seule solution consiste à introduire la prise en compte du niveau propositionnel. C'est ce que nous faisons en cherchant à délimiter les propositions : lors de l'entrée dans une nouvelle proposition, une marque de début de proposition est définie. Elle permet d'ouvrir un domaine propositionnel au sein duquel les extensions du graphe situées dans des domaines propositionnels différents sont inaccessibles aux relations non trans-propositionnelles, le temps du traitement des syntagmes internes au domaine.

Dans l'exemple 5, lors du traitement de «*que*», nous définissons une marque de début de proposition qui a pour effet de masquer l'extension de la relation sujet-verbe définie sur «*La base*».

Ce mécanisme permet la construction de structures non projectives tout en gardant la maîtrise de la combinatoire : les relations de rattachement de subordonnées et d'antécédence étant trans-propositionnelles, elles ne sont pas affectées par la restriction de la visibilité des extensions situées hors du domaine propositionnel courant et les extensions de régissant potentiel et d'antécédent potentiel défini sur *La base* leur restent accessibles.

La délimitation des segments propositionnels repose sur la détection de marques formelles. Le début d'une proposition en français est relativement bien marqué et donne naissance à une marque de début de proposition très sûre. C'est ce type de marque que le processus pose sur «*La base*» et sur «*que*» dans la phrase 5 page précédente, lorsqu'il les rencontre.

La détection des fins de proposition est par contre moins évidente. Dans les cas où elle l'est, le processus pose des marques de fin potentielle de proposition qui lèvent alors le voile sur les extensions des éventuels domaines encore non clos. C'est alors le processus de sélection des extensions concurrentes qui prend le relais en effectuant les meilleurs choix de rattachement. Dans notre exemple, nous posons une marque de fin potentielle sur «*développe*», ce qui a pour effet d'autoriser les syntagmes futurs à être rattaché à *La base* puisqu'il s'agit d'une extension située dans un domaine non clos. Lorsque «*accélérera*» est traité, le système de sélection des extensions ne peut choisir que «*La base*». Le rattachement déclenche la confirmation de la fermeture du domaine propositionnel défini par «*que la SSII développe*».

En conclusion, pour rester cohérent, le graphe doit être construit avec une prise en compte simultanée du niveau syntagmatique et du niveau propositionnel. C'est actuellement l'alliance des contraintes relationnelles et des délimiteurs de segments propositionnels qui garantit la cohérence globale du graphe de relations.

3.3.7 Catégorisation des syntagmes en contexte

En section 2.5 page 101, nous avons vu que les contextes internes des syntagmes pouvaient ne pas être suffisants pour permettre leur catégorisation. Lorsque le cas se présente, une catégorie montrant l'ambiguïté leur est associée. Le processus de mise en relation doit alors les catégoriser en contexte.

« *[Le fichier] [d'incidents] [complète] [le dispositif].* » (6)

Dans l'exemple 6, l'étude du contexte syntagmatique ne peut lever l'ambiguïté verbo-adjectivale de *complète*. Lors de son intégration dans le graphe, étant donné le point d'extension de la relation sujet-verbe défini sur «*Le fichier*» et l'absence de point d'extension d'une relation régissant accordé en genre et en nombre, une relation sujet-verbe est instanciée et le syntagme est catégorisé verbal. Notons que la catégorisation de syntagmes en contexte entraîne la catégorisation des mots internes. En l'occurrence, le mot «*améliore*» est catégorisé verbe. La catégorisation de «*de*», préposition ou partitif, et de *que*, conjonction ou pronom objet, ainsi que la propagation d'attributs morphologiques entre syntagmes reliés sont fondées sur le même principe.

3.4 Regard sur la méthode

Après une étude sur corpus, nous avons remarqué que certaines relations ne s'instanciaient que dans le domaine propositionnel (e.g., la relation sujet-verbe). Sachant que l'unité d'entrée est une phrase, représentée sous la forme d'une suite de syntagmes, nous partitionnons cette suite de syntagmes de manière à faire apparaître les domaines propositionnels au sein desquels les relations intra-propositionnelles doivent être instanciées.

La proposition ne définit cependant pas toujours un domaine continu. Il est en effet commun de rencontrer des insertions de propositions à l'intérieur d'autres propositions. Ce phénomène est appelé «composition interne» par Hervé DÉJEAN (1998a). La partition est donc un peu particulière car elle ne met pas en évidence une suite de propositions mais une suite de segments propositionnels dont certains sont des propositions et d'autres des fragments de propositions.

L'obtention de la partition passe par l'identification de marques de début et de fin de segments propositionnels. Une fois cette partition établie et les propositions identifiées, nous utilisons la position à l'intérieur du domaine propositionnel pour instancier les relations conformément au domaine dans lequel elles doivent être instanciées : la relation sujet-verbe est instanciée dans le domaine propositionnel, la relation d'antécédence est instanciée entre deux domaines propositionnels.

Cette méthode est schématisée figure 3.1 page suivante. Notons que les trois opérations (b), (c) et (d) sont exécutées simultanément, en une seule passe, de la gauche vers la droite, le texte étant considéré comme un flux.

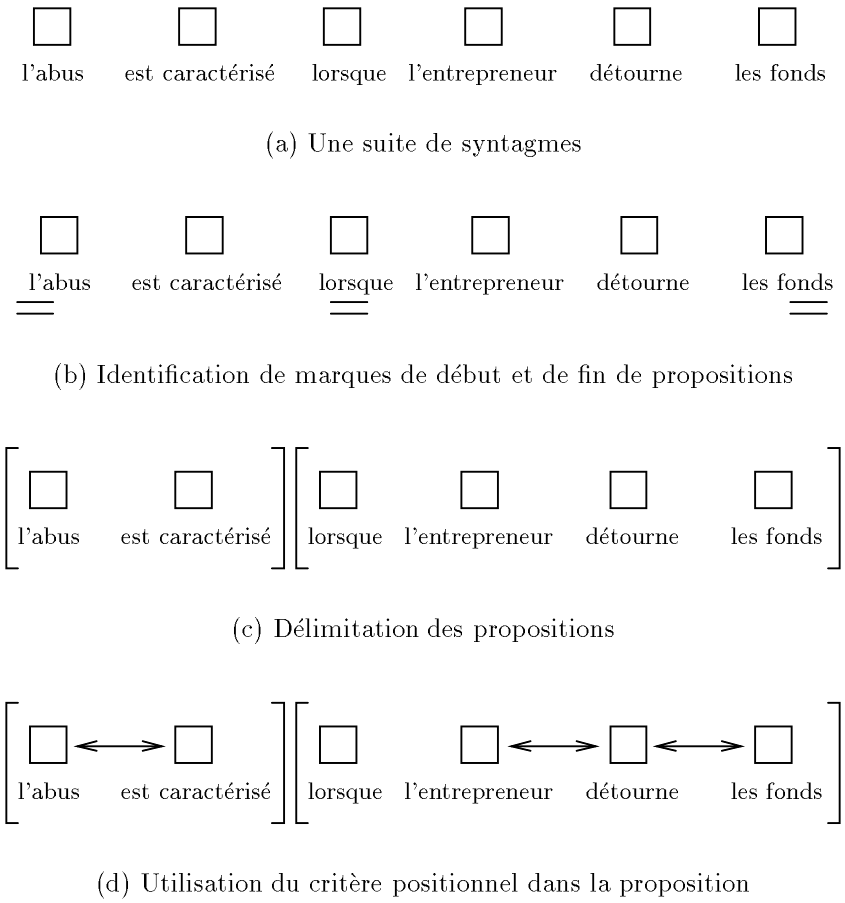


FIG. 3.1 - *Partition en proposition pour une application pertinente des connaissances relationnelles*

3.5 Cadre conceptuel pour une nouvelle implémentation

Le processus de mise en relation a émergé des recherches de Jacques VERGNE sur la syntaxe (1993, 1994, 1995). Il s'est enrichi et affiné par notre coopération sur la sélection des extensions concurrentes et sur l'introduction de la prise en compte des domaines propositionnels (GIGUET et VERGNE, 1997b) et nous avons pu montrer la pertinence du modèle en réalisant une évaluation sur un corpus large (c.f., annexe D.1 page 182).

L'implémentation actuelle (VERGNE, 1999) est née de l'observation de corpus et de la modification progressive de plusieurs prototypes dédiés à son analyse. Cette implémentation n'a jusqu'à présent fait l'objet d'aucune formalisation et nous en ressentons aujourd'hui les limites. Plutôt que de nous lancer dans sa description qui nous paraît un peu obsolète, nous préférons bénéficier de l'élan que constitue la description abstraite précédemment exposée pour définir un cadre formel adéquat à une nouvelle implémentation.

3.5.1 Modélisation du problème par un graphe

Notre problème est modélisable par un graphe orienté, ordonné et étiqueté. Ce graphe est défini comme suit :

$$G = (S, Succ, A, E, f, F, s_0) \text{ où :}$$

- S est l'ensemble des sommets $S = \{s_1, ..., s_n\}$ et $Succ$ est la relation successeur associée à l'ordre $s_1 < s_2 < ... < s_n$;
- s_0 est un sommet additionnel spécial, $s_0 \notin S$;
- E est l'ensemble fini des étiquettes des sommets et f est une fonction $f : S \to E$, qui, à chaque sommet sauf s_0, associe une étiquette ;
- A est un ensemble d'arcs étiquetés $(s_i, s_j, r), i, j \in [0, 1...n], r \in F$ où F est l'ensemble fini des étiquettes des arcs. Remarque : on notera le rôle des arcs adjacents au sommet spécial s_0 : ces arcs modélisent des relations potentielles.

3.5.2 Sémantique du graphe

Le graphe G est le graphe des relations syntaxiques d'un énoncé :

- les sommets $s_1, ..., s_n$ représentent les n syntagmes de l'énoncé dans leur ordre d'apparition ; *Succ* est la relation successeur correspondante ;
- chaque syntagme $s_i \in S$ est étiqueté grâce à la fonction $f : S \rightarrow E$;
- tout arc étiqueté (s_i, s_j, r) interne à G (c'est-à-dire avec $s_i, s_j \in S$) modélise une relation syntaxique dont la nature (e.g., sujet-verbe) est donnée par l'étiquette r ;
- tout arc adjacent à s_0 (s_i, s_0, r) ou (s_0, s_i, r) marque une relation potentielle pour le sommet s_i (i.e., c'est une «attente») : cette relation sera ou non «réalisée» au fur et à mesure de la construction du graphe G.

3.5.3 Principes de construction du graphe

La construction du graphe nécessite la définition d'une pile P de sommets à laquelle sont associés les opérateurs standards de manipulation d'une pile, à savoir *empiler*, *dépiler*, *sommet* et *pilevide*. Cette pile sert à la gestion des domaines propositionnels.

La construction du graphe de relations syntaxiques consiste en la construction de A. L'ensemble A est construit par un parcours séquentiel de tous les sommets $s_1...s_k...s_n$ du graphe. Cette construction passe par la création et la suppression de relations potentielles représentées par les arcs adjacents à s_0.

Initialement, A est vide. Un pas de l'algorithme consiste en le traitement d'un sommet s_k selon la nature du syntagme qu'il représente, selon le graphe partiel G déjà construit sur les sommets $s_0, s_1...s_k$, et selon l'état de la pile P.

Le pas de l'algorithme contient la connaissance linguistique adéquate à l'intégration cohérente du syntagme représenté par le sommet s_k. Ce pas est défini partiellement et évolue en fonction de nos connaissances sur la langue, de nos études sur corpus. Les concepts invoqués dans cette intégration ont été présentés dans la section précédente. Nous allons maintenant montrer qu'ils peuvent tous être transposés en termes d'opérations sur le graphe.

3.5.4 Transposition des concepts dans le modèle

L'extension contrôlée du graphe peut être assurée par les opérations suivantes :

- mémoriser un syntagme s_i comme point d'extension de la relation r se traduit par l'ajout d'un arc (s_i, s_0, r) ou (s_0, s_i, r).

- instancier une relation r entre les syntagmes s_i et s_j se traduit par la création d'un arc (s_i, s_j, r) ou (s_j, s_i, r) avec $s_i, s_j \in S$ et à la suppression des arcs (s, s_0, r') et (s_0, s, r') quel que soit s situé entre s_i et s_j et quel que soit r'.

- supprimer un point d'extension de la relation r sur le syntagme s_i se traduit par la suppression d'un arc (s_i, s_0, r) ou (s_0, s_i, r).

La gestion des domaines propositionnels s'effectue de la sorte :

- ouvrir un domaine propositionnel à partir du syntagme s_i se traduit par l'empilement de $s_i \in S$ sur P.

- fermer un domaine propositionnel se traduit par le dépilement d'un sommet sur P.

Définition des ensembles d'extensions concurrentes :

- l'ensemble des extensions concurrentes d'une relation d'étiquette r donnée correspond, selon l'orientation souhaitée, à l'ensemble $S'(r) = \{s_i \mid s_i \in S, (s_i, s_0, r) \in A \text{ ou } (s_0, s_i, r) \in A\}$;

- l'ensemble des extensions concurrentes, internes au domaine propositionnel courant, d'une relation d'étiquette r donnée, correspond, selon l'orientation donnée, au même ensemble avec $S''(r) = \{s_i \mid s_i \in S, ((s_0, s_i, r) \in A \text{ ou } (s_0, s_i, r) \in A\}) \text{ et sommet(P)} \leq s_i\}$.

3.5.5 Le processus par l'exemple

Pour concrétiser le déroulement du processus, nous allons reprendre pas à pas son exécution sur l'exemple 4 page 123. La segmentation en syntagmes minimaux produit le découpage suivant :

«*[Le système]$_N$[de production]$_{GP}$[améliore]$_V$[les rendements]$_N$*»

La figure 3.2 page ci-contre permet de suivre graphiquement la construction du graphe. En voici les explications associées :

pas n° 1 : le sommet s_1 est empilé sur P pour ouvrir un nouveau domaine propositionnel. Aucune extension du graphe n'est accessible donc le processus ne peut relier s_1. s_1 est nominal et non relié donc le processus crée des extensions pour un nominal non relié dans A (e.g., sujet potentiel et régissant potentiel).

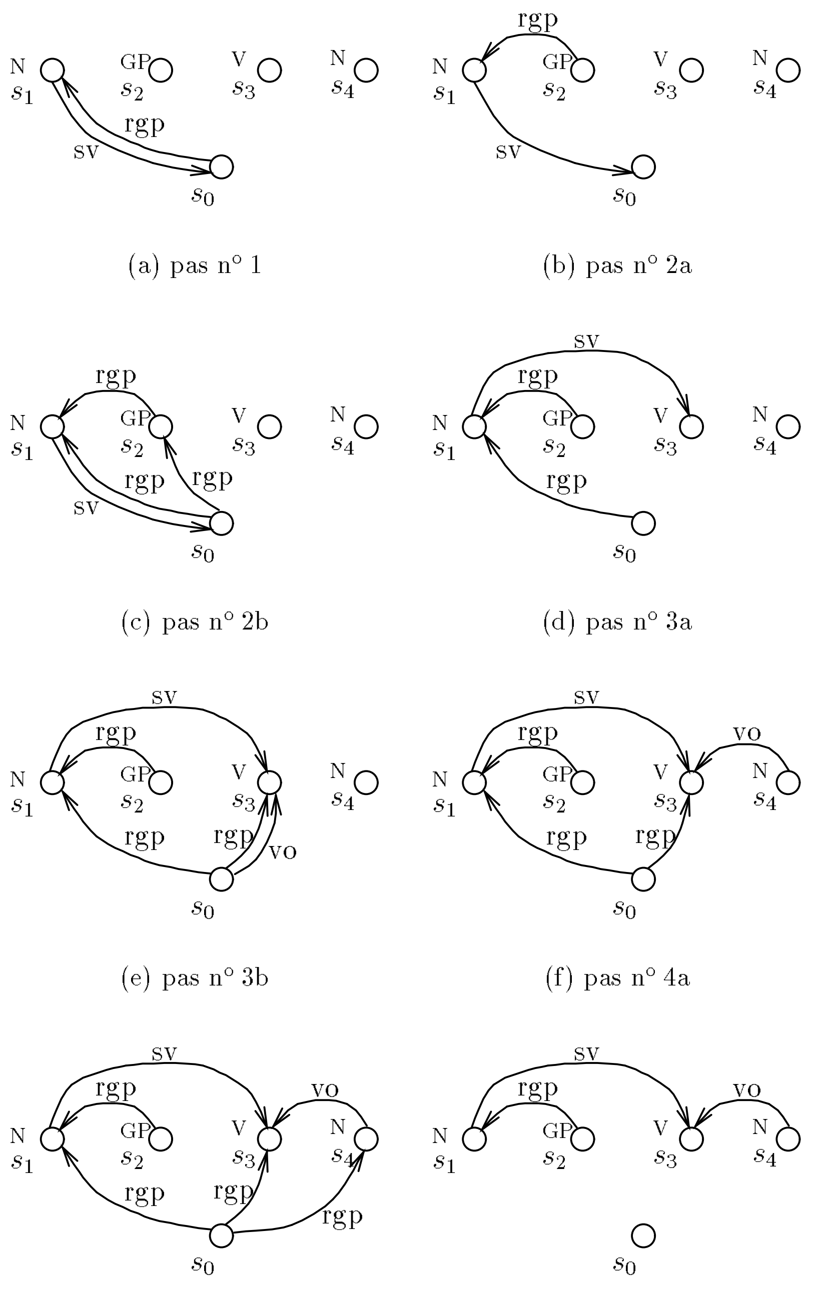
pas n° 2 : (a) le sommet s_2 est prépositionnel, le processus recherche les extensions de régissant dans le domaine propositionnel courant $S''(rgp) = \{s_i \mid s_i \in S, (s_0, s_i, rgp) \in A\}$. $S' = \{s_1\}$, le processus ajoute l'arc (s_2, s_1, rgp) à A, l'extension satisfaite (s_0, s_1, rgp) disparaît. (b) le processus crée de nouvelles extensions sur s_2 et sur s_1 (i.e., il peuvent tous les deux régir).

pas n° 3 : (a) s_3 est verbal, le processus recherche un sujet dans le domaine propositionnel courant $S''(sv) = \{s_1\}$, il instancie la relation, l'extension sujet sur s_1 est satisfaite, toutes les extensions situées entre s_1 et s_3 sont supprimées. (b) le processus crée de nouvelles extensions sur s_3 (e.g., il est transitif donc il peut attendre un objet, il peut régir).

pas n° 4 : (a) s_4 est nominal et n'est pas relié, le processus recherche une extension objet. $S' = \{s_3\}$, la relation est instanciée. (b) le processus définit de nouvelles extensions sur s_4 (e.g., il peut être régissant potentiel mais pas sujet potentiel car il est objet).

fin : tous les arcs adjacents à s_0 sont supprimés.

«[Le système]$_N$[de production]$_{GP}$[améliore]$_V$[les rendements]$_N$»



(a) pas n° 1
(b) pas n° 2a
(c) pas n° 2b
(d) pas n° 3a
(e) pas n° 3b
(f) pas n° 4a
(g) pas n° 4b
(h) fin

FIG. 3.2 - *Exemple d'exécution pas à pas du processus*

3.6 Conclusion

3.6.1 De la pertinence du processus

Dans ce chapitre, nous avons présenté un processus original pour le calcul des relations syntaxiques entre syntagmes minimaux. Ce processus est né du constat que l'étude de la forme des éléments à relier n'est pas suffisante à la définition d'une relation syntaxique : une relation n'a d'existence qu'au travers de l'existence des autres relations. Pour rendre compte de cette caractéristique, nous avons procédé à une reformulation faisant apparaître la dynamique essentielle à la définition d'un processus : toute instanciation de relation est contrainte par les relations précédemment créées et une fois créée, toute relation devient contrainte pour la création des relations à venir.

