Concepts et algorithmes pour la découverte des structures formelles des langues

Hervé Déjean

UFR Sciences **Université de Caen Basse Normandie**
École doctorale SIMEM

Concepts et algorithmes pour la découverte des structures formelles des langues

THÈSE

présentée et soutenue publiquement le 18 décembre 1998

pour l'obtention du

Doctorat de l'université de Caen

(spécialité informatique)

par

Hervé Déjean

Composition du jury

Président :	Daniel Kayser, professeur d'université	Université de Paris 13
Rapporteurs :	Pierre Lafon, directeur de recherche au CNRS	ENS Fontenay Saint-Cloud
	Fathi Debili, directeur de recherche au CNRS	CNRS-CELLMA IRMC
Examinateurs :	Didier Bourigault, chargé de recherche au CNRS	Université de Paris 13
	Khaldoun Zreik, professeur d'université (directeur)	Université de Caen
	Jacques Vergne, maître de conférences	Université de Caen

Mis en page avec la classe TheseCRIN.

à maman

Table des matières

Table des figures

Liste des tableaux

Introduction

Que peut-on apprendre sur la structure d'une langue à partir d'un texte écrit dans cette langue, et ceci sans connaissance particulière sur celle-ci et avec l'aide (disons l'utilisation) d'un ordinateur ? Voilà la question à laquelle nous allons essayer de répondre.

Le terme "apprendre" nous a d'abord conduit vers le monde de l'apprentissage en informatique (le "machine learning"), à la recherche de méthodes et algorithmes nous permettant de mener à bien ce travail. De par la nature des données manipulées, très différentes des données manipulées par ces méthodes, ces recherches nous ont semblé assez infructueuses.

Ce constat nous a alors conduit à nous tourner vers les données. Nous sommes entrés dans une phase d'observation de celles-ci, ce que nous appelons "partir des données". Nous avons constaté que, dans beaucoup de travaux en apprentissage, le travail portait sur les algorithmes, légitime en soi, mais que les données étaient souvent oubliées. Pour mettre au point des méthodes permettant de traiter efficacement des données, il nous semble qu'il faille les considérer comme premières et centrales dans le cas de notre problème. La principale activité de ce travail, très fructueuse, consiste à étudier un texte dans une langue que l'on ne parle pas (donc que l'on ne comprend pas) et à essayer de trouver les relations qu'il peut exister entre les séquences de mots, et une fois une relation trouvée, essayer d'expliquer le pourquoi de celle-ci. Cette activité a eu pour conséquence un changement de terminologie : nous ne parlions plus d'apprentissage mais de *découverte* (d'émergence) de structures. En fait ce travail est un exemple de ce que l'on peut appeler la "linguistique assistée par ordinateur".

Notre crainte, à un moment donné, a été de penser qu'une telle méthode conduise à une absence de formalisation dans les résultats, et qu'elle ne débouche que sur un ensemble de procédures ad hoc. Nous espérons avoir palié ce problème en mettant au point un formalisme de représentation de la structure des langues permettant une certaine prédiction des structures pouvant être rencontrées, ainsi qu'une identification des problèmes théoriques et la mise au point de mécanismes de résolutions de ceux-ci.

Quelles sont les connaissances linguistiques qui peuvent ainsi être découvertes ? Les différentes classes de mots, les notions d'accords, de structures prédicatives ? Les résultats obtenus sont, nous semble-t-il, très intéressants. Ce travail n'a pas découvert de nouvelles unités ou de nouveaux concepts : les notions de morphème, de syntagme simple, de proposition, ou de structures marquées à leur frontières sont connues depuis longtemps. *Mais ce travail présente une mé-*

thode de détection et de génération automatique de ces structures à partir d'un simple texte d'une langue donnée, sans connaissance sur cette langue. Ce travail met aussi en avant des propriétés structurelles des langues, assez générales et montre les limites, mais aussi les possibilités, d'un traitement se basant uniquement sur des critères formels. Notons que ce travail ne porte pas sur le problème de savoir quelle est l'information qui est transmise dans un texte, mais de savoir *comment* cette information est transmise. Nous pouvons trouver l'organisation, la structure utilisée dans telle ou telle phrase, mais jamais nous ne pouvons dire de quoi "parle" cette phrase (quelle information est transmise). Les résultats présentés ici ne concernent que le plan formel de la langue. Qu'entendons nous par la forme d'une langue et comment y accéder ? Comme nous l'avons déjà signalé précédemment, une méthode essentielle est de travailler sur des textes écrits dans des langues que nous ne parlons pas. Impossible donc d'accéder au sens de ces textes. Notre seule information accessible est une suite de symboles. Ce sont les propriétés de cette suite de symboles que nous appellerons les caractéristiques *formelles* de la langue, propriétés générales aux langues et qui permettront la construction de la structure de ces langues.

Dans cet ouvrage, lorsque nous utilisons le terme *la structure de(s) la langue(s)*, nous désignons la hiérarchie structurelle utilisée dans ce travail (figure 4.7). Le terme indéfini de *structure* désigne les différents niveaux de cette hiérarchie (morphème, syntagme, proposition, et couples de ces trois niveaux). Il faut toujours sous-entendre au terme structure, l'adjectif *formel*.

Cet ouvrage s'organise autour de quatre parties. La première partie de cette thèse décrit la problématique, définit ce que nous entendons par procédure de découverte et la méthodologie ainsi que les données que nous avons utilisées.

La deuxième partie concerne le travail au niveau morphologique : découverte des morphèmes, émergence des séquences morphologiques, finalement et surtout la limite de l'utilisation seule de ce critère.

La troisième partie introduit le concept sur lequel ce travail repose : l'idée que les structures formelles des langues peuvent être découvertes grâce à des marqueurs de frontières. Le début et la fin de telles structures sont indiqués par des éléments linguistiques (mots, morphèmes). Ces éléments permettent la construction d'une hiérarchie structurelle à trois niveaux : le morphème, élément de base et donc indécomposable sur le plan structurel, le syntagme simple et la proposition. La découverte de toutes ces structures est essentielle pour mener à bien ce travail.

Une fois les structures possibles identifiées, la quatrième partie explique la manière dont elles sont construites pour une langue donnée. À partir d'un simple texte, nous commençons par générer automatiquement la liste de certains marqueurs de frontières. Ces marqueurs servent alors de point de départ au processus de catégorisation des mots et morphèmes du texte. L'utilisation des structures décrites dans la troisième partie permet de réaliser la construction des contextes distributionnels servant à la catégorisation des mots et morphèmes.