C'est dire si la chronologie tient une place prépondérante dans notre approche. La prise en compte de la dimension chronologique se traduit par l'assimilation de la phrase à un flux : les syntagmes de la phrase sont intégrés séquentiellement à la structure relationnelle et lors du traitement d'un syntagme, la structure s'enrichit de toutes les relations qu'il est possible de déduire de l'occurrence de cet élément à cette position.

Tout au long de la construction du graphe relationnel évolue l'ensemble de ses extensions cohérentes. Cet ensemble représente les relations que le processus est autorisé à créer sur le graphe sans que sa cohérence globale ne s'en trouve affectée. Toute instanciation de relation dans le graphe s'accompagne d'une mise à jour de l'ensemble de ses extensions cohérentes : certaines extensions précédemment valides ne le sont plus selon la nouvelle configuration, d'autres, au contraire, deviennent légitimes au regard de la nouvelle structure.

Cette propagation de contraintes relationnelles permet une maîtrise de la combinatoire et garantit la cohérence globale des analyses proposées.

Le processus se trouve empiriquement validé par l'évaluation du calcul de la relation sujet-verbe sur un corpus du journal «Le Monde», corpus composé de phrases complexes (c.f., annexe D.1 page 182). Sur ce corpus comptant 738 relations sujet-verbe, le système actuel, bien qu'expérimental, a détecté 94,04% des relations et établi correctement 96,39% d'entres elles.

3.6.2 De la nécessité du domaine propositionnel

Cette étude sur la mise en relation des syntagmes nous a permis de constater que, contrairement à nos intuitions, le calcul des relations syntaxiques entre syntagmes ne pouvait être effectué correctement en considérant la phrase comme unité hiérarchique immédiatement supérieure au syntagme. Nous remarquons en effet que certaines relations ne peuvent être instanciées que dans le domaine propositionnel, alors que d'autres peuvent être instanciées entre deux domaines propositionnels différents. Nous ne pouvons donc négliger la proposition comme structure intermédiaire.

Pour prendre en compte le domaine propositionnel, nous avons été obligé d'introduire un mécanisme de délimitation des propositions. Les marques de début et de fin de proposition que nous calculons ne partitionnent cependant pas la phrase en propositions mais en segments propositionnels. Le domaine propositionnel n'est en effet pas toujours continu car une proposition peut être insérée dans une autre proposition. Le segment propositionnel délimité est un sous-domaine propositionnel continu.

Une fois les segments propositionnels localisés et le concept de proposition intégré dans le processus, nous remarquons que les relations sont toujours créées de manière pertinente. Ce constat suffit à considérer la proposition comme unité hiérarchiquement supérieure au syntagme. Il est intéressant de noter que la proposition, concept majeur en linguistique, émerge ici syntaxiquement par nécessité et non par projection d'un attendu *a priori* logique.

3.6.3 De la formalisation du processus

Lors de la formalisation de notre problème, nous avons proposé de modéliser la structure syntaxique par un graphe orienté, étiqueté et ordonné. Le problème de l'analyse syntaxique se traduit alors par la construction de ce graphe. Nous avons défini les opérations autorisées sur le graphe en transposant les opérations linguistiques que nous avions préalablement identifiées lors de nos études sur corpus.

Le cadre formel que nous proposons est un *cadre ouvert*. Il dévoile par cette caractéristique sa capacité à évoluer en fonction des connaissances que nous devrons y incorporer pour modéliser les phénomènes syntaxiques souhaités. Le principe de construction du graphe ainsi que les opérations autorisées sur le graphe constituent *le noyau* du modèle, c'est-à-dire ce que nous tenterons de préserver lors de la prise en compte de nouveaux phénomènes. Par la définition d'un noyau et la capacité d'extension de ce noyau, on peut voir cette formalisation comme la définition d'un méta-modèle.

Bien entendu, au travers de cette formalisation, nous réinvestissons tout un pan de la culture en Intelligence Artificielle. Nous intégrons en effet des concepts très forts tels que les systèmes ouverts, la propagation de contraintes. Mais nous effectuons également une remontée vers la théorie des graphes et la théorie des automates. Le graphe que nous manipulons est en effet une structure algorithmique qui a fait l'objet d'études nombreuses et le principe de construction du graphe par l'intermédiaire d'un sommet additionnel s'inscrit tout à fait dans la théorie des automates.

Chapitre 4

Regard sur la méthode

Dans ce chapitre faisant office de conclusion de partie, nous souhaitons mettre l'accent sur la méthode ayant permis le calcul de la structure syntaxique.

4.1 Présentation de la méthode

Pour construire les syntagmes minimaux et pour mettre en relation ces syntagmes, une méthode unique a été utilisée : c'est cette méthode qui est à l'origine de la qualité des analyses obtenues.

La méthode définit comment construire une suite d'unités segmentales structurées à partir d'une suite d'unités dites «de base» (i.e., une suite de syntagmes à partir d'une suite de mots, une suite de segments propositionnels à partir d'une suite de syntagmes). La méthode consiste à repérer les unités segmentales dans la suite d'unités de base, à calculer une première approximation de leur structure à partir de la forme et de la position des unités de base internes, puis à affiner cette structure en étudiant la position et le rôle de l'unité segmentale dans la structure qui l'intègre.

La méthode réalise donc une structuration de chaque segment en deux temps : dans un premier temps à l'aide d'indices formels et positionnels internes à un segment, dans un second temps à l'aide d'indices formels et positionnels externes au segment.

4.2 Utilisation d'indices internes

Le principe de la première étape consiste à ne calculer des relations syntaxiques entre les unités de base que dans le cadre de l'unité segmentale souhaitée résultante. Pour la catégorisation des mots d'un énoncé, il faut se placer dans le cadre du syntagme minimal. Pour la mise en relation des syntagmes, il faut se placer dans le cadre du segment propositionnel.

Pour calculer des relations syntaxiques entre unités de base, uniquement dans le cadre des unités segmentales, nous marquons les limites de ces dernières sur la suite d'unités de base. La technique consiste à identifier des marques formelles de début et de fin d'unités segmentales. Pour délimiter les mots aussi bien que pour délimiter les propositions, les marques formelles de début et de fin d'unités segmentales sont déduites des contiguïtés de catégories et d'attributs morphologiques des unités de base.

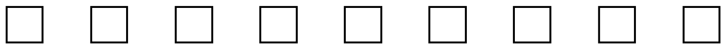
Une fois les unités segmentales délimitées, le calcul des relations entre les unités de base qui la composent n'utilise que des indices internes à ce segment, à savoir la forme des unités et leur position relative. Dans le cadre de la catégorisation des mots, un mot n'est catégorisé qu'à l'aide d'indices extraits du domaine syntagmatique. Dans le cadre du processus de mise en relations, les relations intra-propositionnelles (e.g., la relation sujet-verbe) ne sont instanciées qu'à l'aide d'indices extraits du domaine propositionnel.

Cette première étape, qui consiste à calculer des relations syntaxiques entre des unités de base situées dans une même unité segmentale tout en n'utilisant que des indices internes à cette unité, peut être représentée par le schéma 4.1 page ci-contre.

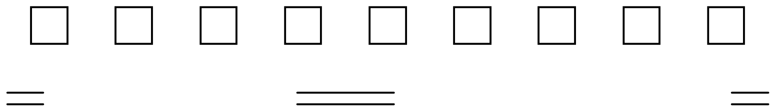
4.3 Utilisation d'indices externes

La première étape n'est cependant pas suffisante. Lorsque le domaine défini par l'unité segmentale n'est pas suffisamment informatif, certaines relations ne peuvent être établies de manière fiable. Dans le cadre de la catégorisation des mots, il n'est par exemple pas possible de catégoriser un mot s'il est à la fois polycatégoriel et seul dans son syntagme.

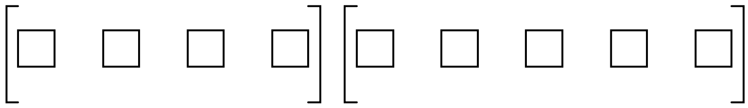
La seconde étape de la méthode consiste alors à étudier les relations entre



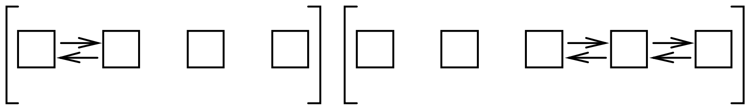
(a) Une suite d'unités de base en entrée



(b) Identification des marques de début et de fin des unités segmentales



(c) Délimitation des unités segmentales

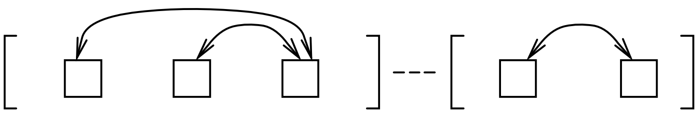


(d) Calcul de relations entre unités de base avec contexte d'étude restreint au domaine défini par l'unité segmentale

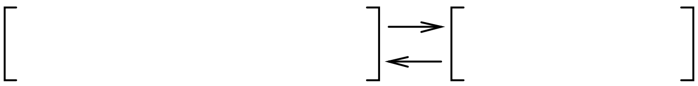
FIG. 4.1 - *Calcul de relations entre unités de base à l'aide d'indices internes*

l'unité segmentale partiellement construite et les autres unités segmentales construites. Il s'agit là d'un des points méthodologiques fondamentaux de la méthode. Il faut accepter de manipuler une unité segmentale dont la structure interne n'est pas totalement stabilisée : c'est en étudiant le comportement des unités segmentales dans leur structure que l'on peut enrichir la structure de chacune de ces unités.

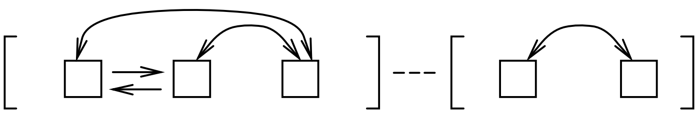
À titre d'illustration, nous avons vu que le processus de mise en relation pouvait manipuler des syntagmes qui n'avaient été que partiellement catégorisés et structurés par le processus de construction des syntagmes minimaux et que le processus de mise en relation pouvait catégoriser en contexte ces syntagmes, ce qui permettait par la suite d'affiner la structure interne des syntagmes partiellement construits. Cette seconde étape est schématisée figure 4.2.



(a) Une suite d'unités de base après délimitation des unités segmentales et calcul des relations internes



(b) Changement d'échelle unité de base/unité segmentale et calcul des relations entre unités segmentales



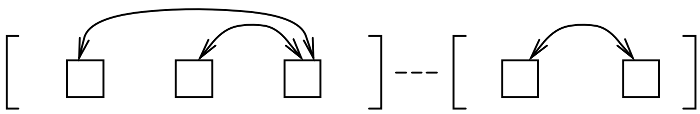
(c) Exploitation des relations entre unités segmentales pour compléter les relations entre unités de base

FIG. 4.2 - *Calcul de relations entre unités de base à l'aide d'indices externes*

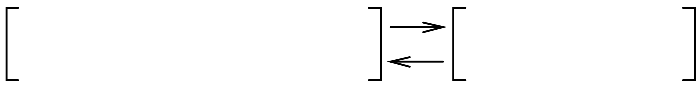
4.4 Mise en œuvre de la méthode

La méthode dans son ensemble n'a pour le moment été implémentée que dans le cadre de la catégorisation des mots : (1) les syntagmes sont délimités à l'aide d'une catégorisation minimale des mots puis, une fois les syntagmes délimités, la catégorisation des mots est effectuée en respectant le domaine syntagmatique, ce que préconise la première étape de la méthode, (2) puis, selon la seconde étape de la méthode, les syntagmes sont mis en relation, et, de ces relations est extraite la connaissance nécessaire à la finalisation de la catégorisation des mots. Le suivi de la méthode dans son ensemble a permis d'obtenir le système de catégorisation des mots actuel le plus performant sur corpus français.

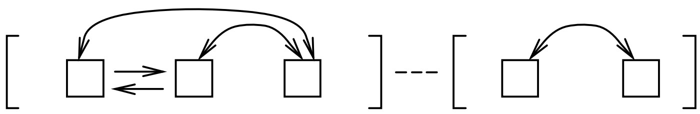
L'utilisation des indices externes pour la catégorisation des mots se schématise comme suit :



(a) Une suite de mots après délimitation des syntagmes et catégorisation des mots internes



(b) Changement d'échelle mot/syntagme et calcul des relations entre syntagmes

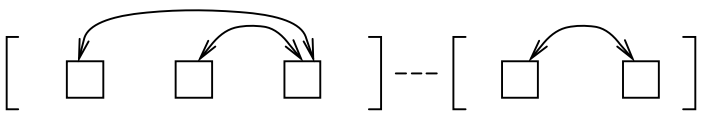


(c) Exploitation des relations entre syntagmes pour affiner la catégorisation des mots

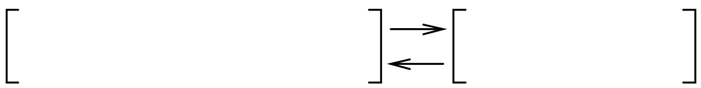
FIG. 4.3 - *Calcul de la catégorie des mots à l'aide d'indices externes*

Le processus de mise en relation a entre autres pour ambition de construire la structure intra-propositionnelle, il effectue donc la délimitation des propositions et calcule les relations intra-propositionnelles. Comme nous l'avons vu, les résultats obtenus sont de très bonne qualité. Mais il faut noter que seule la première étape de la méthode est actuellement implémentée : l'utilisation d'indices internes. Il nous faut aujourd'hui travailler sur la seconde étape de la méthode, à savoir l'utilisation d'indices externes. Il nous faut donc accepter la qualité actuelle de la mise en relation des syntagmes et passer à la mise en relation des segments propositionnels qui va fournir la connaissance nécessaire pour affiner les relations entre syntagmes.

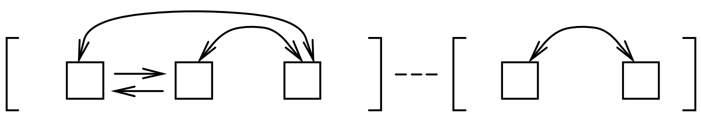
L'utilisation des indices externes pour la mise en relation des syntagmes se schématise comme suit :



(a) Une suite de syntagmes après délimitation des segments propositionnels et calcul des relations internes



(b) Changement d'échelle syntagme/proposition et calcul de la structure propositionnelle



(c) Exploitation de la structure propositionnelle pour affiner les relations entre syntagmes

FIG. 4.4 - *Calcul de relations entre syntagmes à l'aide d'indices externes*

4.5 Identité des processus

L'identité de la méthode ayant permis la catégorisation des mots et de la méthode ayant permis la mise en relation des syntagmes suggère que le processus de construction des syntagmes minimaux et le processus de mise en relation des syntagmes sont amenés à devenir à terme identiques.

Si l'on fait abstraction du type des unités manipulées, on constate en effet que les deux processus travaillent sur des entrées identiques, que les deux processus ont des objectifs identiques, que les deux processus utilisent des moyens identiques pour parvenir à ces objectifs. Développons ces trois aspects.

Les deux processus ont des entrées identiques : une suite d'unités de base. Il s'agit d'une suite de mots pour le premier processus et d'une suite de syntagmes pour le second.

Les deux processus ont des objectifs identiques : construire la structure relationnelle interne des unités segmentales. Le premier processus construit la structure relationnelle interne du syntagme minimal[1], le second processus construit la structure relationnelle interne des propositions.

Les deux processus utilisent des moyens identiques pour parvenir à ces objectifs : délimiter les unités segmentales et réaliser à l'intérieur de ces unités une propagation de contraintes relationnelles sur les unités de base qui la composent. Le premier processus délimite les syntagmes et réalise à l'intérieur de chaque syntagme une propagation de contraintes relationnelles sur les mots. Le second délimite les propositions et réalise à l'intérieur de chaque proposition une propagation de contraintes relationnelles sur les syntagmes.

Pour les deux processus, la propagation de contraintes relationnelles établit des relations entre les unités et peut les catégoriser en contexte. Pour le premier processus, la propagation de contraintes établit des relations entre mots et catégorise les mots ambigus en contexte. Pour le second processus, la propagation de contraintes établit des relations entre syntagmes et catégorise les syntagmes ambigus en contexte.

Pour les deux processus, les propagations de contraintes relationnelles

1. La catégorisation des mots dans un syntagme par un jeu de catégories distributionnel correspond au calcul de la structure relationnelle du syntagme. La relation n'est pas explicite mais codée implicitement dans les catégories distributionnelles.

s'effectuent de la gauche vers la droite et partagent des objectifs structurels identiques : la création de structures planaires pour les relations de portée restreinte au domaine segmental. Pour le premier processus, les relations intra-syntagmatiques sont planaires dans le syntagme et pour le second, les relations intra-propositionnelles sont planaires dans la proposition.

De ces remarques, nous concluons qu'il y a stricte identité du processus de construction des syntagmes minimaux et du processus de mise en relation de ces syntagmes, au détail près du type des unités. Le cadre formel que nous avons défini pour une nouvelle implémentation du processus de mise en relation peut donc être utilisé pour le processus de construction des syntagmes minimaux, il suffit de le paramétrer par l'unité hiérarchique manipulée, le mot dans le premier processus et le syntagme dans le second, et par les règles qui définissent les relations entre ces unités.

4.6 Le processus d'analyse syntaxique

Le processus d'analyse que nous avons proposé dans le premier chapitre de cette partie (c.f., section 1.5 page 88), met en œuvre, dans sa dynamique, la méthode que nous venons de décrire. Les deux étapes de la méthode sont réalisées par les sous-processus de construction des syntagmes et de mise en relation qui s'exécutent simultanément et interactivement sur le flux textuel. Voyons comment s'organise la coopération entre ces deux processus.

La construction de la structure interne des syntagmes commence par une propagation de déductions contextuelles sur les mots. Effectuée par le premier processus, cette propagation doit, conformément à la méthode, respecter les domaines syntagmatiques. Les limites de syntagmes ne sont pas précalculées puisque le texte est traité comme un flux, mais recherchées sur le front de progression des déductions contextuelles. Lors de la détection d'une fin de syntagme, ce dernier est catégorisé et transmis au processus de mise en relation. Cette première phase correspond à la première étape de la méthode : la construction de la structure des syntagmes à l'aide d'indices internes.

Le processus de mise en relation prend alors en charge le syntagme et l'utilise pour enrichir la structure relationnelle qu'il gère. Les relations sont créées par un mécanisme de propagation de contraintes relationnelles qui, conformément à la méthode, respecte les domaines propositionnels lors de l'instanciation de relations intra-propositionnelles. Les limites de segments propositionnels ne sont pas précalculées mais recherchées dynamiquement à chaque prise en compte d'un nouveau syntagme.

Les syntagmes auxquels sont associés des catégories ambiguës sont catégorisés dynamiquement lors de l'instanciation des relations ; l'instanciation d'une relation est également le moment de la propagation d'attributs morphologiques entre syntagmes reliés. Ce complément d'information sur les syntagmes va permettre au premier processus d'affiner ses structures relationnelles ambiguës. Cette seconde phase correspond à la seconde étape de la méthode : la construction de la structure des syntagmes à l'aide d'indices externes.

Le processus d'analyse est adéquat à la méthode puisqu'il en est le fidèle reflet. Il a tout naturellement fait ses preuves, comme le prouvent les évaluations qui ont été effectuées sur des corpus de phrases complexes. Il est cependant clair qu'il doit être étendu pour parfaire et étendre les phénomènes syntaxiques calculés, l'extension à laquelle on pense immédiatement étant la construction de la structure propositionnelle.

Le chemin parcouru

Chapitre 1

Méthode pour l'analyse automatique de structures formelles

1.1 La méthode

La méthode a été le leitmotiv de cette thèse. Quelle que soit la structure formelle à calculer, une seule méthode a toujours été employée. Cette méthode, dont on sait la puissance au regard de la qualité des résultats obtenus à la fois pour l'identification de la langue et pour l'analyse syntaxique automatique d'un énoncé, définit comment construire une suite d'unités segmentales structurées à partir d'une suite d'unités dites «de base».

La méthode consiste à repérer les unités segmentales, ou segments, dans la suite d'unités de base, à calculer une première approximation de leur structure en utilisant des critères formels et positionnels des unités uniquement internes, puis à affiner cette structure en étudiant la position et le rôle de l'unité segmentale dans la structure qui l'intègre.

Notre méthode réalise donc le calcul d'une structure formelle en utilisant deux types de ressources : (1) des ressources internes, à savoir les caractéristiques formelles et positionnelles des unités composant la structure à créer, (2) des ressources externes, à savoir la position et le rôle de la structure formelle dans l'unité linguistique à laquelle elle est intégrée.

On retrouve dans notre approche une distinction méthodologique posée par exemple par Bloomfield entre relations endocentriques (à l'intérieur de l'unité) et relations exocentriques (à l'extérieur des unités, soit dans le domaine hiérarchique immédiatement supérieur). La même idée est formulée par Martinet en termes d'expansions à l'intérieur d'un groupe, et de relations fonctionnelles à l'intérieur de la proposition.

1.2 Méthode et structures formelles

La méthode que nous proposons permet de construire et structurer des suites d'unités linguistiques de type varié sur critères formels, par exemple des suites de caractères, des suites de mots, des suites de syntagmes. Elle est applicable dès lors que le traitement automatique d'un énoncé est basé sur une étude formelle, l'identification de la langue et l'analyse syntaxique d'un énoncé conservant cependant chacun leur particularité.

Pour identifier la langue d'un énoncé monolingue représenté sous la forme d'une suite de caractères, la méthode construit une suite de mots (i.e. tokens) à partir de la suite de caractères puis vérifie positionnellement des propriétés morphologiques internes à chaque mot. Seuls des indices internes au mot ont été utilisés pour réaliser le diagnostic mais nous avons mentionné que la connaissance du contexte dans lequel apparaît le mot pouvait permettre d'améliorer encore les résultats. La prise en compte de ce contexte correspond à l'utilisation d'indices externes.

Dans l'analyse syntaxique automatique, si l'on observe la répartition des tâches entre l'analyse morpho-lexicale et la catégorisation des mots en contexte, il est agréable de constater que c'est exactement cette méthode qui a été utilisée : l'étude morpho-lexicale permet par indices internes de proposer une première catégorisation des mots, puis l'étude du mot dans son contexte (le syntagme minimal) vient affiner cette catégorie. Il s'agit là de l'utilisation d'indices externes.

La méthode a permis également de construire une suite de syntagmes à partir d'une suite de mots : le calcul de la structure relationnelle entre mots internes au syntagme est réalisé en utilisant la position et la catégorie des mots à l'intérieur du syntagme puis en affinant cette structure relationnelle en étudiant le rôle et la position du syntagme dans la structure qui l'intègre. Il s'agit également d'une utilisation d'indices internes et d'indices externes.

Enfin, la méthode a été en partie appliquée pour calculer une suite de segments propositionnels à partir d'une suite de syntagmes : le calcul de la structure relationnelle entre syntagmes s'effectue en tenant compte de leur position dans la proposition. Il s'agit là de l'utilisation d'indices internes permettant d'organiser la structure propositionnelle.

1.3 Méthode et analyse automatique

La présentation de la méthode est née de l'étude *a posteriori* à la fois des réalisations informatiques que nous avons effectuées, à savoir l'identification de la langue, et des développements dans lesquels nous nous sommes intégré, à savoir les travaux de Jacques Vergne en analyse syntaxique automatique. En cela, la méthode, bien que présentée de manière générale et abstraite, est attachée à une réalité informatique et se trouve naturellement compatible avec un traitement automatisé.

Généralisée et abstraite, la méthode est une contribution importante pour la conception de processus d'analyse automatique de structures formelles. Elle offre en effet un processus abstrait d'analyse automatique de structures formelles au sein duquel le type de l'unité linguistique manipulée et les règles de construction de l'unité segmentale sont des paramétres.

Une seconde contribution notable pour l'analyse automatique de structures formelles est la démonstration de la possibilité et de l'intérêt de manipuler des unités linguistiques dont la structure est incomplète. Nous avons en effet vu qu'en étudiant en contexte les unités ayant une structure incomplète, il était possible d'acquérir des connaissances sur leur fonction et d'utiliser cette information pour affiner leur structure interne.

Cette stratégie d'analyse s'inscrit en tant qu'alternative aux stratégies classiques qui, refusant la manipulation de structures incomplètes, annulent toutes les ambiguïtés structurelles dès leur détection au moyen de choix plus ou moins aléatoires et s'engagent alors irrémédiablement dans un coûteux processus de révision de ces choix (i.e., *backtrack*) lorsque ceux-ci entraînent un échec de l'analyse en cours.

Dans notre démarche, lorsque nous manipulons des structures incomplètes, nous retardons en fait certaines décisions jusqu'au moment où le contexte, étendu de manière cohérente, devient assez informatif pour prendre directement la bonne décision. C'est donc toute une source de combinatoire purement artificielle qui est évitée.

1.4 Une passerelle entre linguistique et informatique

La définition d'une méthode d'analyse automatique de structures formelles ainsi que la formalisation d'un processus d'analyse automatique de telles structures a pour ultime objectif de définir un espace au sein duquel les communautés linguistique et informatique puissent discuter. Nous avons en effet veillé à ce que les exigences de chacune des communautés soient respectées.

Côté linguistique, les unités que nous manipulons ont une existence réelle, confortée par leur validité multilingue ; la définition de ces unités par critères formels et positionnels internes et externes est acceptable car elle trouve son équivalence dans des théories linguistiques reconnues.

Côté informatique, les méthodes de résolution relevant de l'intelligence artificielle sont appliquées (e.g., propagation de contraintes, diviser pour régner, séparation du moteur et des ressources linguistiques) ; les principes de conception de processus satisfont les exigences du génie logiciel (e.g., généricité, réutilisabilité).

Nous considérons donc que le but poursuivi est atteint et, au regard de la qualité des résultats qu'une telle alliance nous a permis d'obtenir, nous pensons qu'il serait souhaitable de continuer dans cette voie.

Annexes

Annexe A

L'identificateur de langue

A.1 Évaluation détaillée

Avant de présenter les évaluations individuelles de chacune des quatres langues traitées, nous fournissons à titre indicatif les performances de l'identificateur en terme de vitesse. Sachant que le module d'identification a été compilé de manière optimisée (-O3) avec le compilateur g++ version egcs-2.91.57 et exécuté sur une station UltraSparc monoprocesseur à 167MHz doté d'une mémoire de 128Mo, système d'exploitation Solaris 2.6, le temps d'analyse moyen est de 250 000 caractères à la seconde. Ce temps pourrait être amélioré par l'utilisation d'automates à états finis.

A.1.1 Corpus français

Capacité à converger vers une seule langue

nb de mots	nb de segments monolingues	Expérience n° 1 : suffixes désactivés %L	%H	%I	Expérience n° 2 : suffixes activés %L	%H	%I
1	227	8,81	19,82	71,37	46,70	20,70	32,60
2	98	45,92	25,51	28,57	81,63	14,29	4,08
3	99	59,69	33,33	7,07	92,93	7,07	0
4	89	77,53	21,35	1,12	94,38	5,62	0
5	87	80,46	17,24	2,30	96,55	3,44	0
6	87	95,40	4,60	0	100	0	0
7	104	96,15	3,85	0	100	0	0
8	111	97,30	2,70	0	100	0	0
9	90	100	0	0	100	0	0
10	93	100	0	0	100	0	0
...	...	100	0	0	100	0	0

TAB. A.1 - *Capacité à converger vers une seule langue (corpus français)*

Capacité à identifier la bonne langue

Étant donné le peu d'erreurs, nous citons exhaustivement les segments ayant posé problèmes :

– segments étiquetés «anglais» à tort : «*Bofors ?*» et «*mouvement social*» ;

langue du segment	nb de segments catégorisés	nb de segments catégorisés correctement	incorrectement
français	2 802	?	?
anglais	2	0	2
espagnol	5	0	5
allemand	5	0	5

Lecture : parmi les 2 segments monolingues catégorisés anglais, aucun n'était effectivement anglais. La colonne «catégorisation correcte» met en évidence le multilinguisme du corpus (supposé) monolingue.

TAB. A.2 - *Détail de l'identification, corpus français, suffixes activés*

– segments étiquetés «espagnol» à tort : «*Morbihan*», «*, affirme Istvan Schein.*», «*, affirme le Coran.*», «*un timbre Nadar*» et «*Joseph Castaldini, 6, square de Colmar, 68 490 Chalampé*» ;

– segments étiquetés «allemand» à tort : «*post*» (3 occurrences), «*riposte*» et «*Etats-Unis*».

A.1.2 Corpus anglais

Capacité à converger vers une seule langue

nb de mots	nb de segments monolingues	Expérience n° 1 : suffixes désactivés %L	%H	%I	Expérience n° 2 : suffixes activés %L	%H	%I
1	226	10,18	3,98	85,84	16,82	14,60	68,58
2	95	30,53	3,16	66,31	51,58	20	28,42
3	63	74,60	9,53	15,87	88,89	7,94	3,17
4	59	83,05	11,87	5,08	93,22	6,78	0
5	67	92,54	5,97	1,49	100	0	0
6	69	100	0	0	100	0	0
7	79	100	0	0	100	0	0
8	91	98,90	1,10	0	100	0	0
9	112	99,11	0	0,89	100	0	0
10	110	100	0	0	100	0	0
...	...	100	0	0	100	0	0

TAB. A.3 - *Capacité à converger vers une seule langue (corpus anglais)*

Capacité à identifier la bonne langue

langue du segment	nb de segments catégorisés	nb de segments catégorisés correctement	incorrectement
français	24	20	4
anglais	2 728	?	?
espagnol	26	21	5
allemand	10	0	10

TAB. A.4 - *Détail de l'identification, corpus anglais, suffixes activés*

Étant donné le peu d'erreurs, nous citons exhaustivement les segments ayant posé problèmes :

- segments étiquetés «français» à tort : «*Sir W.*», «*or little housebuilder*», «*a Trigonocephalus, or Cophias*» et «*Observaciones Geologicas*» ;

- segments étiquetés «espagnol» à tort : «*E.*» (2 occurrences), «*Linnaean Trans,.* », «*Jour., Jan 1830*» et «*Arrow-head, antiquarian Relic*» ;
- segments étiquetés «allemand» à tort : «*ST.*» (2 occurrences), «*Montagne, in Comptes Rendus, etc., Juillet, 1844;*», «*Labillardiere, vol.*», «*Monte Video*», «*Molothrus niger*», «*August 11th.*», «*Chem.*», «*Martens shot an ostrich;*», «*[2] Travels in Africa p.*».

A.1.3 Corpus espagnol

Capacité à converger vers une seule langue

nb de mots	nb de segments monolingues	Expérience n° 1 : suffixes désactivés %L	%H	%I	Expérience n° 2 : suffixes activés %L	%H	%I
1	25	20	52	28	32	56	12
2	51	86,28	5,88	7,84	94,12	3,92	1,96
3	103	92,23	7,77	0	97,09	2,91	0
4	118	94,92	5,08	0	99,15	0,85	0
5	205	98,54	1,46	0	99,51	0,49	0
6	168	97,62	2,38	0	99,40	0,60	0
7	188	99,47	0,53	0	99,47	0,53	0
8	191	99,48	0,52	0	100	0	0
9	164	100	0	0	100	0	0
10	185	99,46	0,54	0	100	0	0
...	...	100	0	0	100	0	0

Tab. A.5 - *Capacité à converger vers une seule langue (corpus espagnol)*

Capacité à identifier la bonne langue

langue du segment	nb de segments catégorisés	nb de segments catégorisés correctement	incorrectement
français	9	0	9
anglais	2	0	2
espagnol	3 296	?	?
allemand	0	0	0

Tab. A.6 - *Détail de l'identification, corpus espagnol, suffixes activés*

Étant donné le peu d'erreurs, nous citons exhaustivement les segments ayant posé problèmes :

– segments étiquetés «français» à tort : «*¡Quién diera fuese la tarde!*», «*A tu simiente la daré.*», «*¡Moisés, Moisés!*», «*¿qué les responderé?*», «*Perseguiré, prenderé, repartiré despojos;*», «*A tu simiente la daré:*», «*Oye, Israel: (2 fois)*» et «*Extendiste tu diestra;*» ;

– segments étiquetés «anglais» à tort : «*decirle has:*» et «*Peregrino he sido en tierra ajena;*».

A.1.4 Corpus allemand

Capacité à converger vers une seule langue

nb de mots	nb de segments monolingues	Expérience n° 1 : suffixes désactivés %L	%H	%I	Expérience n° 2 : suffixes activés %L	%H	%I
1	53	3,77	39,62	56,61	52,83	35,85	11,32
2	37	37,84	13,51	48,65	91,89	5,41	2,70
3	29	62,07	13,79	24,14	79,31	17,24	3,45
4	50	90	8	2	96	4	0
5	68	98,53	0	1,47	100	0	0
6	81	98,77	1,23	0	100	0	0
7	67	100	0	0	100	0	0
8	83	100	0	0	100	0	0
9	80	100	0	0	100	0	0
10	100	100	0	0	100	0	0
...	...	100	0	0	100	0	0

TAB. A.7 - *Capacité à converger vers une seule langue (corpus allemand)*

Capacité à identifier la bonne langue

langue du segment	nb de segments catégorisés	nb de segments catégorisés correctement	incorrectement
français	25	0	25
anglais	1	1	0
espagnol	0	0	0
allemand	2 185	?	?

TAB. A.8 - *Détail de l'identification, corpus allemand, suffixes activés*

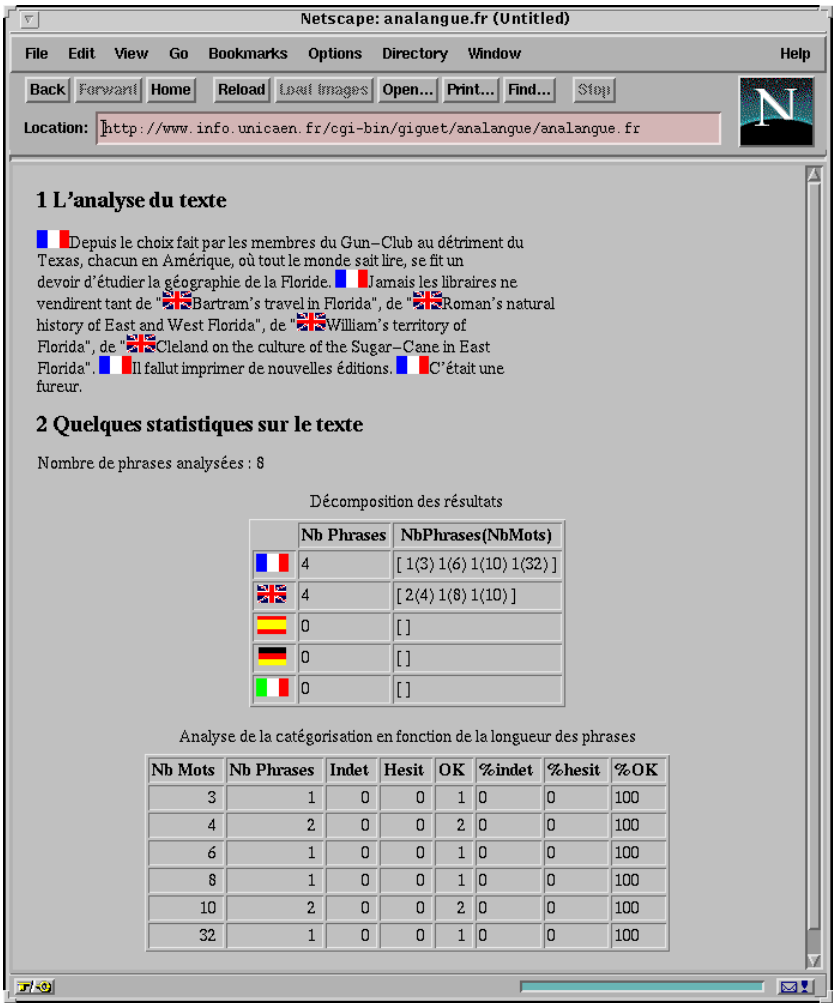
Étant donné le peu d'erreurs, nous citons exhaustivement les segments ayant posé problèmes :

- segments étiquetés «français» à tort : «*Artikel*» (24 occurrences) et «*Dezember 1995 Sache des Bundes.*» ;

A.2 Le diagnostiqueur de langue sur internet

http://www.info.unicaen.fr/~giguet/diagnostic-fr.html

FIG. A.1 - *L'interface du diagnostiqueur de langue*


Netscape: analangue.fr (Untitled)

File Edit View Go Bookmarks Options Directory Window Help

Back Forward Home Reload Load Images Open... Print... Find... Stop

Location: http://www.info.unicaen.fr/cgi-bin/giguet/analangue/analangue.fr

1 L'analyse du texte

Depuis le choix fait par les membres du Gun–Club au détriment du Texas, chacun en Amérique, où tout le monde sait lire, se fit un devoir d'étudier la géographie de la Floride. Jamais les libraires ne vendirent tant de "Bartram's travel in Florida", de "Roman's natural history of East and West Florida", de "William's territory of Florida", de "Cleland on the culture of the Sugar–Cane in East Florida". Il fallut imprimer de nouvelles éditions. C'était une fureur.

2 Quelques statistiques sur le texte

Nombre de phrases analysées : 8

Décomposition des résultats

	Nb Phrases	NbPhrases(NbMots)
(drapeau français)	4	[1(3) 1(6) 1(10) 1(32)]
(drapeau britannique)	4	[2(4) 1(8) 1(10)]
(drapeau espagnol)	0	[ ]
(drapeau allemand)	0	[ ]
(drapeau italien)	0	[ ]

Analyse de la catégorisation en fonction de la longueur des phrases

Nb Mots	Nb Phrases	Indet	Hesit	OK	%indet	%hesit	%OK
3	1	0	0	1	0	0	100
4	2	0	0	2	0	0	100
6	1	0	0	1	0	0	100
8	1	0	0	1	0	0	100
10	2	0	0	2	0	0	100
32	1	0	0	1	0	0	100

FIG. A.2 - *Exemple de sortie du diagnostiqueur de langue*

Annexe B

La loi de Zipf

B.1 Présentation

George Kingsley ZIPF (1949) a montré qu'en classant les différents mots d'un texte par fréquence décroissante, on observe alors que leur fréquence est inversement proportionnelle à leur rang. La loi de ZIPF stipule que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. Cette loi peut s'exprimer de la manière suivante :

$$\text{Fréquence d'un mot de rang } i = \frac{\text{Fréquence du mot de rang 1}}{i}$$

De cette loi, on constate qu'en traçant pour chaque mot d'un texte le couple (rang,fréquence) dans un repère logarithmique, alors le nuage de points paraît linéaire. La répartition réelle des points montre que la linéarité n'est qu'approximative mais le phénomène est intéressant car indépendant des locuteurs, types de texte et langues.

En étudiant la liste des fréquences, on note que les mots les plus fréquents sont les mots grammaticaux, et que leur ordre d'apparition dans la liste est stable d'un texte à l'autre lorsque la longueur du texte est quelque peu conséquente. Les mots lexicaux apparaissent ensuite, ceux thématisant le document arrivant en tête.