Les parties une et deux peuvent se lire indépendamment. La lecture de la conclusion de la deuxième partie suffit comme pré-requis pour les parties suivantes. La lecture de la troisième partie est recommandée avant celle de la quatrième partie.

Dans cet ouvrage, les exemples portent sur plusieurs langues. *Ces exemples sont tous extraits des corpus décrits en annexe.*

Nous avons jugé que nos travaux étaient assez éloignés des travaux et des méthodes existants pour ne pas consacrer une partie entière à ceux-ci. Les références à ces travaux se trouvent incorporées à différents endroits du document.

Première partie

Liminaires

Introduction

Cette partie est composée d'un ensemble de remarques générales relatives à l'analyse distributionnelle et comprend aussi quelques points méthodologiques.

Nous allons d'abord présenter ce travail et ses objectifs initiaux. Nous présenterons ensuite la méthode distributionnelle et la notion de distribution, qui ont servi de cadre méthodologique dans ce travail, ainsi que les critiques méthodologiques ou pratiques adressées à cette méthode. Nous ferons aussi le parallèle entre notre travail et deux autres types de recherches : celles des universaux des langues et le travail réalisé par les déchiffreurs de langues et d'écritures. Nous verrons quelles différences existent entre ces types de travaux et le nôtre. Puis nous présenterons notre méthodologie de travail, ainsi qu'un descriptif des données utilisées. Ce point permettra de préciser l'importance d'un travail sur corpus et d'une approche multilingue, c'est-à-dire le travail sur plusieurs langues variées.

Les citations utilisées dans cette partie assez "polémique", pour illustrer les idées des auteurs, peuvent parfois simplifier celles-ci. Nous ne pouvons qu'encourager les lecteurs à une lecture plus approfondie des ouvrages cités.

Chapitre 1

Quelques points méthodologiques

Sommaire

1.1 L'objectif du travail

Ce travail est parti d'une question assez simple (peut-être naïve) : que peut-on apprendre sur une langue en étudiant un texte (corpus) de cette langue ? Question assez vague au premier abord. Quels étaient les objectifs à atteindre ou envisageables ? Nous ne le savions pas. L'analyse distributionnelle nous a fourni un premier cadre méthodologique dans ce travail, et nous avons repris un certain nombre de points méthodologiques de cette analyse. Cela a orienté très fortement la suite de nos recherches. Cette procédure (décrite dans la section suivante) travaille sur un texte ou un enregistrement sonore d'une langue donnée et essaie de découvrir la structure de cette langue, ceci sans utiliser le sens du

texte, en se basant uniquement sur des régularités formelles. Nous nous sommes alors placé dans ce cadre de travail, une étude portant sur la structure formelle de la langue. La question se reformulait donc ainsi : que peut-on apprendre de la *structure formelle* d'une langue en étudiant un corpus de cette langue. Nous insistons sur le fait que ce travail ne concerne en rien un travail d'analyse syntaxique. Pour bien comprendre le problème auquel nous nous confrontons, un simple exercice suffit : prenez un texte dans un langue donnée, de taille aussi grande qu'il vous plaira, et essayez de trouver quels sont les mots en relation les uns avec les autres. C'est ce type d'exercice que nous avons pratiqué pour mettre au point la méthode de découverte des structures. Cette notion de structure formelle était alors assez floue. Elle s'est affinée au fur et à mesure du travail. Nous avons essayé de partir avec le moins d'a priori possibles, mais comme toujours, ceux-ci sont loins d'être nuls. Au commencement, nous reprenions l'idée traditionnelle de deux types de structures : paradigmatiques et syntagmatiques. Autrement dit, il existe des catégories d'éléments et des relations entre celles-ci. De plus, la structure était vue comme étant hiérarchique, c'est-à-dire que les structures d'un niveau donné forment les éléments du niveau suivant (ou supérieur). Nous reviendrons sur cette notion plus en détail dans le chapitre 4. Nous avions donc deux objectifs : trouver ces catégories et ces relations. Les catégories mises à jour correspondent assez bien aux catégories traditionnelles. En fait, notre problème n'est pas seulement de trouver quelles sont les relations entre les éléments, mais de trouver les indices formels qui marquent ces relations. En effet, il ne suffit pas de savoir que dans telle langue, il existe une relation entre un substantif et un adjectif, mais de pouvoir déterminer quel élément est un substantif, quel autre est un adjectif et quelle est la marque (si elle existe) qui marque la relation entre ces deux éléments. C'est l'identification de ces marques qui nous permet de sélectionner ou non certaines structures. La question qui a guidé ce travail est donc : *quelles sont les marques formelles qui permettent d'établir une relation entre deux éléments et ainsi de définir une structure composée de ces deux éléments.* Une partie du travail a donc consisté à identifier ces marques (comme la notion de début et de fin que nous avons manipulée assez tôt dans ce travail), une autre partie a été de pouvoir les utiliser correctement. Il nous a fallu plusieurs mois avant de "comprendre" comment utiliser ces notions et à quoi elle correspondaient. En fait, il nous a fallu attendre la construction de la structure intégrant le niveau propositionnel pour pouvoir mettre au point des algorithmes de catégorisation vraiment efficaces.

Pourquoi vouloir entreprendre un tel travail et quel peut en être l'intérêt ? Alors qu'en intelligence artificielle, un courant de travail cherche à simuler informatiquement les différents processus humains, dans le but de modéliser ceux-ci, notre démarche est inverse : lorsque nous programmons une machine (ici un ordinateur) pour réaliser une tâche, essayons d'utiliser ses points forts en ayant conscience de ses points faibles et de ses limites, sans chercher à les dépasser mais seulement à les identifier. Nous ne disons pas que la simulation informatique des processus humains est une mauvaise voie, au contraire, mais que ce n'est pas celle qui a été choisie pour ce travail. Pourquoi vouloir traiter la langue (un texte) par des moyens formels, c'est-à-dire qui ne prennent pas en compte le sens d'un énoncé mais les propriétés de sa construction physique ? *Parce que ces*

propriétés sont facilement accessibles et utilisables d'un point de vue informatique, puisqu'elles sont contenues dans les données fournies et qu'elles peuvent en être extraites. L'intérêt de ce travail de découverte est donc de rechercher dans les données des marques, des particularités formelles qui nous donnent des indications sur les structures, non pas que l'on veut construire, mais que l'on *peut* construire (qui, en pratique, se recoupent). En se mettant dans la "boite" de la machine (en travaillant sur des langues que nous ne comprenons pas, ce qui permet un réel travail formel sans recours au sens), nous pouvons recenser les opérations facilement réalisables en utilisant les ressources formelles des langues, et donc mettre au point des processus assez simples et ne demandant pas de grandes ressources. *Ainsi, il est par exemple plus facile de segmenter une séquence en propositions que de mettre en relation certains syntagmes de ces propositions.* De plus, la segmentation en propositions peut se révéler indispensable à la mise en relation de certaines syntagmes. Nous voyons donc que la difficulté d'une tâche n'est pas en relation avec le niveau hiérarchique des éléments qui la composent : à chaque niveau (morphémique, syntagmatique, propositionnel,...), certaines opérations sont facilement réalisables avec des ressources formelles, et d'autres très délicates, voire impossibles avec ces mêmes ressources.

1.2 Le distributionnalisme

You shall know a word by the compagny it keeps. [Firth, 1957]

Que peut-on apprendre sur une langue (ou plus exactement sur sa structure) à partir de l'étude d'un texte écrit dans cette langue ? Une première réponse nous a été fournie par les travaux de l'école dite distributionaliste américaine dont la figure emblématique était Zellig S. Harris. Cette école doit son nom à l'utilisation de la notion de *distribution*, expliquée à la section 1.3. [Harris, 1951] présente l'ensemble des "méthodes de recherche utilisées en linguistique descriptive ou, plus exactement, structurale" [Harris, 1951, page 1]. Schématiquement la méthode consiste à construire un échantillon d'une langue, appelé *corpus*, et à étudier les régularités de ce corpus, afin de décrire la structure de cette langue. L'étude des régularités se base sur la notion de *distribution*. La distribution d'un élément (phonème, morphème, séquence de morphèmes) est la somme des environnements de cet élément. Ce seul critère est utilisé pour catégoriser les éléments. Le sens n'intervient pas dans la démarche. La recherche de régularité se fait en segmentant les séquences du corpus pour mettre à jour des régularités entre les éléments ainsi segmentés. Les différentes procédures proposées par Harris seront décrites dans le chapitre 2 et la section 3.3. Elles ont fourni un excellent point de départ à notre travail. L'expérimentant et arrivant aux limites de celles-ci, il nous a fallu introduire d'autres notions et d'autres procédures afin d'aller un peu plus loin dans ce travail. Une des grandes difficultés de cette méthode est de s'être trop intéressée aux petites unités de la structure (phonèmes et morphèmes), faute que Halliday considérera comme le quatrième péché de la méthode "bloomfieldienne" [Halliday, 1961, page 280]. De ces travaux, nous avons retenu trois points importants :

– l'utilisation de corpus

– la notion de distribution
– l'utilisation de la forme seule, sans recours au sens

Nous verrons, dans la section suivante, les principales critiques qui ont été adressées à cette méthode.

On trouve dans [Harris, 1954] une présentation générale de la méthode distributionnelle, et dans [Harris, 1951] un exposé très détaillé des procédures utilisées. La lecture de l'introduction de [Harris, 1951] resitue bien quel est l'intérêt d'un tel travail pour Harris, qui est beaucoup plus méthodologique que pratique. Un de ses intérêts (partagé par quelques autres comme [Pitman, 1948]) était de fournir aux linguistes des outils afin de systématiser le travail réalisé, et ainsi de permettre une meilleure comparaison entre les différents résultats obtenus. Le travail de Harris est à considérer sur le plan méthodologique beaucoup plus que sur le plan opérationnel. D'ailleurs n'écrit-il pas dans cette introduction :

> The particular methods described in this book are not essential. They are offered as general procedures of distributional analysis applicable to linguistic material [Harris, 1951, page 6].

Si l'on en croit [Nevin, 1993], Harris n'a jamais prétendu que la méthode qu'il propose permettait de générer une grammaire[1] à partir de textes. La lecture de l'introduction nous conduit aussi à cette analyse ainsi que la lecture de son dernier ouvrage [Harris, 1990]. Tout au long de notre travail, il nous semble avoir suivi la philosophie harrisienne, et les résultats obtenus nous semblent valider celle-ci.

1.3 La notion de distribution et ses problèmes

La méthode distributionnelle repose sur une notion centrale : la *distribution* d'un élément. L'observation de Harris sur la distribution des éléments est simple :

> Les parties d'une langue n'apparaissent pas arbitrairement relativement les unes aux autres ; chaque élément se rencontre dans certaines positions par rapport aux autres.[Harris, 1954]

De cette notion de distribution découle tout le processus de découverte des structures. Voici la définition que Harris en donne :

> la *distribution* d'un élément sera définie comme la somme de tous les environnements de cet élément. L'environnement d'un élément A est la disposition effective de ces *co-occurrents*, c'est-à-dire des autres éléments, chacun dans une position déterminée, avec lesquels figure A pour produire un énoncé.[Harris, 1954, page 13]

Ce critère est utilisé pour catégoriser les éléments d'un corpus. Deux éléments ayant une même distribution (le critère de *similarité*) sont considérés comme appartenant à une même classe dite distributionnelle (*regroupement par similarité*).

Nous allons voir que cette notion de distribution, si elle est centrale comme le montre notre travail, est néanmoins problématique. Quiconque commence à

[1] Par grammaire, nous entendons description des structures.

vouloir effectuer une analyse distributionnelle doit apporter une réponse aux questions suivantes : comment construire les contextes distributionnels et sélectionner les bons contextes, et comment classer les mots ?

Comment construire les contextes ? Le premier problème rencontré est celui de la définition du contexte. Nous avons vu que les mots sont regroupés par classes distributionnelles, c'est-à-dire que les mots partageant une même distribution sont regroupés dans une même classe. Quelle est la distribution d'un mot ? Les phrases dans lesquelles il apparaît ? Dans ce cas, aucun mot n'a de distribution semblable et aucun regroupement ne peut se faire. Il faut donc réduire la taille de la distribution. Celle utilisée habituellement dans les algorithmes de catégorisation est de quelques mots avant et/ou après. Les essais (voir annexe D et section 3.3.1) montrent que la catégorisation obtenue ne varie que très peu en fonction du nombre de mots. Le tableau 1.1 montre que cette approche n'est pas adéquate puisque parfois le contexte gauche est à utiliser, parfois le contexte droit est préférable, ceci pour une même langue (ce tableau est sans valeur si l'on considère que les mots *la* et *sa*, ainsi que *dans* et *avec* n'appartiennent pas à une même classe).