B.2 Illustrations

À titre d'illustration, nous présentons quelques expériences exhibant les propriétés linguistiques mentionnées dans la section précédente. Ces expériences ont été rendues possibles grâce à une démonstration que nous avons conçue. Implémentée sous la forme d'une applet java (c.f., figure B.3 page 176), elle est accessible sur internet à l'URL suivante :

http://www.info.unicaen.fr/~giguet/java/zipf.html

B.2.1 Indépendance envers le type du texte

Notre première illustration a pour objectif la monstration de la stabilité relative de l'ordre d'apparition des mots d'un texte, quel qu'en soit le type.

Pour cela, nous avons fixé la langue des textes, le français, choisi de manière à faciliter la compréhension de certaines caractéristiques.

Quatre extraits de textes ont été sélectionnés : un extrait de livre scientifique «*La reconnaissance des formes*» (L. MICLET), un extrait d'article scientifique «*Rôles et transformations des pigments caroténoïdes dans les réseaux trophiques marins*» (M. VINCENT), un extrait de texte philosophique «*Discours sur l'origine et les fondements de l'inégalité parmi les hommes*» (mis à disposition par le CRI Philosophie), et un extrait de roman «*De la Terre à la Lune*» (Jules VERNE).

La liste des mots, classée par fréquence décroissante, a été calculée pour chacun des extraits. La définition du mot retenue est celle d'une séquence de caractères comprise en deux espaces ou ponctuations. Nous avons comptabilisé les séquences de ponctuations avec les mots afin d'en laisser apprécier la fréquence.

Ces listes, restreintes aux mots les plus fréquents dans le tableau B.1 page suivante, laisse apparaître la forte représentativité des mots grammaticaux et leur ordre relativement stable. On s'aperçoit que les seuls mots lexicaux présents reflètent la thématique des textes. On note l'importance systématique de la ponctuation, notamment du point et de la virgule, les autres ponctuations étant plus liées au type du texte (parenthèses et deux points dans les textes scientifiques ; tirets de dialogue, points d'exclamation, points d'interrogation et guillemets dans le roman).

Nous avons tracé, pour chacun des extraits, les nuages de points correspondants aux différents couples (rang,fréquence) pour en apprécier la linéarité (c.f., figure B.1 page 172 et page 173). Sur chaque graphique sont inscrites à titre indicatif l'équation de la droite de régression linéaire et la fréquence du mot de rang 1 afin de juger l'approximation de l'équation générale :

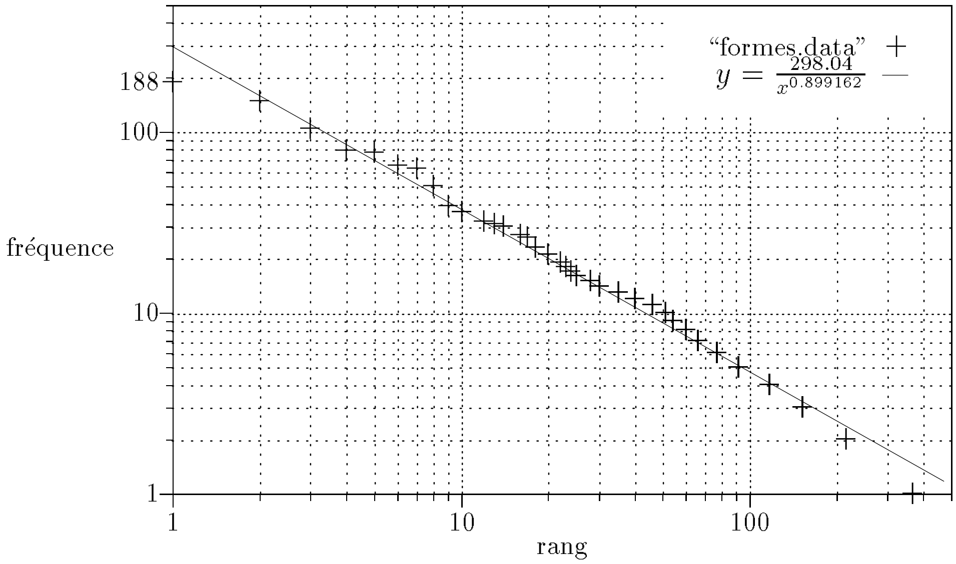
$$\text{Fréquence d'un mot de rang } i = \frac{\text{Fréquence du mot de rang 1}}{i}$$

Livre scientifique (3195 mots)	Article scientifique (1873 mots)	Discours philosophique (3353 mots)	Roman (2794 mots)
de	,	,	,
,	de	de	de
.	la	et	.
la	.	la	les
des	les	les	et
les	des	à	à
à	et	que	!
et	(	.	en
un	)	le	le
en	pigments	des	des
est	caroténoïdes	qui	la
une	dans	ne	un
dans	ou	dans	–
(	sont	;	du
)	à	en	pas
par	chez	l'homme	qui
le	]	plus	dans
que	[	se	ne
pour	le	qu'il	se
ou	que	est	une
sur	en	un	il
formes	Les	avec	par
représentation	est	pour	au
on	:	point	que
très	par	il	plus
se	une	si	?
plus	qui	ses	ces
deux	formes	nature	Maston
:	plus	pas	«
d'un	peut	leur	ce

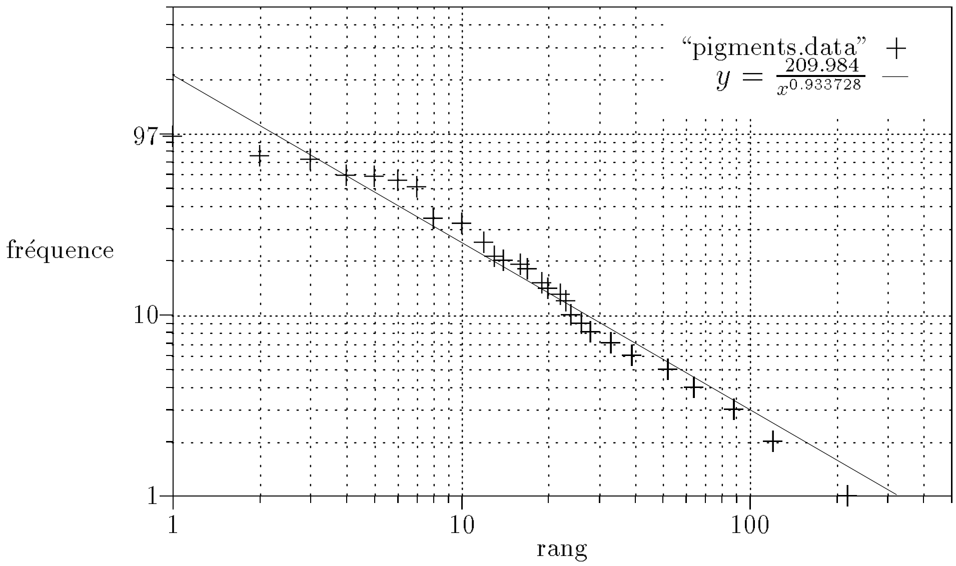
TAB. B.1 - *Mots les plus fréquents de quatre textes français de type différent*

B.2.2 Indépendance envers la langue du texte

Notre seconde illustration réitère l'opération précédente en faisant varier non plus le type des textes mais leur langue. Nous avons sélectionné quatre langues, l'anglais, l'espagnol, l'italien et l'allemand et tracé les nuages de points et la droite de régression linéaire suivant le protocole décrit ci-dessus. Les résultats sont donnés figure B.2 page 174 et page 175.

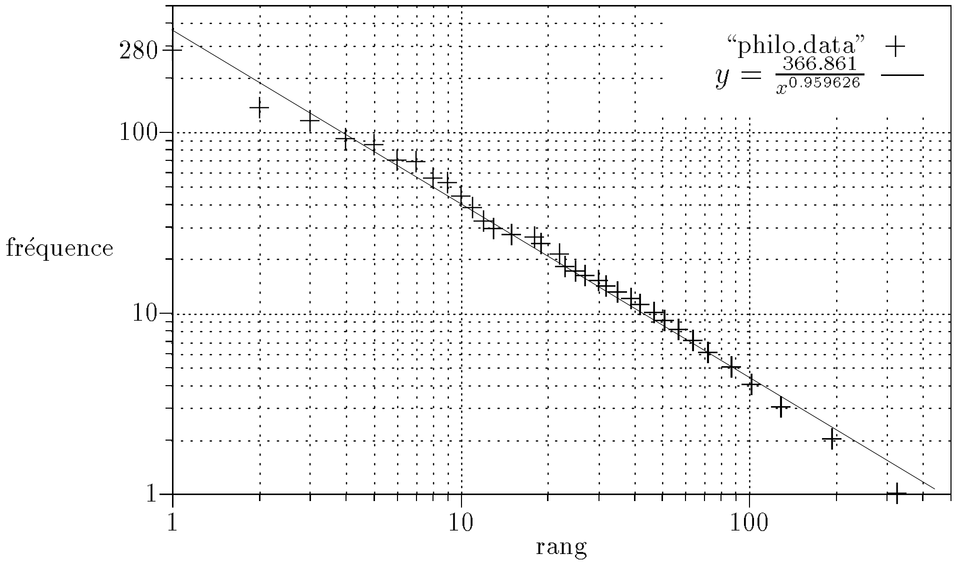


(a) Livre scientifique (3195 mots)

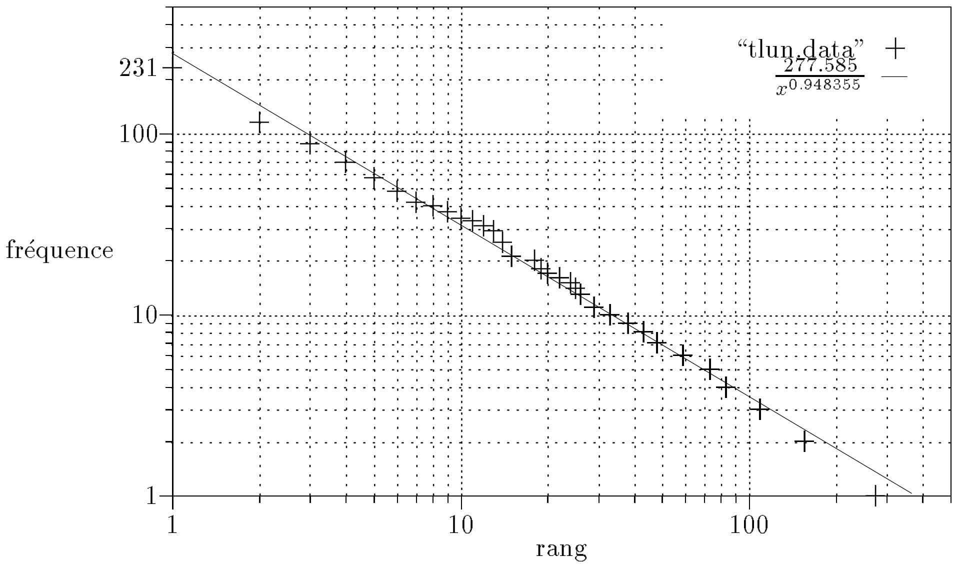


(b) Article scientifique (1873 mots)

FIG. B.1 - *Indépendance du type du texte*

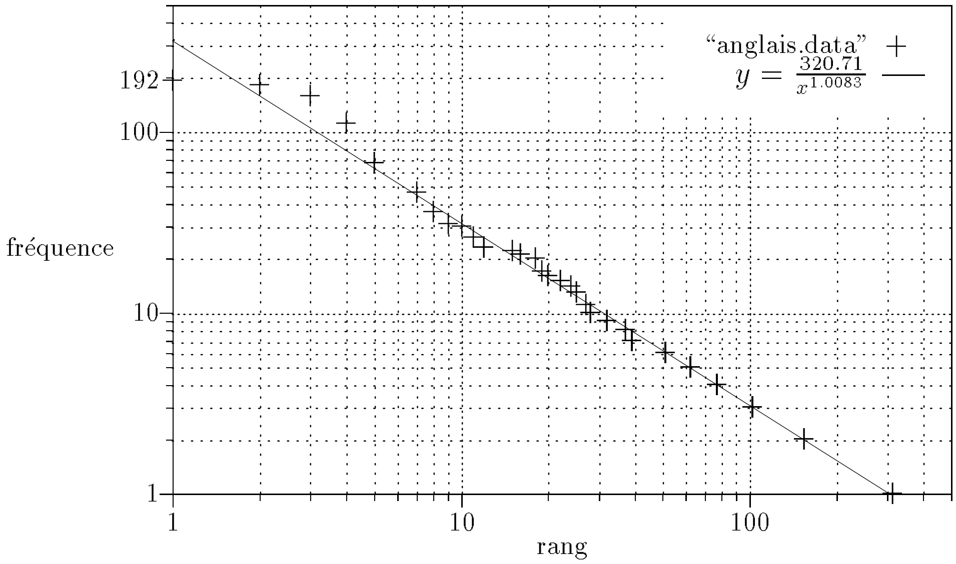


(c) Discours philosophique (3353 mots)

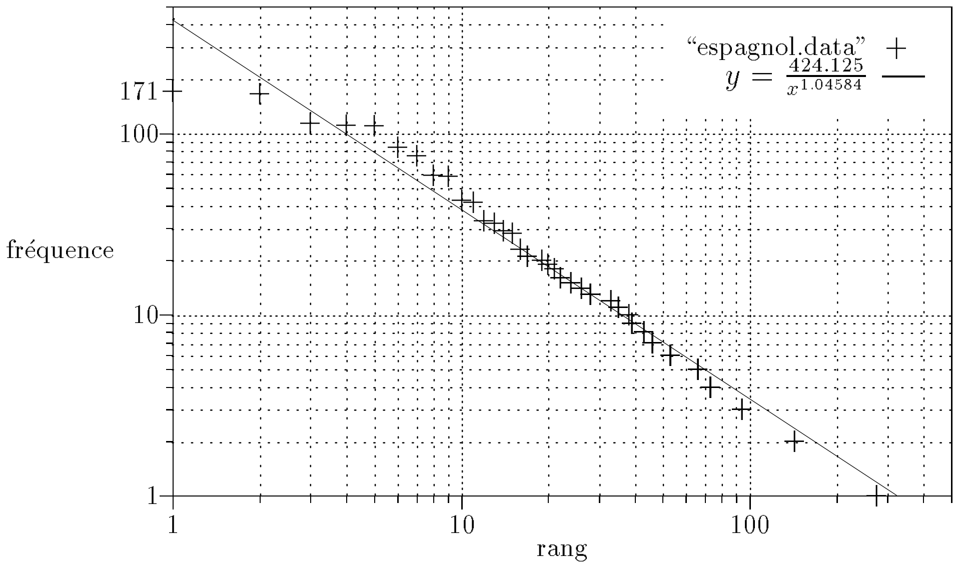


(d) Roman (2794 mots)

FIG. B.1 - *Indépendance du type du texte*

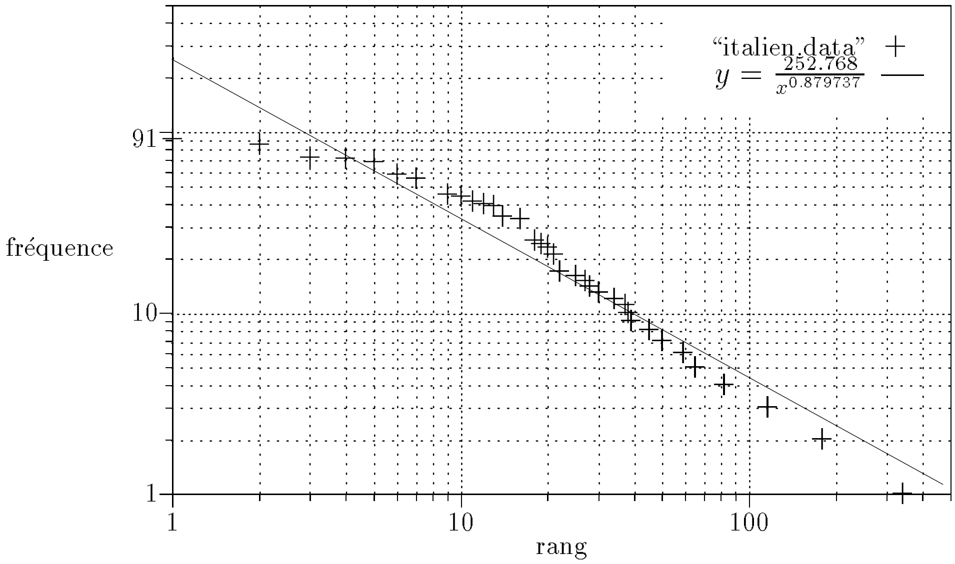


(a) Anglais (2841 mots)

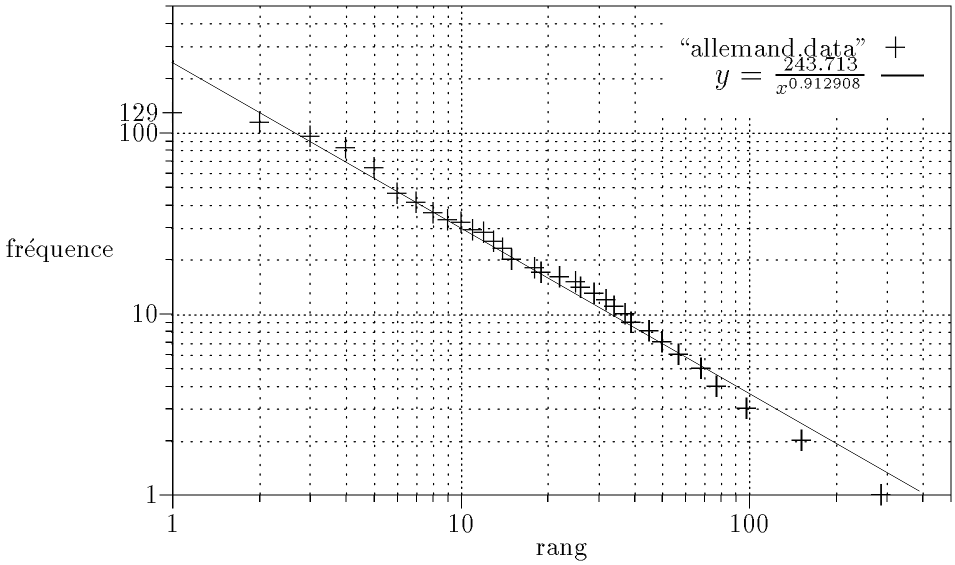


(b) Espagnol (2758 mots)

FIG. B.2 - *Indépendance de la langue du texte*

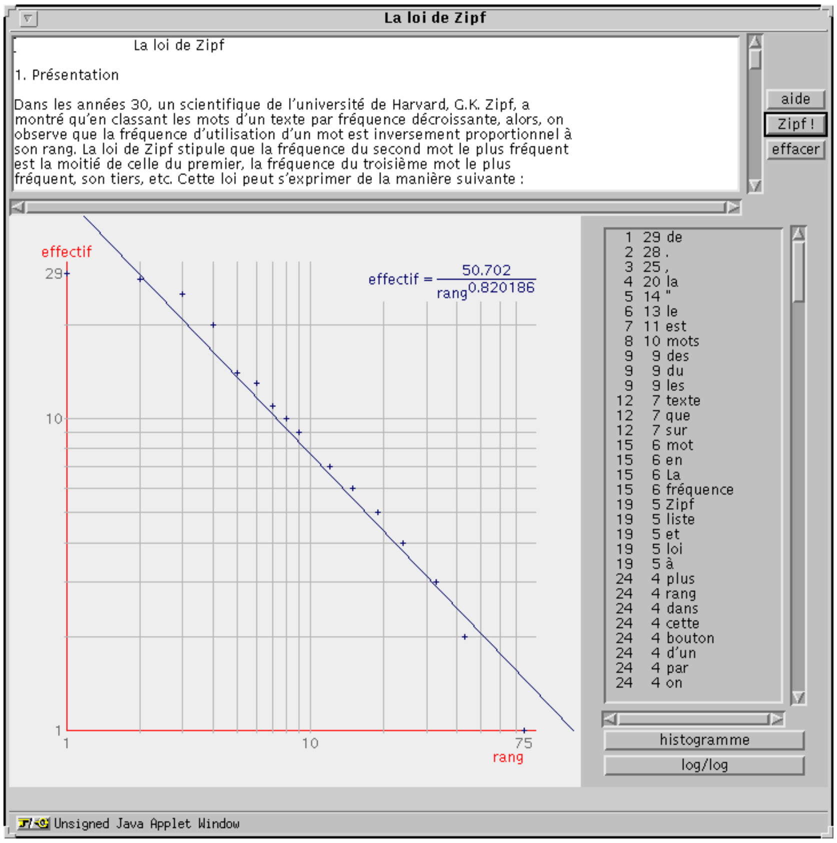


(c) Italien (3204 mots)



(d) Allemand (2424 mots)

FIG. B.2 - *Indépendance de la langue du texte*



http://www.info.unicaen.fr/~giguet/java/zipf.html

FIG. B.3 - *L'interface d'illustration de la loi de* ZIPF

Annexe C

Exemple de jeu de catégories distributionnel

C.1 Introduction

Nous présentons dans cette annexe le jeu de catégories distributionnel que nous utilisons dans notre analyseur. Ce jeu n'est donné qu'à titre d'exemple. Il peut certainement être amélioré mais dans son état actuel, il concrétise correctement le concept de distribution allié au concept de syntagme minimal.