Mot précédent	Mot	Mot suivant
de, à, dans, et, sur,	**la**	première, commission, fin, France, vie
de, dans, à, et, pour,	**sa**	part, vie, mort, mère, femme
et, que, c'est, pas, notamment, place	**dans**	le, les, la, un, une
relations, alliance, contact, désaccord, coopération	**avec**	le, la, les, un, une

TAB. 1.1 – Contextes gauche et droite. Les mots *la* et *sa*. Alors que le contexte gauche est quasiment identique (4 mots sur 5), le contexte droit est totalement différent. L'inverse se produit pour les mots *dans* et *avec*.

Le fait d'augmenter la taille de la distribution n'est pas suffisant, la validité d'un contexte ne dépendant pas de sa taille. Dans notre corpus *français01*, le triplet de mots le plus fréquent est *il y a*. Mais le mot suivant peut appartenir à de nombreuses catégories (préposition, déterminant, verbe, adverbe, substantif, pronom). On peut penser que prendre un contexte gauche et droit renforce les contraintes et permet ainsi d'obtenir une catégorisation correcte, mais il n'en est rien. Les trois environnements (les mots en gras du tableau 1.2) permettent la catégorisation des éléments *la*, *leur*, *sa* et *notre*. Le tableau 1.3 illustre le cas d'un mauvais contexte constitué aussi d'un mot précédent et d'un mot suivant.

Comment savoir que le premier contexte est correct et le second ne l'est pas, étant donné que le critère de validation ne peut faire intervenir que des connaissances formelles ? La réponse à cette question, centrale à la méthode, ne peut être donnée qu'en ayant une connaissance de la structure formelle de la langue et non en augmentant aveuglément la taille des contextes. Harris traite de ce problème avec ce qu'il nomme la notion de *domaine* :

de	*la*	**fédération**
de	*leur*	**fédération**
de	*notre*	**fédération**
de	*sa*	**fédération**

TAB. 1.2 – Contexte distributionnel "correct".

de	*l'est*	**pas**
de	*même*	**pas**
de	*ne*	**pas**
de	*victor*	**pas**

TAB. 1.3 – Contexte distributionnel "incorrect".

> Toutes les règles sur la dépendance et la substituabilité s'appliquent à l'intérieur d'un domaine défini, ce domaine étant déterminé soit par sa nature (ainsi le "silence" avant ou après un énoncé), soit par les types d'environnements à l'intérieur desquels il y a une régularité (par exemple l'étroite restriction distributionnelle de *hood* concerne seulement ce qui le précède et, dans cette direction, seulement le premier morphème). [...] Le mot, le syntagme et la proposition sont des types courants de domaines. [Harris, 1954, page 31]

Si nous partageons ce point de vue, le problème reste entier : comment définir distributionnellement ces domaines ? Comment trouver que tel ou tel contexte correspond à un syntagme ou une proposition ? Nous apportons une réponse à ce problème au chapitre 4. Tant qu'une définition précise (et opératoire) du contexte n'est pas donnée, il est inutile de continuer un tel travail.

Comment classer les mots ? Le deuxième écueil de la méthode concerne la variété de contextes dans lesquels un mot peut apparaître. Si nous reprenons le tableau 1.2, nous voyons que les mots *la* et *notre* apparaissent dans le contexte [*de X fédération*]. Ils sont donc regroupés dans la même catégorie grâce à ce contexte. Mais ces deux mots ne partagent pas tous les contextes dans lesquels ils apparaissent, et donc, n'ont pas exactement la même distribution. Le problème est contourné en regroupant les mots qui partagent un contexte "assez" proche. La difficulté consiste alors à définir la distance de ressemblance entre deux mots. Certains mots se ressemblent plus que d'autres, ce qui produit une hiérarchie dans les classes obtenues. Ces points sont développés à la section 3.3.1.

Ce problème ne se pose que si nous raisonnons au niveau des mots. Les contextes que nous avons mis au point (chapitre 4), ne font pas appel aux mots mais à des concepts formels tels que des marqueurs de frontière. En fait, notre classification ne consiste pas à recenser les contextes dans lesquels un mot apparaît et à le regrouper avec les autres mots apparaissant dans un contexte ressemblant (ce qui est traditionnellement fait), mais à construire un contexte pour chaque classe distributionnelle, et ainsi de considérer qu'un mot apparais-

sant dans tel contexte appartient à telle classe. Il faut donc inverser le point de vue, et travailler avec les contextes, ce qui n'est habituellement pas fait, puisque ces contextes nécessitent une théorie formelle de la langue. *Le travail central est bien la construction des contextes distributionnels.*

1.4 Les critiques de principe adressées à la méthode distributionnelle

Plusieurs sortes de critiques ont été adressées à cette méthode. Certaines d'ordre méthodologique, comme celles de Noam Chomsky, d'autres d'ordre pratique.

Les critiques de Noam Chomsky Le linguiste Noam Chomsky, élève de Harris, a très fortement contesté l'intérêt d'un tel travail. Il condamne assez fortement le travail basé sur la notion de procédure de découverte et sur l'étude de corpus. Sur ce premier point il écrit :

> Nous pensons qu'il est déraisonnable d'attendre d'une théorie linguistique qu'elle fournisse plus qu'une procédure pratique d'évaluation des grammaires.[...] Autrement dit, elles [les propositions] essaient de formuler des méthodes d'analyses dont un chercheur pourrait réellement se servir, *s'il en avait le temps*[2], pour construire une grammaire d'une langue directement à partir des données brutes. Il me parait douteux que cet objectif puisse être atteint d'une manière intéressante, et je crains que toute tentative de cet ordre ne conduise à un dédale de procédures analytiques de plus en plus complexes et raffinées, qui laisseront sans solution beaucoup de problèmes importants concernant la nature de la structure linguistique.[Chomsky, 1969b, page 60]

> Les allusions à des "procédures de découvertes" ou "méthodes objectives" présumées bien connues ne font que masquer les conditions effectives où le travail linguistique doit se poursuivre pour le moment.[Chomsky, 1965, pages 35 et 36]

S'il est vrai qu'une génération automatique de grammaire à partir d'un corpus semble un défi assez difficile, les résultats obtenus en essayant de le relever peuvent être très intéressants. Quant au "dédale de procédures analytiques de plus en plus complexes et raffinées", cela est vrai et il nous semble difficile d'y échapper. D'ailleurs le travail de Chomsky semble illustrer parfaitement son propre propos.

Pour Chomsky, le travail à partir d'un corpus ne peut servir de base à un travail linguistique. Il base sa méthode de travail en interrogeant le locuteur sur sa langue et en faisant confiance à son intuition linguistique.

> Il y a, tout d'abord, la question de la manière dont on peut obtenir des informations sur la compétence du locuteur-auditeur, sur sa

[2]Mis en valeur par nous.

> connaissance de la langue. Comme la plupart des faits intéressants et importants, celui-ci [celle-ci ?] n'est pas accessible à l'observation directe et ne saurait être extrait des données par des procédures inductives d'aucune espèce bien connue.[...] En bref, il se trouve malheureusement qu'on ne connaît aucune technique formalisable adéquate pour obtenir une information solide touchant les faits de la structure linguistique (et cela n'a rien de spécialement surprenant) [Chomsky, 1965, page 36]

Pour resituer ces propos dans leurs contextes, nous devons insister sur le fait que l'objet d'étude de Chomsky (et selon lui de la linguistique) semble être principalement la *compétence* du locuteur-auditeur, c'est-à-dire la connaissance que ce dernier a de sa langue. Mais cette pratique peut être elle aussi critiquée.

> L'exigence de la référence à un corpus défini est donc d'abord une exigence de rigueur élémentaire, car on risque toujours de penser décrire une langue alors qu'on ne décrit que son propre usage, voire le sentiment qu'on en a. [François, 1968, p. 176]
>
> Il faut noter qu'en général, il n'y a pas d'opposition entre la description d'un corpus et le recours aux questionnaires ou interrogatoires, dans la mesure où ceux-ci ne se fondent pas sur l'hypothèse trompeuse selon laquelle les sujets seraient parfaitement conscients de la langue qu'ils parlent. [François, 1968, p. 176]
>
> On ne doit pas en conclure qu'il y a une différence de nature entre l'étude d'un corpus et l'étude de la langue.[François, 1968, p. 177]

Nous nous sommes aperçu, durant notre travail, qu'il y avait un phénomène que le locuteur maîtrise très mal : la fréquence des éléments et des structures dans la langue. C'est pourtant une caractéristique essentielle et une aide précieuse lorsque l'on travaille sur corpus, même si elle est à manipuler avec précaution (section 1.6). Quant au problème de la finitude du corpus, le recours au locuteur ne résout pas la question, puisqu'il n'a accès lui aussi qu'à une partie des structures existantes. Le problème de la représentativité du corpus et de sa constitution est discuté à la section 1.10. En fait, ce débat semble maintenant quelque peu dépassé aux vues des résultats fournis par le travail sur corpus [Habert et al., 1997].

Le problème du sens La deuxième critique concerne le "rejet" du sens dans cette méthode. Une des caractéristiques de la méthode distributionnelle est de remplacer l'utilisation du sens par la notion de distribution. Le sens des éléments n'intervient donc pas. Ce point suit la remarque de Leonard Bloomfield :

> La description du signifié est [...] le point faible de l'étude du langage [Bloomfield, 1933, page 140].

Mais la "condamnation" du sens chez Harris est beaucoup moins forte [Harris, 1954, page 26]. Il nous semble clair que le rejet du sens dans tous les domaines de la linguistique est absurde. Le problème est de bien définir le champ d'étude des travaux, ce que fait Harris : son objectif est de proposer des méthodes en linguistique descriptive, et pour lui la linguistique descriptive "ne concerne pas

l'ensemble des activités de la parole, mais les régularités dans certaines caractéristiques de la parole" [Harris, 1951, page 5]. Il nous semble que cette approche offre une méthodologie très intéressante en ce qui concerne les travaux sur la structure formelle des langues. Dans une perspective opératoire en traitement automatique des langues (maintenant TAL), il nous semble aussi important de voir quelles sont les limites théoriques des travaux se basant sur de simples ressources formelles, et d'un autre côté, quels sont les problèmes que de telles ressources peuvent résoudre (chapitre 7).

L'impossibilité pratique de la méthode La troisième sorte de critique est d'ordre pratique : il nous suffit de citer [Mahmoudian, 1981] :

> On constate qu'une analyse distributionnelle au sens strict du terme n'a jamais été effectuée, pour une langue. Les applications que l'on connaît sont des descriptions où, guidé par l'*intuition sémantique*[3], le linguiste opère des segmentations et des classements ; mais les arguments qu'il avance en faveur de ces opérations sont de nature distributionnelle. Or les phénomènes distributionnels sont nombreux d'une part, et d'autre part ils ne sont pas tous pris en compte de façon systématique. Il s'en suit que dans l'ensemble des faits de distribution, il y en a qui étaieraient une description, mais on en trouve aussi qui iraient à l'encontre de cette même description.
>
> L'analyse distributionnelle dans l'acception stricte du terme (c'est-à-dire sans critère sémantique) est une utopie. [Mahmoudian, 1981, page 149]

La critique est simple mais pertinente. La réponse aussi. Devant la complexité de la tâche qui peut s'étonner de ce résultat ? Et personne ne contredit ces remarques, même Harris y souscrit : l'introduction de [Harris, 1951, page 1] va dans ce sens :

> These procedures also do not constitute a necessary laboratory schedule in the sense that each procedure should be completed before the next is entered upon. In practice, linguists take unnumbered short cuts and intuitive or heuristic guesses, and keep many problems about a particular language before them at the same time [...]

Nous verrons qu'en axant la procédure sur les structures, la prise en compte systématique des faits peut être réalisée sans aucune contradiction. Il est vrai qu'un des problèmes de cette méthode a été de savoir trier les "bonnes" régularités des "mauvaises". Cela a été fait, et en fait, a été assez facile à réaliser (section 1.6).

Mais même si une automatisation totale de l'analyse distributionnelle est utopique (et nous pensons qu'elle ne l'est peut être pas autant que cela, même si nous ne l'avons pas réalisée), les essais, le travail tendant vers cette automatisation ne peut être que bénéfique à la connaissance que l'on a des langues et

[3] Mis en valeur par nous.

de leurs structures. Nous prendrons en exemple, la recherche de la pierre philosophale. Bien qu'aucun alchimiste ne soit parvenu (à notre connaissance !) à la réalisation d'une telle pierre, les travaux de ces chercheurs ont énormément fait progresser les connaissances en chimie. De plus, il suffit d'avoir conscience (ou plus exactement de découvrir) les limites d'un travail se basant sur des faits formels, et de ne pas attendre plus que l'on ne peut espérer. Si les résulats décrits au chapitre 6 posent les limites d'un tel travail, ils montrent aussi ses possiblités.