Certains verront dans ce jeu de catégories des possibilités de factorisation, pouvant se traduire par l'ajout de traits sous-catégorisants, cela afin de diminuer sa taille. Nous n'avons cependant pas choisi cette option, d'une part parce que seul le critère distributionnel a déterminé la distinction entre les catégories, et d'autre part parce que cela nous permet de coder chaque catégorie par un caractère unique, et, qu'une fois les mots étiquetés, il nous est alors possible de raisonner sur une représentation abstraite résultant de la concaténation des étiquettes des mots. Concrètement, dans l'exemple 5 page 127, délimité en syntagmes, «*[La base] que [la SSII] [développe] [accélérera] [les traitements].*», après étiquetage nous pouvons travailler sur une représentation abstraite très simple, à savoir :

`dS O dS V V dS`

C.2 Le jeu de catégories distributionnel

- **Catégories apparaissant dans un contexte verbal :**

- verbe transitif infinitif
- verbe transitif conjugué
- verbe transitif participe présent
- verbe transitif participe passé
- verbe intransitif infinitif
- verbe intransitif conjugué
- verbe intransitif participe présent
- verbe intransitif participe passé
- auxiliaire avoir infinitif
- auxiliaire avoir conjugué
- auxiliaire avoir participe présent

- auxiliaire avoir participe passé
- auxiliaire et copule être infinitif
- auxiliaire et copule être conjugué
- auxiliaire et copule être participe présent
- auxiliaire et copule être participe passé
- auxiliaire pouvoir et vouloir infinitif
- auxiliaire pouvoir et vouloir conjugué
- auxiliaire pouvoir et vouloir participe présent
- auxiliaire pouvoir et vouloir participe passé
- négation préverbale
- négation postverbale
- adverbe antéposé d'attribut ou de participe passé
- adverbe postposé de verbe
- pronom préverbal sujet
- pronom préverbal objet
- pronom préverbal non-objet
- pronom postverbal
- adjectif attribut
- préposition introduisant un infinitif
- préposition introduisant un participe présent

• Catégories apparaissant dans un contexte apparenté verbal :

- participe passé sans auxiliaire ou disjoint de l'auxiliaire
- adjectif épithète disjoint

•Catégories apparaissant dans un contexte nominal :

- substantif
- nom propre
- pronom tonique
- adjectif épithète
- adjectif antéposé (indéfini, numéral ou cardinal)
- déterminant

- partitif
- préposition introduisant du nominal
- coordination d'épithètes
- adverbe d'épithète
- adjectif ou nom
- pronom relatif sujet
- pronom relatif objet
- pronom relatif groupe prépositionnel

- **Catégories externes aux contextes nominal et verbal :**

- conjonction de subordination
- coordination
- parenthèse ouvrante ou deux points
- parenthèse fermante
- virgule
- adverbe de phrase ou de syntagme

Annexe D

Le processus de mise en relation

D.1 Évaluation du processus de mise en relation

D.1.1 Objectif de l'évaluation

L'évaluation que nous proposons a pour but d'étudier la pertinence du processus de mise en relation basé sur la création et la suppression d'attentes de relations syntaxiques. Cette évaluation est restreinte au calcul de la relation sujet-verbe car aucun corpus français annoté en terme de relations syntaxiques n'est actuellement disponible.

Cette évaluation pourrait paraître trop restrictive pour être représentative. Elle satisfait cependant notre objectif car, l'ensemble des relations interagissant au travers de créations et de suppressions d'attentes, une relation telle que la relation sujet-verbe ne peut être correctement instanciée que si le calcul des autres relations a permis de suffisamment réduire la combinatoire des candidats potentiellement concurrents et si ces calculs ont été assez correctement effectués pour ne pas déclencher la suppression du bon candidat. Par ailleurs, la délimitation des domaines propositionnels étant nécessaire au calcul de la relation sujet-verbe, nous pouvons vérifier simultanément s'ils ont été respectés.

D.1.2 Le corpus testé

L'évaluation de l'analyseur a été mené sur un ensemble d'articles extraits du journal «Le Monde», corpus n'ayant pas servi à la mise au point des règles de mise en relation. Vingt-quatre articles traitant de sujets aussi variés que la politique, l'économie, la mode, la haute-technologie et la vie quotidienne, composent le corpus.

La figure D.1 page ci-contre présente l'histogramme des longueurs de phrase du corpus. Il contient 474 phrases de longueur moyenne 24,43 mots, la longueur maximale étant de 82 mots. Notons que les points-virgule et les deux-points sont comptés comme séparateurs de phrase.

738 relations sujet-verbe ont été manuellement comptabilisés. Nous les avons catégorisées en 4 classes selon la nature du sujet : les relations impli-

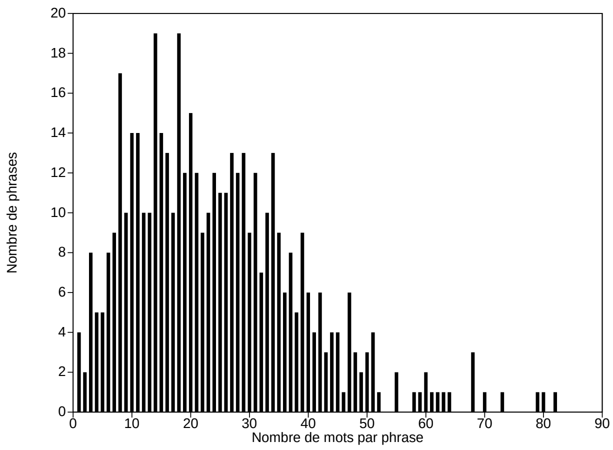


FIG. D.1 - *Longueur des 474 phrases du corpus*

quant (1) un sujet nominal, (2) un sujet infinitif verbal, (3) un sujet pronom relatif et (4) un sujet pronom personnel.

nature du sujet	nb
nominal	458
infinitif verbal	2
pronom relatif	85
pronom personnel	193
Total	738

La figure D.2 page suivante montrent la répartition des relations sujet-verbe dans les phrases : 39 phrases ne comportent pas de relation sujet-verbe et le nombre maximal de relations sujet-verbe par phrase est de 6.

Nous présentons une autre métrique : la distance entre un verbe et son sujet. La figure D.3 page suivante illustre cette distance pour les sujets nominaux. Elle montre que la distance entre un verbe et son sujet peut atteindre 45 mots pour une relation standard et 8 mots dans une relation inversée, d'où l'intérêt d'effectuer le calcul des relations sur des syntagmes minimaux, ce qui réduit le nombre de candidats potentiels pour une relation. Notons qu'une relation sujet-verbe impliquant un sujet pronom relatif ou pronom personnel n'implique pas nécessairement la contiguïté des éléments reliés, certaines relations peuvent être de plus longue distance, par exemple en cas d'énumération de verbes ou d'insertion de groupe prépositionnel.

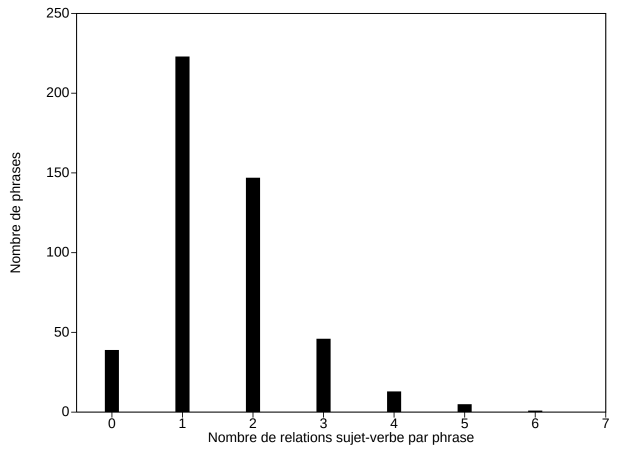


FIG. D.2 - *Les relations sujet-verbe dans le corpus*

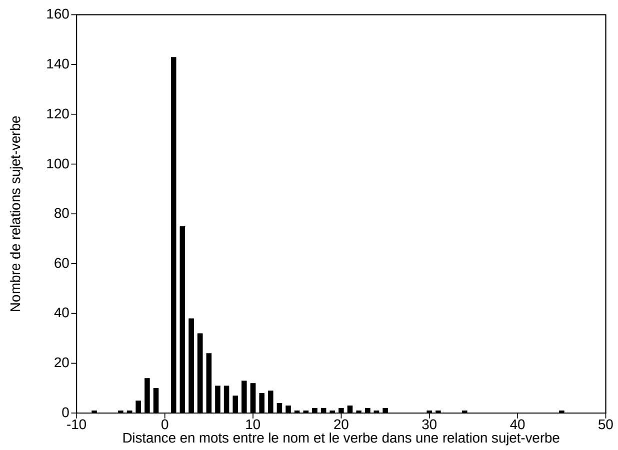


FIG. D.3 - *Distance entre un sujet nominal et son verbe dans le corpus*

D.1.3 Évaluation du calcul de la relation sujet-verbe

Protocole d'évaluation

La fonction d'évaluation est basée sur le principe suivant : chaque verbe ne doit être relié qu'à un seul sujet. De ce principe, nous considérons 3 situations : une relation est dite *correcte* si le verbe est relié au sujet attendu et que les têtes des deux syntagmes sont correctement calculées, *incorrecte* dans le cas contraire et *manquante* si un sujet était attendu mais qu'aucun n'a été calculé.

Dans le cas de coordination de sujets, chaque verbe ne doit être relié qu'à la tête de la coordination, c'est-à-dire à l'élément apparaissant le plus à gauche dans l'ordre linéaire. Dans le cas de coordination de verbes, chaque verbe de la coordination doit être relié au sujet attendu.

Résultat de l'évaluation

Les résultats sont présentés tableau D.4. La *précision* correspond au nombre de relations correctes par rapport au nombre de relations calculées. Le *rappel* indique le nombre de relations correctes par rapport au nombre de relations attendues.

Les taux que nous obtenons sont très encourageants car le système est en cours de développement. Sachant que toutes les relations interagissent au travers des mémoires, il faut noter que calculer une relation sujet-verbe implique ne pas faire trop d'erreurs dans le calcul des autres relations. On notera particulièrement le cas du calcul des relations sujet-verbe sur verbes coordonnés qui est une simple déduction de l'existence de la relation de coordination des verbes mais qui implique une résolution correcte de la coordination.

Nature du sujet	nb.	correcte	incorrecte	manquante	précision	rappel
nominal	458	418	26	14	94.14%	91.27%
infinitif verbal	2	2	0	0	100.00%	100.00%
pron. relatif	85	85	0	0	100.00%	100.00%
pron. personnel	193	191	0	2	100.00%	98.96%
Total	738	694	26	16	96.39%	94.04%

FIG. D.4 - *Résultat de l'évaluation du calcul de la relation sujet-verbe*

Ces résultats peuvent bien entendu être améliorés car cette évaluation est la première menée sur large corpus. Les 42 relations manquantes et incorrectes ont pour origine 5 classes de causes : (1) une implémentation incorrecte de la vérification de l'accord en personne et en nombre, (2) des syntagmes minimaux mal construits, (3) des coordinations non ou mal détectées, (4) des relations sujet-verbe inversées non détectées dans des discours rapportés, (5) des syntagmes nominaux mal étiquetés.

Cette évaluation contribue à la validation du processus dans la mesure où l'on ne note que peu d'erreurs mettant en cause directement le processus de mise en relation. Dans la majorité des cas, l'interaction des différentes relations est satisfaisante puisqu'elles autorisent le calcul de la bonne relation sujet-verbe. Les domaines propositionnels sont quant à eux relativement bien respectés. Le comportement du processus est toujours homogène, quelle que soit la complexité de la phrase. Nous notons quelques rares cas de réelles dégradations lorsque la frontière de deux propositions juxtaposées n'a pas été détectée car toutes les attentes de la première proposition se trouvent alors ouvertes à toutes les unités de la proposition suivante.

Aujourd'hui, pour améliorer le calcul d'une relation, c'est d'une part vers l'amélioration du calcul de l'ensemble des relations qu'il faut se tourner puisqu'elles interagissent les unes avec les autres, mais surtout vers la prise en compte réelle du niveau propositionnel que nous avons jusqu'à présent trop négligé en nous restreignant à une délimitation un peu légère de ces unités. Il nous faudra notamment étudier plus finement les marques de fin d'unités propositionnelles.

Pour avoir une idée de la pertinence du processus quant à sa capacité à calculer les autres relations, nous vous invitons à parcourir les nombreux corpus analysés mis à disposition sur notre site internet :

http://www.info.unicaen.fr/~giguet/syntaxique.html

Performances d'autres systèmes

Xerox a évalué un analyseur du français basé sur les technologies à états finis (Aït-Mokhtar et Chanod, 1997). L'évaluation porte sur la détermination des sujets syntaxiques. Sur un corpus extrait du journal «Le Monde» et constitué de 279 phrases, leur analyseur obtient une précision de 92.6% et un rappel de 82.6%.

Tapanainen et Järvinen (1997) ont évalué leur analyseur de l'anglais basé sur le modèle de la dépendance. Pour la relation sujet-verbe, les taux de précision et de rappel sont de 95% et 89% sur un corpus journalistique (136 phrases de «The independent»), 98% et 92% sur un corpus littéraire (195 phrases de «British Book data»), 95% et 89% sur un corpus de radio-diffusion (244 phrases de «American National Public Radio»).

D.2 Exemples d'analyses syntaxiques

Afin de concrétiser les possibilités offertes par un tel processus de mise en relation, nous avons sélectionné des phénomènes syntaxiques variés, résolus de manière uniforme. Ces phénomènes sont la relation sujet-verbe, la relation de coordination, la catégorie syntaxique de «*que*» (conjonction de subordination, conjonction de coordination, pronom relatif ou adverbe), la catégorie syntaxique de «*de*» (préposition ou partitif).

Pour chaque phénomène, nous vous invitons à examiner une sélection de phrases analysées, extraites de corpus réels. Les analyses sont prises en sortie d'analyseur, sans aucune retouche manuelle ; vous constaterez que certaines relations sont erronées, d'autres sont manquantes.

Les graphes de relations ont été générés par un visualiseur d'analyses que nous avons développé et qui sera rapidement présenté dans la section suivante. Le rendu des différentes analyses a été modifié manuellement de manière à faire apparaître en gras les phénomènes syntaxiques pertinents. Lorsque le phénomène a été mal géré par le processus, l'analyse est préfixée du symbole *.

D.2.1 Notation

Afin d'alléger les sorties et de permettre une focalisation sur les phénomènes syntaxiques, les graphiques sont proposés en version simplifiée. Seules les catégories des groupes minimaux et les étiquettes des relations sont représentées. La notation est la suivante :

Étiquettes des relations	
`SV`	relation sujet-verbe
`VO`	relation verbe-objet
`Ant`	relation d'antécédence
`Coord`	relation de coordination
`DV`	rattachement d'un dépendant du verbe
`DN`	rattachement d'un dépendant du nom
`Sub`	rattachement d'une subordonnée conjonctive
`Rel`	rattachement d'une subordonnée relative

Catégories sur groupes minimaux			
N	nominal	w	adverbe
Aj	adjectival	p	préposition
V	verbal conjugué	o	partitif
I	verbal infinitif	pr	pronom relatif
PP	verbal participe présent	c	coordonnant
cs	conjonction de subordination	pp	pronom personnel sujet

Plusieurs catégories peuvent apparaître simultanément sur un groupe minimal, elles sont alors reliées par le symbole +. Par exemple, un groupe prépositionnel est catégorisé p+N.