Est-ce la place importance qu'occupe Noam Chomsky dans cette partie de siècle, ou une réelle pertinence de ces critiques, toujours est-il, que peu nombreux ont été les chercheurs poursuivant les traces de Harris. Cela est d'autant plus regrettable que l'évolution technique de ces trente dernières années, offre de nouveaux outils (les ordinateurs) et de nouveaux types de données (les textes électroniques) totalement adaptés à ce genre de travail. Il est vrai que le manque de formalisme de la méthode présentée par Harris, rend celle-ci inopérante dans l'état où Harris l'a présentée.

1.5 L'historique du travail

Le seul travail que l'on puisse commencer par en haut, c'est creuser un trou.
(anonyme).

Cette section résume l'historique des travaux. Les objectifs de ce travail n'étaient pas fixés très clairement au départ. Nous étions à la recherche d'une structure. Mais laquelle ? Les travaux de Harris nous ont fourni un premier élément de celle-ci : le morphème. Il s'est avéré que cet élément est l'élément minimal de notre structure. Nous retrouvons cette idée chez de nombreux auteurs [Hockett, 1961], [Harris, 1951], [Halliday, 1961]. Notre première tentative, une fois les morphèmes segmentés (chapitre 2), a été de travailler sur les séquences de morphèmes, et d'essayer de trouver les relations entre elles. Suivant le principe de Harris, nous avons donc entrepris une "montée" de la structure. Nous sommes arrivés au niveau du syntagme. Là, nous nous sommes rendu compte que ce niveau permettait une certaine correction de la segmentation (c'est-à-dire du niveau inférieur : celui des morphèmes) : la maîtrise d'un niveau permet une meilleure compréhension et une meilleure analyse des niveaux inférieurs. D'où l'idée de partir des niveaux supérieurs afin de découvrir toute la hiérarchie. Le problème était que nous ignorions alors quelle était cette structure supérieure. Nous avons pris la phrase et l'"entre-ponctuations" et essayé de descendre dans la hiérarchie de la structure, mais sans succès (section 2.5), ce qui corrobore bien les propos suivants de Harris :

> The procedure outlined [l'analyse en CI] here could be paralleled by a series of substitutions beginning with the whole utterance and working down instead of beginning with simple morphemes and working up. In that case we would have to find formal criteria for breaking the utterance down at successive stages. This is essentially the difficult problem of determining the immediate constituents of an utterance. It is not clear that there exists any general method for successively

> determining immediate constituents, when we begin with the whole utterance and work down. In any case, it would appear that the formation of substitution classes presents fewer theoretical difficulties if we begin with morphemes and work up [Harris, 1946, page 178-179].

Le moyen le plus efficace est bien de partir de l'unité de base : le morphème, puis de gravir les échelons. Le niveau supérieur au morphème est celui du syntagme, construit avec un séquence de morphèmes. Ce niveau a été trouvé grâce à des marqueurs de début et de fin[4]. En appliquant le même principe, (considérer une séquence de syntagmes), nous avions espéré trouver le niveau supérieur au syntagme. Mais là, aucune structure n'apparaissait. Certaines séquences morphologiques étaient faciles à générer (chapitre 3). Ce qui nous préoccupait le plus était que la construction même des syntagmes pouvait se révéler assez difficile pour certaines langues (comme l'allemand). Avions nous déjà atteint la limite de la méthode ? Les informations formelles étaient-elles insuffisantes pour aller plus loin ? La structure de la langue etait alors vue comme une séquence de syntagmes, et chaque syntagme pouvait être marqué par un élément de début ou de fin (figure 1.1).

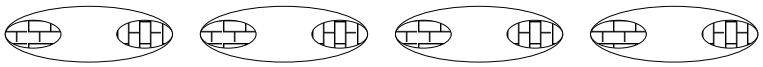


FIG. 1.1 – La première structure de la langue : une séquence d'éléments marqués à leur début et/ou leur fin.

Le problème s'est résolu lorsque nous avons intégré à notre structure le niveau supplémentaire classique : la proposition. Nous nous sommes aperçu que le niveau supérieur au syntagme, la proposition, était accessible directement à partir du niveau morphologique, et qu'il ne fallait pas le constuire à partir du niveau syntagmatique *mais en même temps*. Qui plus est, la connaissance du niveau propositionel est nécessaire à la construction du niveau syntagmatique (section 6.4).

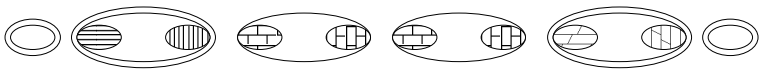


FIG. 1.2 – La deuxième structure de la langue : la proposition, marquée elle aussi par des débuts et des fins est composée d'une séquence de syntagmes.

Nous voyons là une différence entre le processus de découverte des niveaux de la structure, et le processus de construction des niveaux pour une langue donnée (figure 1.3). Le premier est un travail de bas en haut (morphème vers syntagme et proposition), mais le second travail part du niveau le plus haut (la proposition) pour construire le niveau inférieur (le syntagme). Le niveau morphémique étant le niveau de base, il est nécessaire de l'acquérir dès le début.

[4]Ces notions sont expliquées dans le chapitre 4.

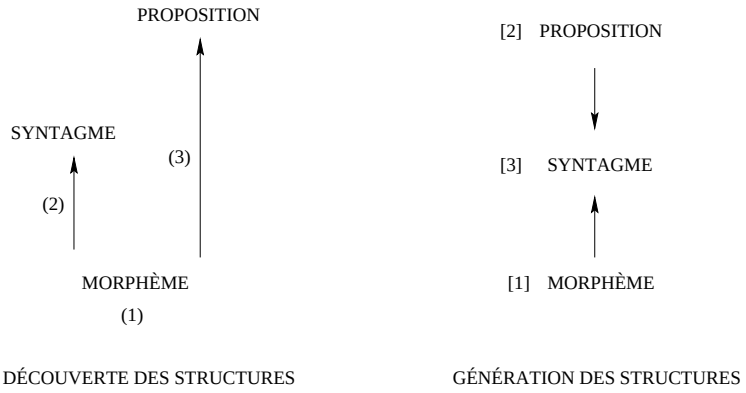


FIG. 1.3 – Chronologie dans la découverte des structures et dans leur génération. La découverte des structures s'est faite en montant dans la hiérarchie. La génération des structures pour une langue donnée se fait en partant du niveau propositionnel.

Comme nous le verrons, sa construction peut se faire, pour l'essentiel, sans recours aux niveaux supérieurs, même si ceux-ci peuvent, par la suite, corriger certaines erreurs. En fait, il nous semble important de noter que la structure mise à jour est assez simple : elle comporte peu de niveaux (morphème, syntagme, proposition), et chaque niveau possède des marqueurs spécifiques. Le travail de découverte n'est donc pas si compliqué qu'il y paraît. Nous reviendrons en détail dans le chapitre 4 sur la notion de structure et de niveau.

1.6 La recherche des régularités

La base d'un apprentissage non supervisé (voir section 1.7) est la recherche de régularité dans les données. Harris l'écrit aussi :

> Le premier fait distributionnel est la possibilité de diviser (de segmenter) toute chaîne parlée en parties, de façon à découvrir certaines régularités d'occurrence de l'une des parties, relativement à d'autres parties de la chaîne parlée.[Harris, 1954, pages 28-29]

Mais quelles sont donc ces "régularités d'occurrences" ? La recherche de ces régularités consiste seulement à remarquer certaines propriétés formelles des contextes dans lesquels ils apparaissent. Elles vont concerner les divers éléments que nous manipulons (mots, morphèmes, syntagmes,...). Tous les comportements ne sont pas pris en compte : ils sont trop nombreux. Ceux retenus devront se retrouver dans toutes les langues étudiées (ou pour le moins, dans une grande partie). Ils seront uniquement formels et seront interprétés à partir d'un modèle théorique.

1.6.1 À la recherche des universaux ?

Lorsque l'on travaille sur un corpus dans une langue donnée, de nombreuses régularités apparaissent. Elles sont souvent spécifiques à une langue donnée. La

recherche de ces régularités dans d'autres langues ne fournit généralement aucun résultat positif. Par exemple, il existe des langues dans lesquelles certains articles définis sont construis selon une structure consonantique donnée, comme l'allemand (*der*, *die*, *das*,*den*, *dem*, *des*), le français, (*le*, *la*, *les*, *leur*), l'anglais (*the*, *this*, *that*, *those*, *these*), etc...On retrouve aussi cette régularité au niveau des pronoms relatifs. Dans les langues étudiées, cette particularité n'existe absolument pas pour une classe comme les prépositions (qui proviennent parfois d'anciens mots lexicaux). Une telle régularité ne peut se trouver qu'après avoir effectué une catégorisation des éléments, comme critère de valisation par exemple (dans une certaine mesure). Elle ne peut absolument pas servir de critère de catégorisation, les coïncidences étant la règle générale. Ainsi la ressemblance entre les mots espagnols suivants *da*, *dan*, *dad*, *dar* ne se base sur aucune régularité structurelle (ou le mot anglais *they* ne fait pas partie de la liste donnée).

Il existe une multitude d'autres spécificités (section 1.8) liées à une ou plusieurs langues, ou plus exactement au système d'écriture utilisé. Dans le cadre de ce travail, nous allons essayer de ne déceler uniquement que les régularités "multilingues". Telle ou telle particularité à une langue donnée ne sera donc pas pris en compte dans la mise au point de la méthode générale. Les propriétés générales (universelles ?) se basent sur une conception simple de l'objet : une séquence linéaire d'unités. Ces unités sont marquées par des indicateurs de frontière. Nous avons retrouvé ce schéma dans toutes les langues étudiées. Nous pouvons donc considérer qu'il est une constante dans les langues, un universel. Cette réflexion nous a conduit à nous intéresser aux universaux de la langue. [Greenberg, 1963] nous donne une liste de 48 universaux *structurels* ou plus exactement 48 propositions que l'on retrouve dans 30 langues des cinq continents. En voici quelques exemples :

1 In declarative sentences with a nominal subject and object, the dominant order is almost always one in which the subject precedes the object.

2 In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it always precedes.

3 Languages with dominant VSO order are always prepositional.

4 If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.

5 Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number.

7 All languages have pronominal categories involving at least three persons and two numbers.

Si ces observations sont bien corroborées par notre expérience, il n'en reste pas moins qu'elles sont inutilisables dans un processus de découverte, au moins au début de celui-ci : *connaître l'existence d'une structure ne résoud pas le problème de l'identification de celle-ci.* Mais ces propositions peuvent être utiles, dans un deuxième temps, pour deux raisons. Premièrement, elles ne contredisent pas notre travail. Elles concernent assez souvent l'ordre des éléments dans une

séquence et des question d'accord (morphologiques), deux notions que nous utilisons dans notre méthode. Deuxièmement, ces propositions peuvent être utilisées pour affiner notre catégorisation. Par exemple pouvoir identifier le sujet de l'objet grâce à la proposition 1, ou des marques d'accords grâce aux propositions 4 ou 5. Nous ne nous sommes pas livrés à ce type de travail. Nos "universaux" sont beaucoup plus généraux que ceux de Greenberg, puisqu'ils ne concernent que les indications qui peuvent permettre une découverte des structures. Ils sont donnés au chapitre 4.

1.6.2 Les critères formels

Nous allons maintenant expliquer comment notre recherche de régularités s'est effectuée. Qu'entendons nous par l'adjectif formel : le critere formel est un critère qui ne prend en compte que des propriétés de la chaîne de symboles qui composent les corpus. Nous opposons donc un critère formel à un critère sémantique (qui utilise la compréhension). Le premier élément que nous avons pris en compte est l'effectif des éléments dans le corpus. Le deuxième concerne la longueur des séquences observées. Ces deux critères ont une particularité très importante : ce sont des *critères visuels*. Ce sont les deux critères que l'on utilise immédiatement lors d'une étude manuelle. Lorsque l'on travaille sur un texte, les premières régularités (les premières "choses" que l'on remarque) sont ces éléments fréquents ou qui apparaissent très souvent avec un autre élément qui n'est pas très loin *visuellement*. Le troisième critère est relatif à la position d'un élément dans une séquence. C'est un critère qui est beaucoup moins immédiat que les deux autres. Et pourtant il est *primordial*.