D.2.2 Analyses de relations sujet-verbe

- Relations de longue distance :
  analyses 1, 2, 3, 4, 5 et 6

- Relations avec un sujet inversé :
  analyses 7 et 8

- Relation avec un sujet pronom relatif :
  analyse 9

- Relations avec des verbes coordonnés :
  analyses 10 et 11

- Relations dans une phrase interrogative :
  analyses 12 et 13

- Relation avec une proposition infinitive sujet :
  analyse 14

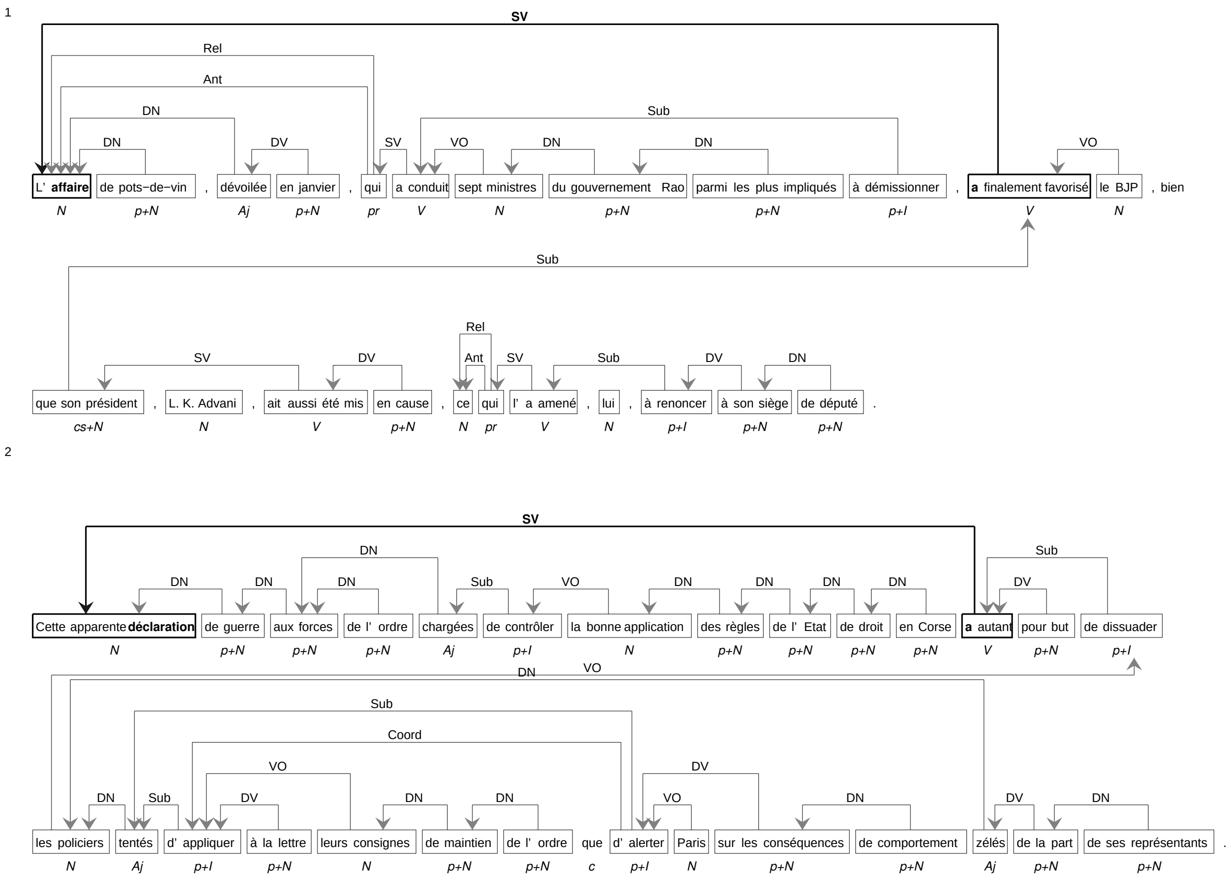


FIG. D.5 - *Analyses de relations sujet-verbe (a)*

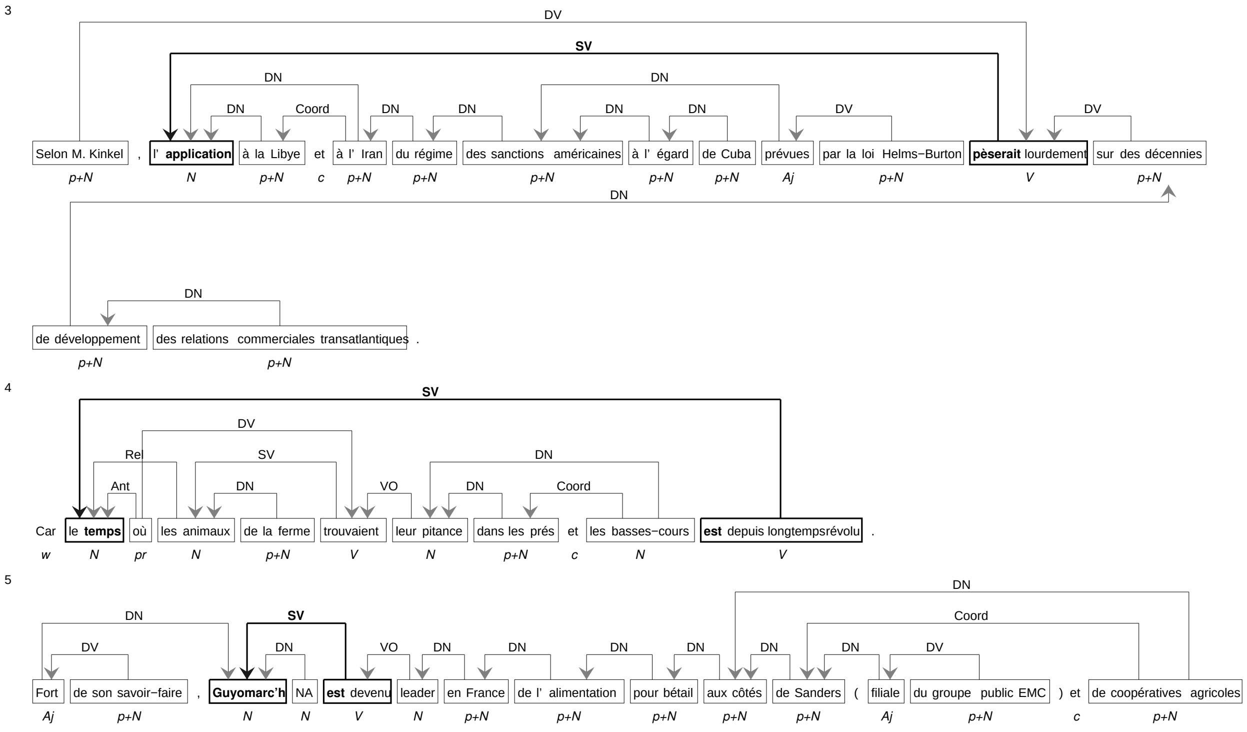


FIG. D.6 - *Analyses de relations sujet-verbe (b)*

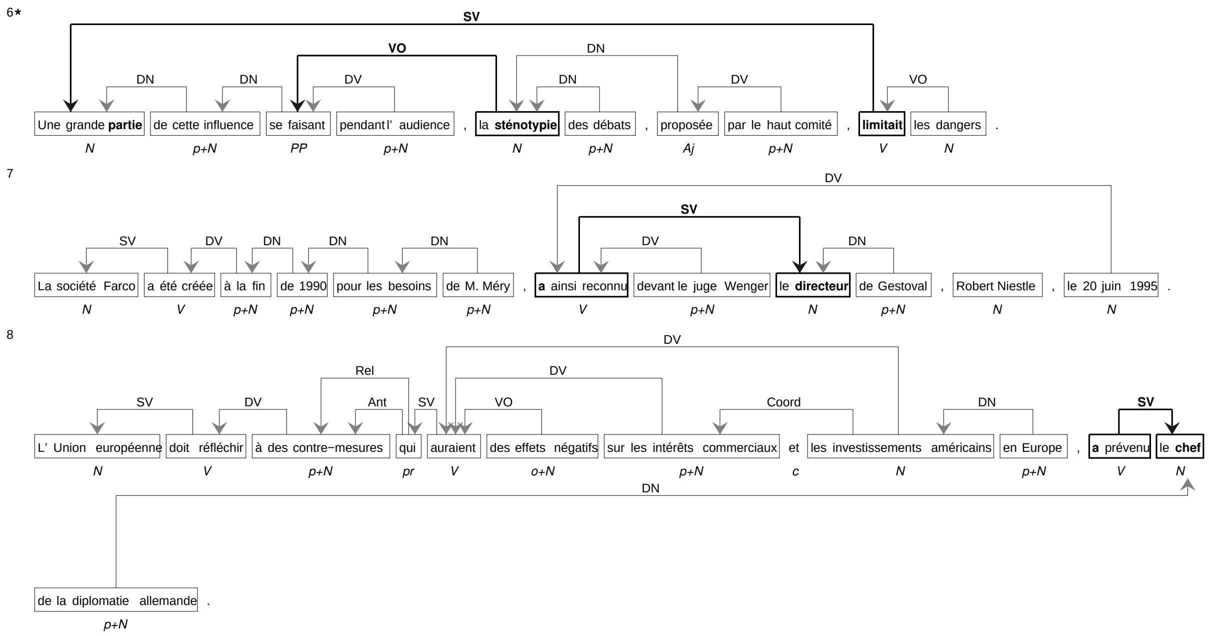


FIG. D.7 - *Analyses de relations sujet-verbe (c)*

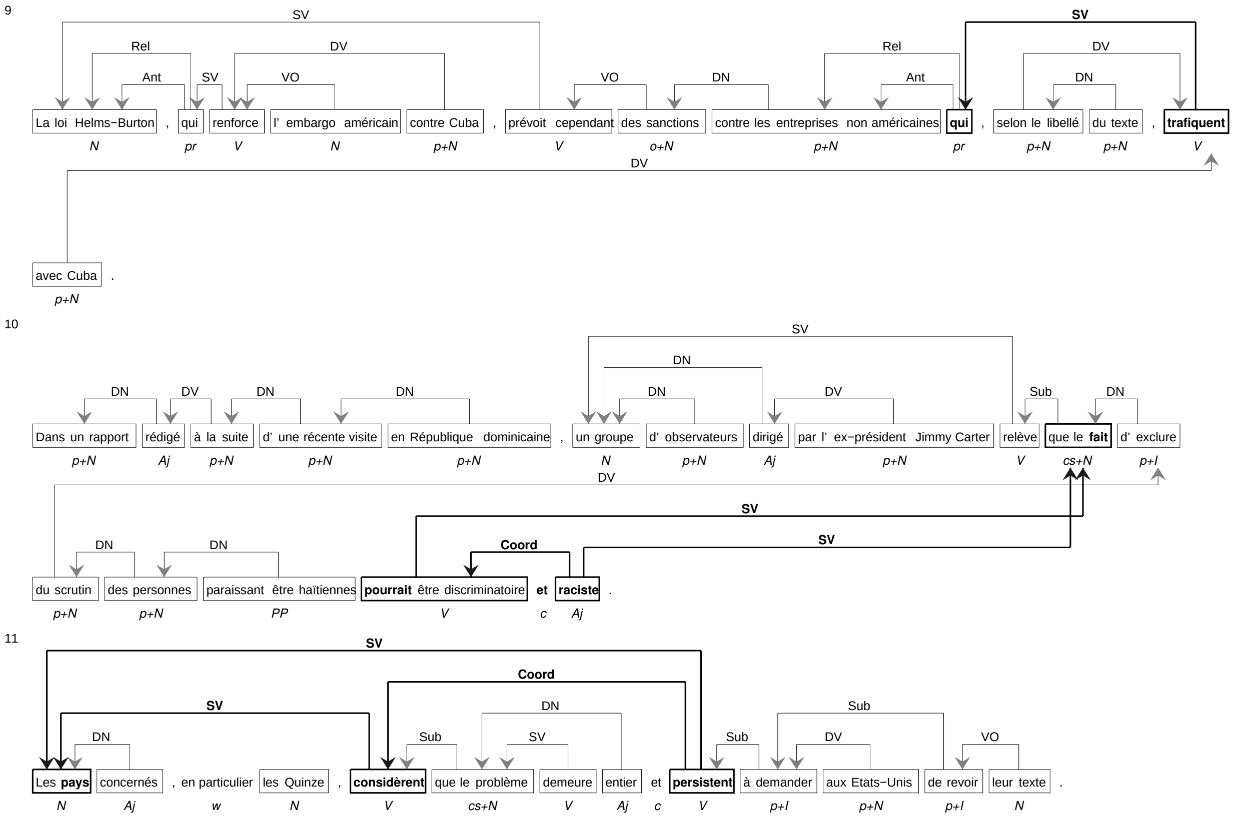


FIG. D.8 - *Analyses de relations sujet-verbe (d)*

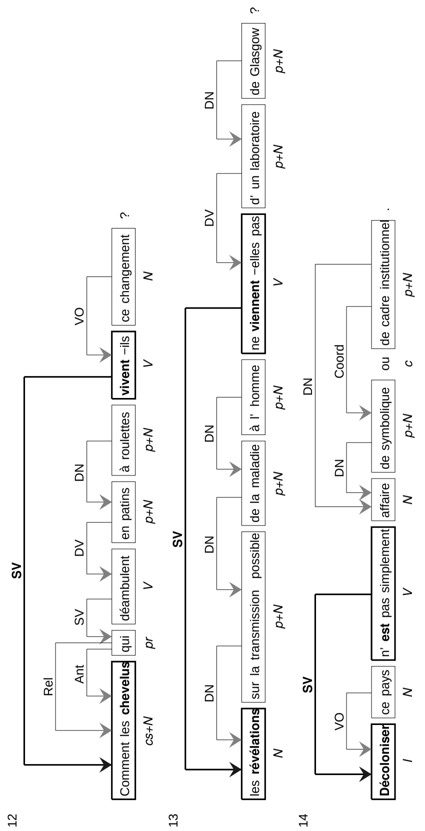


FIG. D.9 - *Analyses de relations sujet-verbe (e)*

D.2.3 Analyses de coordinations

- Coordinations de verbes conjugués :
  analyses 1, 2 et 3
- Coordinations d'objets directs :
  analyses 4 et 5
- Coordination de noms :
  analyse 6
- Coordination de groupes prépositionnels :
  analyse 7
- Coordination de subordonnées relatives :
  analyse 8
- Coordinations de subordonnées conjonctives :
  analyses 9 et 10
- Coordinations de subordonnées infinitives :
  analyses 11 et 12
- Coordination de subordonnées de catégories différentes :
  analyse 13
- Coordinations en énumération :
  analyses 14, 15 et 16

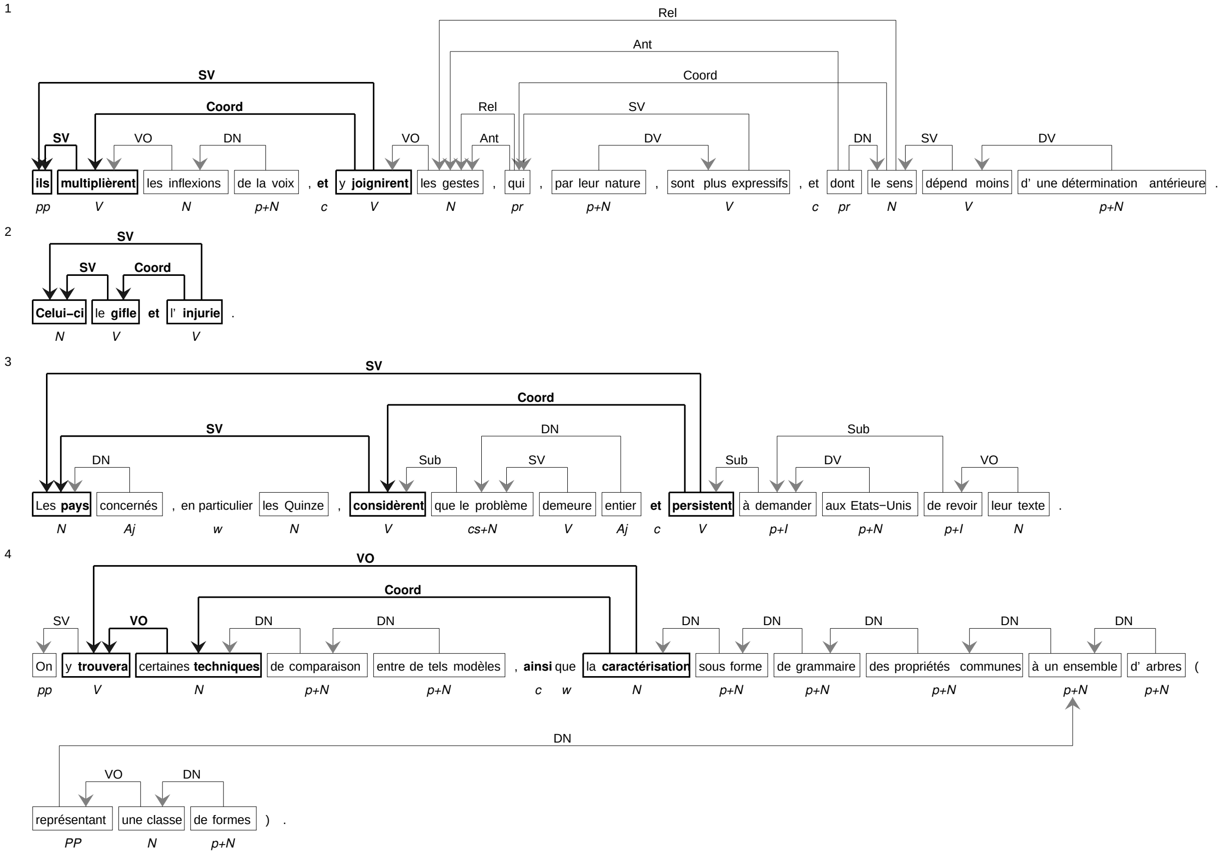


FIG. D.10 - *Analyses de coordinations (a)*

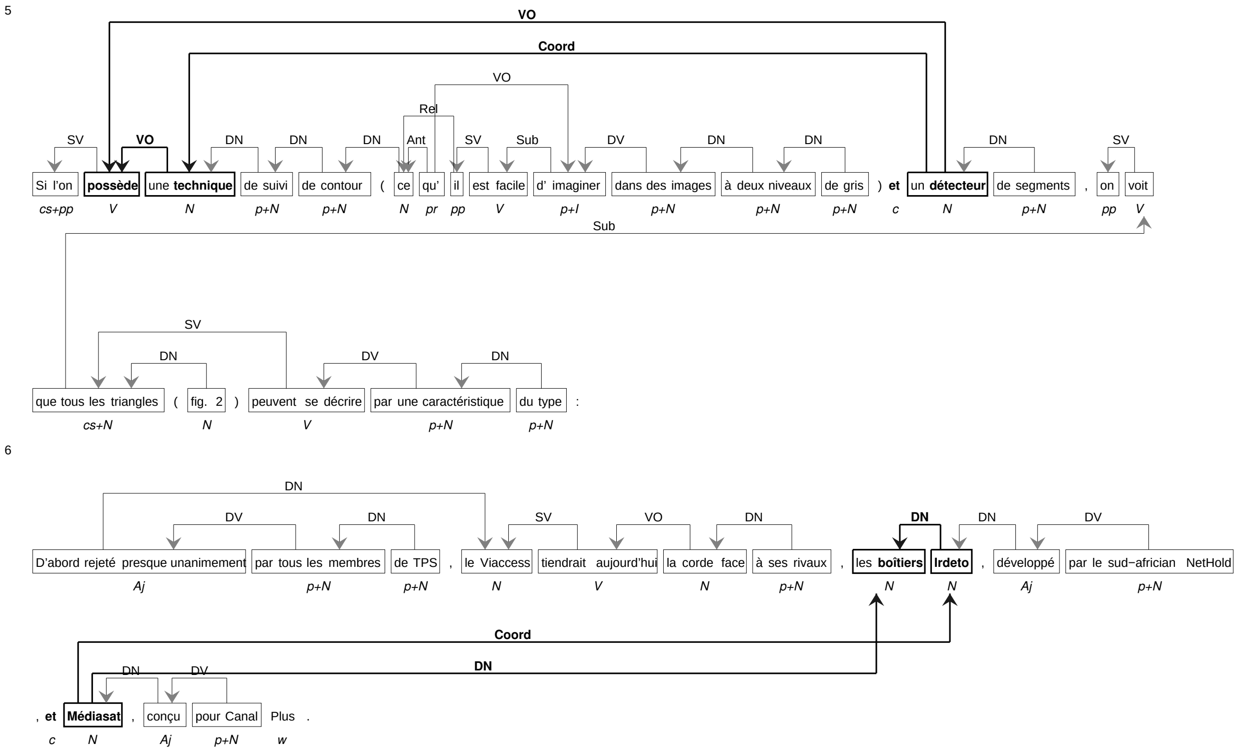


FIG. D.11 - *Analyses de coordinations (b)*

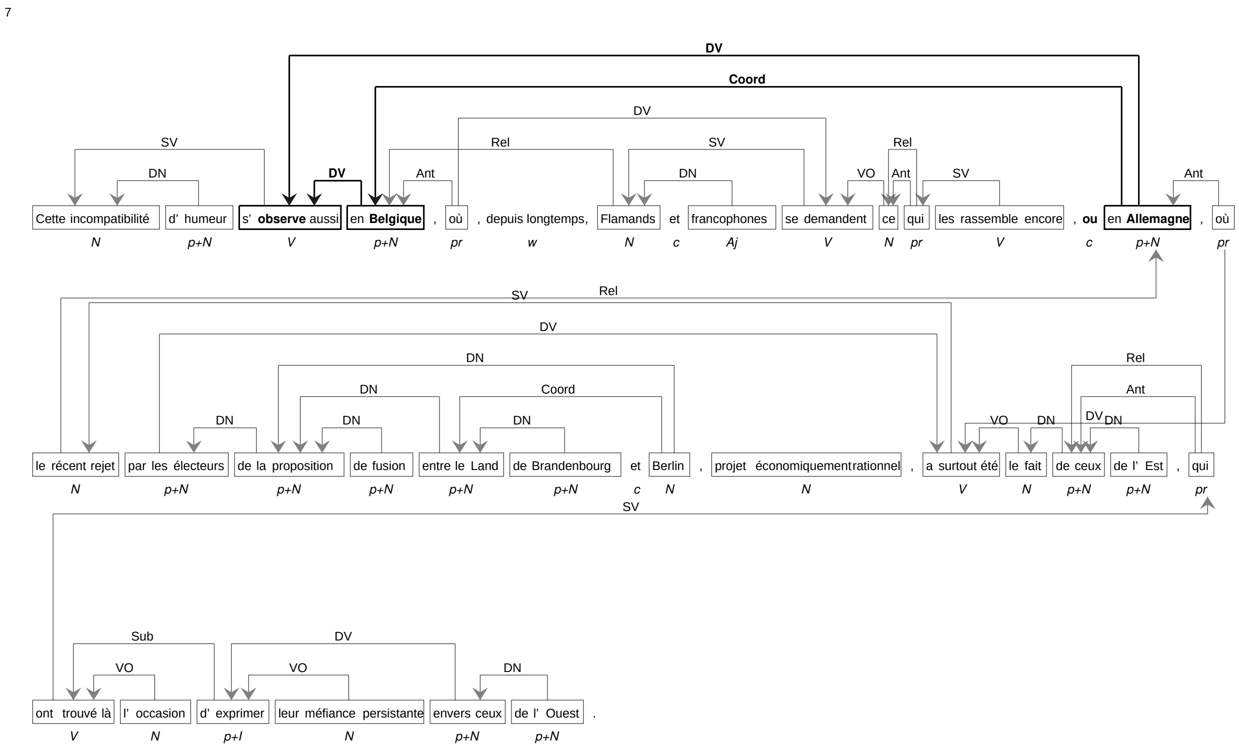


FIG. D.12 - *Analyses de coordinations (c)*

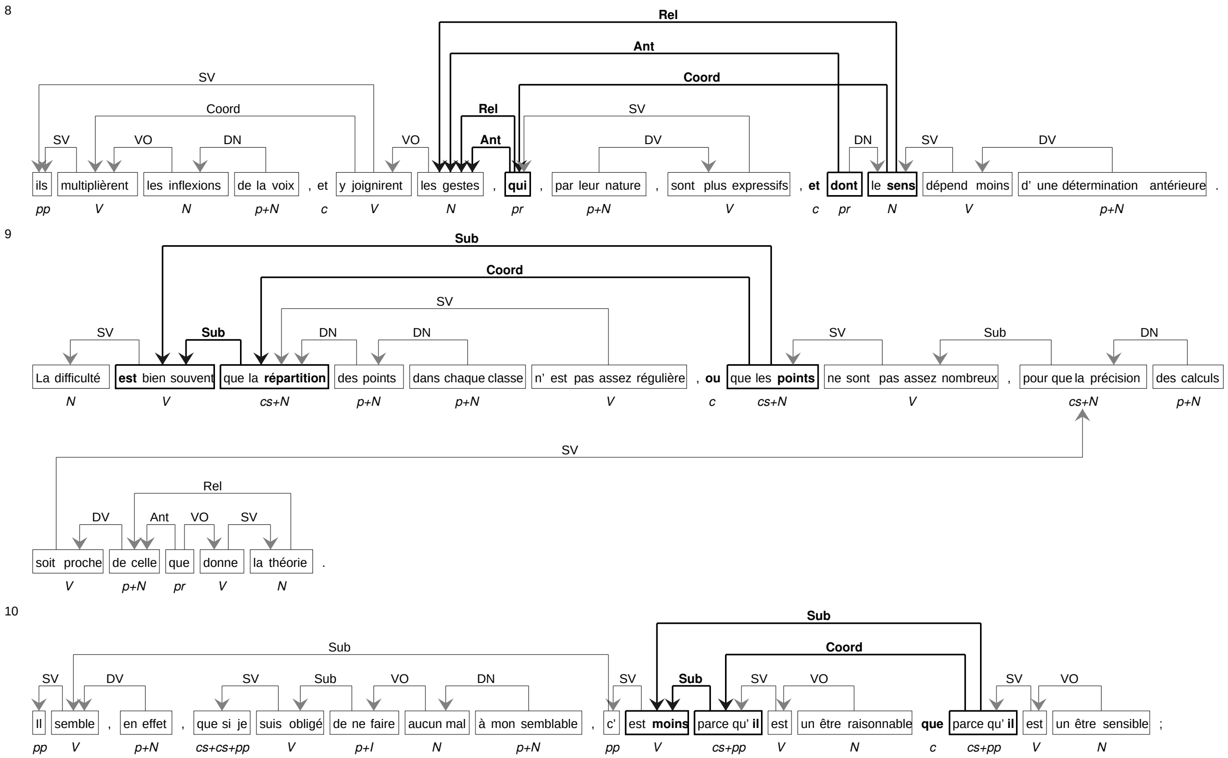


FIG. D.13 - *Analyses de coordinations (d)*

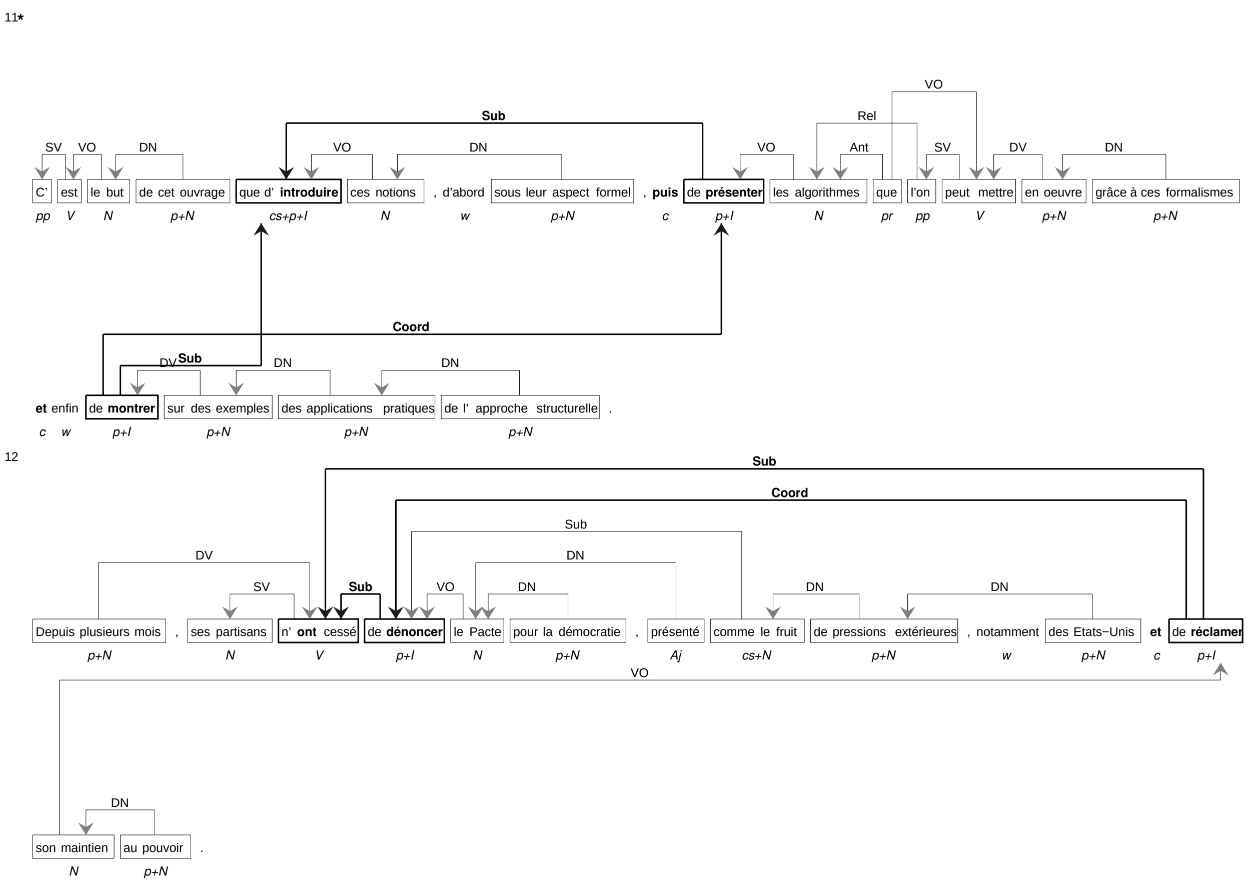


FIG. D.14 - *Analyses de coordinations (e)*

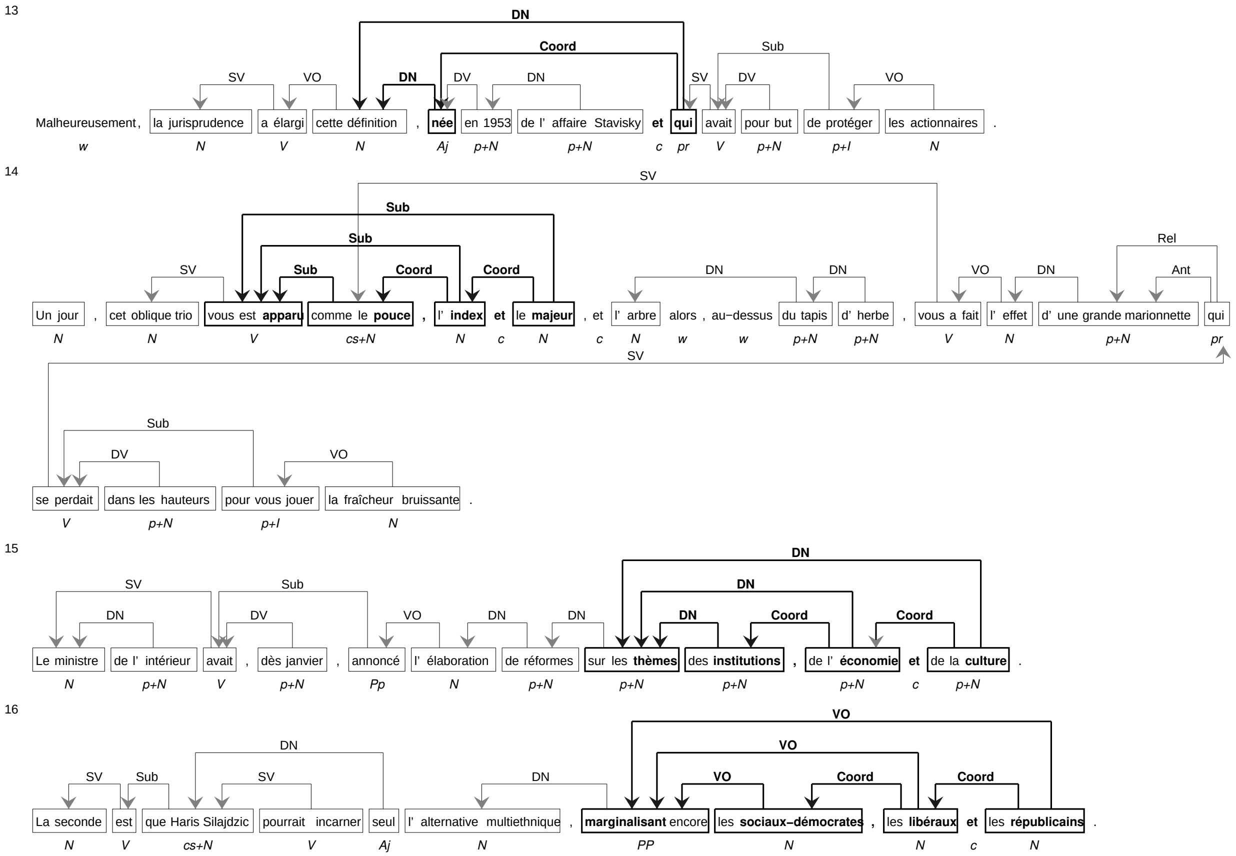


FIG. D.15 - *Analyses de coordinations (f)*

D.2.4 Analyses de la catégorie syntaxique de «*que*»

- conjonction de subordination (cs) :
  analyses 1 et 2
- pronom relatif (pr) :
  analyses 3 et 4
- coordonnant (c) :
  analyses 6, 7 et 8
- adverbe (av) :
  analyses 5, 9, 10 et 11

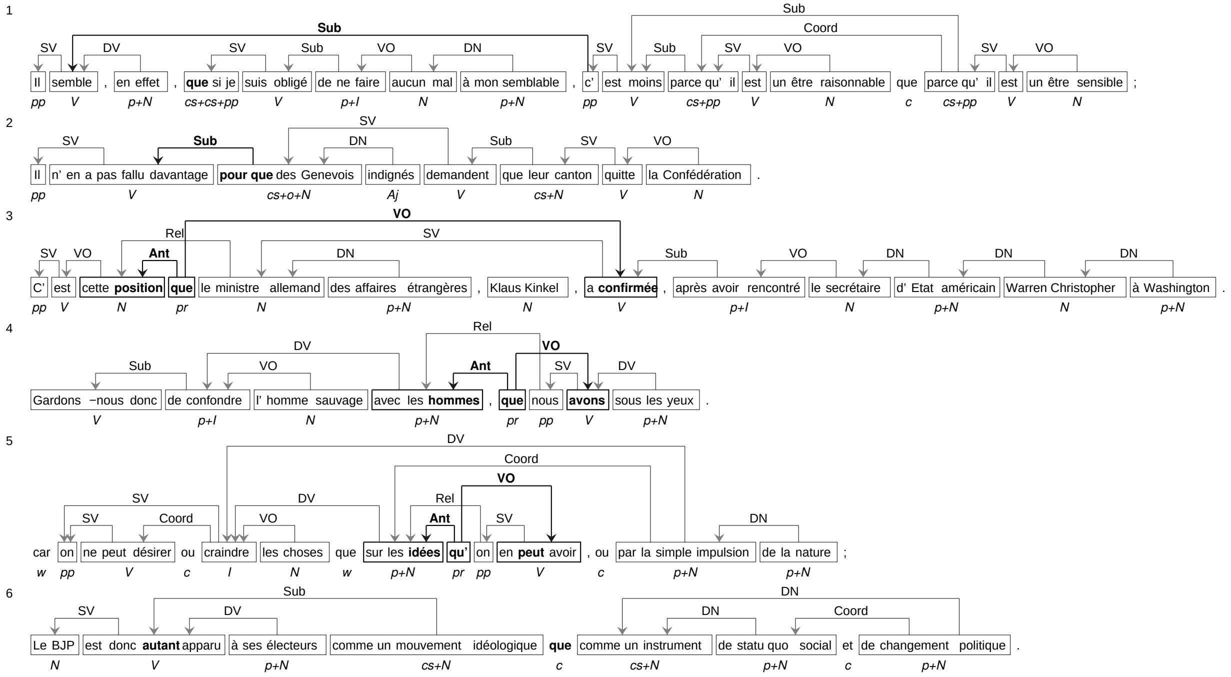


FIG. D.16 - *Analyses de la catégorie syntaxique de «que» (a)*

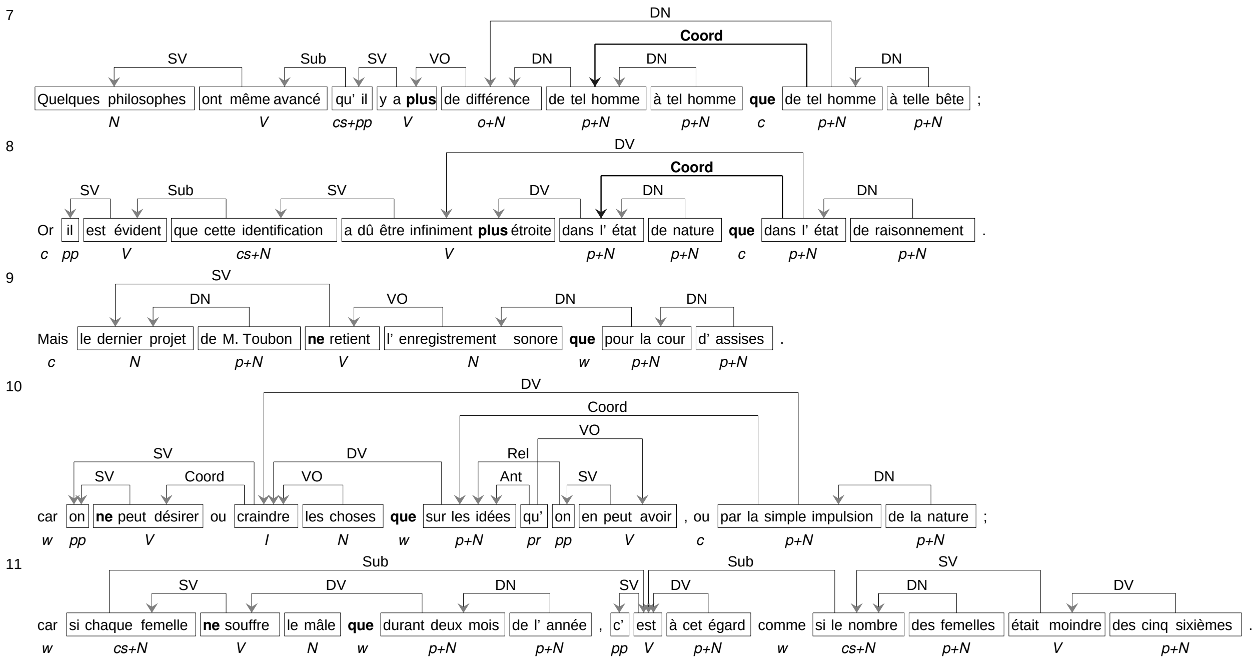


FIG. D.17 - *Analyses de la catégorie syntaxique de «que» (b)*

D.2.5 Analyses de la catégorie syntaxique de «*de*»

Un syntagme minimal est étiqueté *o+N* dans le cas où «*de*» est partitif, et *p+N* dans le cas où il est préposition.

- Dans une relation sujet-verbe :
  analyses 1 et 2
- Dans une relation verbe-objet :
  analyses 3, 4, 5, 6 et 7

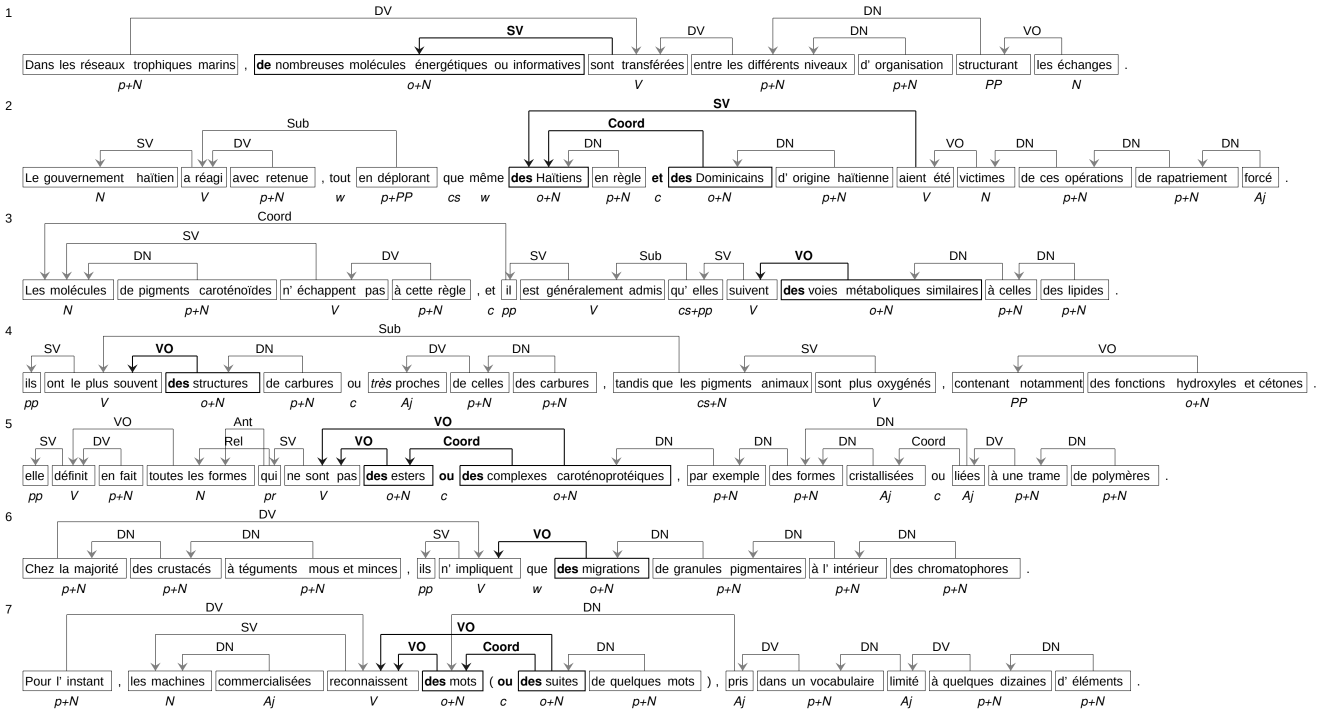


FIG. D.18 - *Analyses de la catégorie syntaxique de «de»*

D.3 Le visualiseur d'analyses syntaxiques

Dans le cadre de cette thèse, nous avons développé un visualiseur d'analyses syntaxiques. Ce logiciel, écrit en Java, est accessible sous forme de démonstration sur notre site internet

http://www.info.unicaen.fr/~giguet/syntaxique.html

(des captures d'écrans de ce visualiseur sont présentées figure D.19 page 208).

Le développement d'outils permettant de manipuler des corpus analysés est essentiel dans une démarche orientée corpus. Il faut pouvoir visualiser très rapidement de grosses quantités de données et avoir un accès pertinent à l'information contenues dans ces données. L'objectif du visualiseur est de répondre à ces nécessités. Ce logiciel offre plusieurs façons d'accéder à l'information :

Par défaut, la navigation dans la liste des analyses s'effectue en demandant l'accès à l'analyse précédente ou à l'analyse suivante. Ce type de navigation est intéressant car il permet un balayage rapide d'un fichier d'analyses et donne un aperçu global de la qualité des résultats. C'est ce que nous attendons d'un mode par défaut.

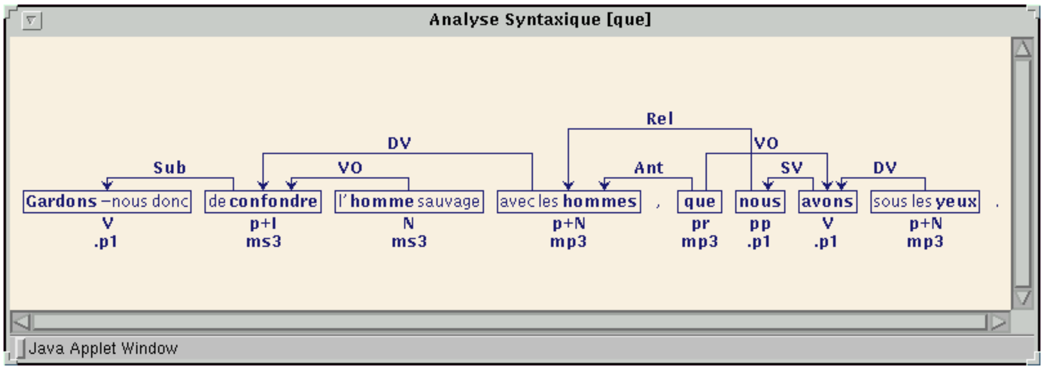
Le deuxième mode de navigation est l'accès direct à une analyse par sa position absolue dans le fichier des analyses. L'accès direct évite le parcours séquentiel du fichier et permet de se positionner sur une analyse jugée pertinente et dont la position a été notée pour pouvoir la retrouver rapidement par la suite.

Sur ce noyau, deux principales extensions ont été greffées :

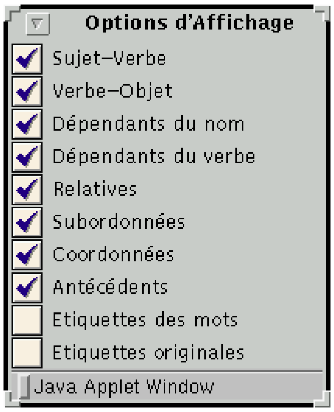
En tant que prestataire de recherche, Christèle Poirier a réalisé un module de comparaison de deux fichiers d'analyses syntaxiques. Ce module permet une visualisation graphique des différences d'analyse. La navigation se fait toujours phrase à phrase mais saute automatiquement les phrases ayant des analyses identiques. L'utilisation que nous en avons est double : il s'agit d'une part de comparer un fichier d'analyses avec un attendu, d'autre part de comparer deux fichiers d'analyses générés par deux versions successives de l'analyseur. La deuxième comparaison est actuellement la plus utilisée car il est important de vérifier à chaque modification de l'analyseur, les répercutions sur le corpus.

En stage de maîtrise, nous avons encadré Frédéric LECOQ et Frédéric SUTER pour le développement d'un module de recherche de configurations syntaxiques dans un fichier d'analyses. Ce module permet d'extraire d'un fichier toutes les analyses répondant à une requête exprimée en terme de relations syntaxiques. Le langage de requêtes qu'ils ont conçu permet d'interroger le graphe de relations syntaxiques dans tout son ensemble. Dans ce mode, la navigation se fait phrase à phrase parmi les phrases répondant aux critères de recherche.

Ce logiciel a fait l'objet d'une publication sous forme de démonstration référencée sous (GIGUET et VERGNE, 1997c).



(a) Fenêtre de visualisation des analyses



(b) Panneau de configuration de l'affichage

(c) Panneau de navigation dans un fichier d'analyses

http://www.info.unicaen.fr/~giguet/syntaxique.html

FIG. D.19 - *Le visualiseur d'analyses syntaxiques*

Bibliographie

Bibliographie

GRACE. Grammaires et Ressources pour les Analyseurs de Corpus et leur Évaluation. http://www.limsi.fr/TLP/grace, 1997. URL testée le 10 octobre 1998.

Steven ABNEY. Parsing By Chunks. Dans Robert BERWICK, Steven ABNEY, et Carol TENNY, editeurs, *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht, 1991.

Steven ABNEY. Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax. Dans Jennifer COLE, Georgia M. GREEN, et Jerry L. MORGAN, editeurs, *Computational Linguistics and the Foundations of Linguistic Theory*, pages 145–164. CSLI, 1995.

Steven ABNEY. Part-Of-Speech Tagging and Partial Parsing. Dans Ken CHURCH, Steve YOUNG, et Gerrit BLOOTHOOFT, editeurs, *An Elsnet Book*, Corpus-Based Methods in Language and Speech. Kluwer Academic, Dordrecht, 1996.

Salah AÏT-MOKHTAR et Jean-Pierre CHANOD. Incremental Finite-State Parsing. Dans *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pages 72–79, Washington, DC USA, avril 1997. Association for Computational Linguistics.

Leonard BLOOMFIELD. *Language*. Holt and Winston, 1933. Traduction française Payot 1970.

William B. CAVNAR et John M. TRENKLE. N-Gram-Based Text Categorization. Dans *Symposium On Document Analysis and Information Retrieval*, pages 161–176, Université du Névada, Las Vegas, USA, 1994.

Jean-Pierre Chanod et Pasi Tapanainen. Statistical and constraint-based taggers for French. Rapport technique MLTT-016, Rank Xerox Research Center, Grenoble Laboratory, novembre 1994.

Jean-Pierre Chanod et Pasi Tapanainen. Creating a Tagset, Lexicon and Guesser for a French Tagger. Dans *Proceedings of the European Chapter of the ACL SIGDAT Workshop "From text to tags : Issues in Multilingual Language Analysis"*, pages 58–64, Dublin, Irlande, mars 1995a.

Jean-Pierre Chanod et Pasi Tapanainen. Tagging French — comparing a statistical and a constraint based method. Dans *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, pages 149–156, Dublin, Irlande, mars 1995b. Association for Computational Linguistics.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge MA, USA, 1965.

Patrick Constant. *Analyse Syntaxique par Couches*. thèse de doctorat, École Nationale Supérieure des Télécommunications, avril 1991.

Michael A. Covington. A dependency Parser for Variable-Word-Order Languages. Rapport technique AI-1990-01, Artificial Intelligence Programs, The University of Georgia Athens, Georgia 30602 USA, janvier 1990.

Hervé Déjean. *Concepts et algorithmes pour la découverte des structures formelles des langues*. thèse de doctorat, Université de Caen, France, décembre 1998a.

Hervé Déjean. Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. Dans *Workshop on Paradigms and Grounding in Natural Language Learning*, pages 295–299, Adélaide, Australie, janvier 1998b.

Ted Dunning. Statistical Identification of Language. Rapport technique MCCS-94-273, Computing Research Lab, New Mexico State University, 1994.

Michel Dupont. Le calcul de la référence dans la compréhension automatique limitée de corpus homogènes. Dans *Actes du onzième colloque international du CerLICO*, pages 237–259, Caen, juin 1997. CerLICO: Cercle Linguistique du Centre et de l'Ouest.

David Elworthy. Tagset Design and Inflected Languages. Dans *Proceedings of the European Chapter of the ACL SIGDAT Workshop "From text to tags : Issues in Multilingual Language Analysis"*, pages 1–9, Dublin, Irlande, mars 1995.

Emmanuel Giguet. Categorization according to Language: A step toward combining Linguistic Knowledge and Statistic Learning. Dans *Proceedings of the International Workshop of Parsing Technologies (IWPT'95)"*, Prague - Karlovy Vary, République Tchèque, septembre 1995a.

Emmanuel Giguet. Multilingual Sentence Categorization according to Language. Dans *Proceedings of the European Chapter of the ACL SIGDAT Workshop "From text to tags : Issues in Multilingual Language Analysis"*, pages 73–76, Dublin, Irlande, mars 1995b.

Emmanuel Giguet. The Stakes of multilinguality: Multilingual text tokenization in Natural Language Diagnosis. Dans *Proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence Workshop "Future issues for Multilingual Text Processing"*, pages 9–13, Cairns, Australie, août 1996.

Emmanuel Giguet. Toward an Adequate Model for Automatic Syntactic Parsing. Dans *Poster Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI'97)*, Nagoya, Aichi, Japan, août 1997.

Emmanuel Giguet et Jacques Vergne. From Part of Speech Tagging to Memory-based Deep Syntactic Analysis. Dans *Proceedings of the International Workshop on Parsing Technologies (IWPT'97)*, pages 77–88, MIT, Cambridge, MA, USA, septembre 1997a.

Emmanuel Giguet et Jacques Vergne. Syntactic analysis of unrestricted French. Dans *Proceedings of the International Conference on Recent*

Advances in Natural Languages Processing (RANLP'97), pages 276–281, Tzigov Chark, Bulgarie, septembre 1997b.

Emmanuel GIGUET et Jacques VERGNE. Syntactic structures of sentences from large corpora. Dans *Demonstration Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 1–2, Washington, USA, avril 1997c. Association for Computational Linguistics.

Gregory GREFENSTETTE. Comparing Two Language Identification Schemes. Dans *Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT'95)*, Rome, Italie, décembre 1995.

Gregory GREFENSTETTE et Pasi TAPANAINEN. What is a word, What is a sentence? Problems of tokenization. Dans *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX'94)*, pages 79–87, Budapest, 1994. Research Institute for Linguistics Hungarian Academy of Sciences.

Maurice GREVISSE. *Précis de grammaire française*. J. DUCULOT S.A., Paris-Gembloux, 1969. 28e .

Blanche-Noëlle GRUNIG. Charges mémorielles et prédictions syntaxiques. *Les Cahiers de Grammaire*, (18):13–29, décembre 1993.

Blanche-Noëlle GRUNIG. Une conception dynamique du contexte. *La linguistique*, 31(2):5–13, 1995.

Claude HAGÈGE. *La structure des langues*. Que sais-je? PUF, 1982.

Zellig HARRIS. From phonemes to morphemes. *Language*, 31(2):190–222, 1955.

Fred KARLSSON. Constraint Grammar as a framework for parsing running text. Dans *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, pages 168–173, Helsinki, Finlande, 1990.

Fred KARLSSON, Atro VOUTILAINEN, Juha HEIKKILÄ, et Arto ANTTILA (eds.). *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, 1994.

Igor A. MEL'ČUK. *Dependency Syntax: theory and practice.* State University of New York Press, Albany, 1988.

David M. W. POWERS. Applications and explanations of Zipf's Law. Dans David M. W. POWERS, editeur, *Proceedings of New Methods in Language Processing and Computational Natural Language Learning (NEMLAP3/CoNLL98)*, pages 151–160, Macquarie University, janvier 1998.

Marie-Paule PÉRY-WOODLEY. Quels corpus pour quels traitements automatiques? *Traitement Automatique des Langues*, 36(1-2):213–232, 1995.

Frédérique SEGOND et Max COPPERMAN. Lexicon Filtering. Dans *Proceedings of the International Conference on Recent Advances in Natural Languages Processing (RANLP'97)*, pages 51–58, Tzigov Chark, Bulgarie, septembre 1997.

Penelope SIBUN et Jeffrey C. REYNAR. Language Identification: Examining the Issues. Dans *Proceedings of the Fifth Annual Conference on Document Analysis and Information Retrieval (SDAIR'96)*, Las Vegas, USA, avril 1996.

Daniel D. SLEATOR et Davy TEMPERLEY. Parsing English with a Link Grammar. Dans *Proceedings of the Third International Workshop on Parsing Technologies (IWPT'93)*, pages 277–292, Tilburg, Durbuy, août 1993.

Pasi TAPANAINEN et Timo JÄRVINEN. A non-projective dependency parser. Dans *Proceedings of the fifth Conference on Applied Natural Language Processing (ANLP'97)*, pages 64–71, Washington, DC USA, avril 1997. Association for Computational Linguistics.

Lucien TESNIÈRE. *Éléments de syntaxe structurale.* Klincksieck, Paris, 1959.

Jacques VERGNE. Syntactic properties of natural languages and application to automatic parsing. Dans *SEPLN'93 congress*, Grenade, Espagne, août 1993. Sociedad Española para el Procesamiento del Lenguaje Natural.

Jacques VERGNE. A non recursive sentence segmentation, applied to parsing of linear complexity in time. Dans *New Methods in Language Processing (NEMLAP'94)*, pages 234–241, juin 1994.

Jacques VERGNE. Esquisse d'une Syntaxe des Langues Concrètes. Dans *Les Cahiers du GREYC*, volume 11. GREYC, Université de Caen, France, 1995.

Jacques VERGNE. Entre arbre de dépendance et ordre linéaire, les deux processus de transformation : linéarisation, puis reconstruction de l'arbre. *Les Cahiers de Grammaire*, (23), décembre 1998.

Jacques VERGNE. Mémoire d'habilitation à diriger les recherches. Université de Caen, 1999. à paraître.

Jacques VERGNE et Emmanuel GIGUET. Regards Théoriques sur le «Tagging». Dans *Actes de la cinquième conférence annuelle «Le Traitement Automatique des Langues Naturelles» (TALN 1998)*, Paris, France, juin 1998.

Bernard VICTORRI. La construction dynamique du sens : un défi pour l'Intelligence Artificielle. Dans *Proceedings du 11ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'98)*, Clermont-Ferrand, janvier 1998. AFCET, AFIA, Univ. Blaise Pascal de Clermont-Ferrand. Conférence invitée.

Atro VOUTILAINEN. Designing a parsing grammar. Rapport technique 22, Departement of General Linguistics, Université d'Helsinki, Helsinki, Finlande, 1994.

Atro VOUTILAINEN, Juha HEIKKILÄ, et Arto ANTTILA. Constraint Grammar of English: A performance-oriented introduction. Rapport technique 21, Departement of General Linguistics, Université d'Helsinki, Helsinki, Finlande, 1992.

Stephan WERMTER. Integration of Semantic and Syntactic Constraints for Structural Noun Phrase Disambiguation. Dans *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, volume 2, pages 1486–1491, Detroit, Michigan USA, août 1989.

Yorick WILKS, Xiuming HUANG, et Dan FASS. Syntax, Preference and Right Attachment. Dans *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, volume 2, pages 779–784, Los Angeles, California USA, août 1985.

George Kingsley ZIPF. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley, 1949.

Méthode pour l'analyse automatique de structures formelles sur documents multilingues

Cette thèse traite de l'analyse automatique de structures formelles de l'écrit. Elle commence par une excursion dans le multilinguisme au cours de laquelle nous présentons les documents dans leur dimension multilingue et montrons la nécessité de les traiter comme tels. Nous étudions leur structure multilingue et développons comment la calculer à l'aide d'un identificateur de langues.

Nous poursuivons par l'exposé d'une méthode originale d'analyse syntaxique automatique d'énoncés français tout-venants. Cette méthode est issue de nos travaux de généralisation et d'abstraction des recherches de Jacques Vergne. Les structures syntaxiques auxquelles nous nous sommes particulièrement intéressé sont le syntagme minimal et la proposition ; deux unités auxquelles il est possible d'associer une définition ayant une validité multilingue, ce qui rend la méthode applicable à diverses langues.

Nous proposons deux processus permettant la construction de ces unités. Ces processus considèrent les énoncés comme des flux textuels et construisent chacun leurs structures syntaxiques par propagation de contraintes relationnelles. Les structures intra-syntagmatique et intra-propositionnelle étant dépendantes, elles sont construites par l'interaction des deux processus, le second processus acceptant de travailler sur des unités partiellement définies. Enfin, nous montrons que les deux processus sont identiques si l'on fait abstraction de la nature de l'unité qu'ils construisent et de la base de règles qu'ils manipulent.

Le fil conducteur de cette thèse est la méthode. À chaque calcul de structure, nous mettons en effet l'accent sur la méthode ayant permis son obtention. Nous montrons que cette méthode est unique. Chaque structure est en effet calculée à partir d'indices formels et positionnels à la fois internes et externes : internes par l'étude des unités qui composent la structure, externes par l'étude du rôle de cette structure dans l'unité qui l'intègre.

Mots-clés : langage naturel, traitement du (informatique)/analyse documentaire/analyse automatique (linguistique)/multilinguisme

Discipline : Informatique

GREYC CNRS UPRESA 6072
Groupe de Recherche en Informatique, Image, et Instrumentation de Caen
Université de Caen — Campus II — F14032 Caen Cedex

A Method for Automatic Parsing of Formal Structures in Multilingual Documents

This thesis deals with automatic parsing of formal structures in written texts. It begins with a presentation of documents in their multilingual dimension and of the necessity to process them in this way. We study their multilingual structure and present how to compute it with the help of a language identification tool.

Then, we present an original syntactic parsing method of unrestricted french sentences. This method is a generalization and an abstraction of Jacques Vergne's researches. The syntactic structures we are interested in are the minimal syntagm and the proposition ; both units can be defined as multilingual units so that the method can be applied to various languages.

We propose two processes which allow the building of these units. Both processes consider texts as flows and build syntactic structures thanks to a relational constraints propagation. As the syntagmatic and propositional structures are dependent, they are built up by the interaction of the two processes. We show that both processes are identical if we disregard the nature of the unit they build up and the rule base they use.

The main thread of this thesis is the method. Each time a process is described, we emphasize the related method. We show that this method is unique. Each structure is computed with the help of formal and positionnal clues : these clues come from the study of the units located inside the structure (internal clues) or from the study of the function of the structure in its upper-level units (external clues).

Keywords: natural language processing (computer science)/parsing (computer grammar)/subject cataloging/multilingualism

GREYC CNRS UPRESA 6072
Groupe de Recherche en Informatique, Image, et Instrumentation de Caen
Université de Caen — Campus II — F14032 Caen Cedex