L'effectif La première opération à effectuer est un recensement de la liste des éléments (mots, morphèmes, syntagmes,...) que l'on manipule, ainsi que leur effectif[5] dans le corpus. Nous verrons dans la section 1.10.3 quelles sont les propriétés "fréquencielles" d'un texte écrit dans une langue. La première idée était de travailler avec les éléments fréquents du corpus. Ce sont ces éléments sur lesquels on possède le plus d'informations. Nous sommes partis de l'idée intuitive que l'effectif d'une séquence de mots était une indication de la mise en relation de ces mots. Ces informations peuvent se révéler d'un côté très utiles ([Kiss, 1972] utilise ces bigrammes pour catégoriser une trentaine de mots), de l'autre inexploitables. Si l'on peut dire que l'effectif entre éléments est une indication d'une relation entre ces éléments, cette indication est à considérer avec précaution. Elle n'indique pas une relation spécifique, mais correspond à toutes les relations de la structure. Prenons les exemples du tableau 1.6.2.

Les premiers couples de mots concernent assez souvent les éléments grammaticaux[6], (les couples les plus fréquents sont composés des mots les plus fréquents

[5]On trouvera souvent dans la littérature française le terme de *fréquence* pour désigner l'effectif d'un élément, ce qui est nous semble être un anglicisme (*frequency* : effectif, *relative frequency* : fréquence).

[6]Nous définissons un élément grammatical comme étant un élément (mot ou affixe) appartenant à une classe de marqueurs de frontières de syntagme et de proposition (sans être lui-même un syntagme).

Couples	Effectif	Rang
de la	2423	1
à la	980	2
et de	463	3
que les	287	7
n'est pas	189	24
le gouvernement	129	46
ministre de	120	52
secrétaire général	65	132

TAB. 1.4 – L'effectif reflète des relations à tous les niveaux de la structure.

en général). Les structures décrites mettent en relation des éléments appartenant à un même syntagme (le cas le plus fréquent), ou entre deux syntagmes (*ministre de*, *secrétaire général*), ou entre deux propositions (*que les*[7]). L'effectif d'un couple ne peut en aucun cas refléter la nature de la relation entre les deux éléments. Mais si l'on arrive à identifier cette nature, alors l'effectif devient un signe de relation entre les éléments. Ainsi, le couple *ministre de* indique une relation entre le syntagme comprenant le lexical *ministre* et le syntagme suivant commençant par *de*.

D'une manière générale, tout phénomène fréquent est une marque qu'il faut étudier et surtout comprendre. Travailler en premier sur les éléments fréquents permet de découvrir les structures fréquentes de la langue. Une fois ces structures traitées, il est alors possible de s'occuper des structures rares. L'inverse me semble très difficile. Ceci explique pourquoi les exemples qui illustrent ce travail concernent surtout les phénomènes fréquents.

La contiguïté Nos données sont constituées d'une séquence de mots compris entre des séparateurs (la ponctuation). Comme nous le verrons dans le chapitre 6, la construction des structures se base sur des séquences contiguës d'éléments. Nous nous sommes toujours restreint à rechercher les régularités dans un espace assez limité, pratiquement un espace de recherche d'un élément précédent et d'un élément suivant, l'élément correspondant au mot (pour la construction des syntagmes) ou au syntagme (pour la construction des structures). Étendre la recherche à toute la phrase, c'est-à-dire générer tous les couples formés de deux mots dans une phrase, ne produit aucun résultat intéressant. Ce traitement avait pour objectif le traitement des structures discontinues de la langue, en particulier la structure *sujet-verbe*. Ce type de travail ne donne que des résultats très limités. En particulier, il permet de mettre en relation des débuts et fins de proposition (comme les accords entre pronoms sujets (en début de proposition) et verbes (fin de proposition) en turc). Mais l'on s'aperçoit alors que la notion de discontinuité est relative, puisque, pour ces éléments, elle n'existe plus au niveau propositionnel et que ces résultats peuvent être obtenus en systématisant la recherche d'accords aussi bien au niveau syntagmatique qu'au niveau

[7] Peut aussi être une relation interne à un syntagme.

propositionnel.

Nous nous sommes donc contenté d'un espace de recherche de régularités d'un élément précédent et suivant l'élément traité aussi bien au niveau syntagmatique qu'au niveau propositionnel, ce qui est suffisant pour découvrir une immense partie des structures des langues. Ceci à pour conséquence de fournir un descriptif des structures sous forme de liste de couples. Cette représentation nous semble suffisante pour la représentation des structures (section 4.11). Nous avons dévelopé deux principes sur la recherche de structures :

- *La recherche des structures composées de plus de deux éléments peut (et doit) se ramener à la rechercher de structures composées de deux éléments, qui sont les seules structures observables.*
- *Toutes les structures composées de deux éléments peuvent être observées grâce à la contiguïté fréquente des deux éléments.*

La position Le critère que nous appelons *positionnel* est sans doute le plus remarquable, puisqu'il est indispensable à la construction de la structure, mais a aussi été le plus délicat à appréhender. Ce critère consiste à observer la position d'un élément dans une séquence. Par position, nous entendons le nombre d'éléments (plus un si l'on veut commencer à zéro) entre le début de la séquence et l'élément concerné. Au début de ce travail, étudiant sur les langues européennes, nous avons remarqué que certains éléments étaient placés assez souvent en début de séquence. Nous avons alors fait le rapprochement entre l'objet linéaire qu'est une séquence de mots et le traitement de l'objet informatique qu'est une pile, c'est-à-dire une séquence d'éléments. Dans une pile, deux éléments sont traités de façon particulière : le premier élément et le dernier. Nous avons alors pensé qu'il en était peut être de même pour la langue (même si les deux objets ne sont pas comparables, l'analogie a été intéressante puisqu'elle nous a permi d'acquérir le concept de symétrie dans les structures.). Et cela a été le cas : les débuts et fins de séquences correspondaient à des éléments aux propriétés caractéristiques. Donc *toutes les positions ne sont pas à étudier*, ce qui aurait été très coûteux (et même inutile), mais seulement les première et dernière positions. Nous reviendrons en détail sur ces observations au chapitre 4.

Maintenant donc ces trois choses demeurent : l'effectif, la contiguïté, la position ; mais la plus grande de ces choses, c'est la position.

1.7 Découverte ou apprentissage ?

> *Pour trouver quelque chose, il faut d'abord savoir ce que l'on cherche.*
> [Ramat, 1985, page 59]

Ce type de travail nous a bien sûr conduit vers les différents travaux réalisés dans le domaine de l'apprentissage en informatique. Deux grands paradigmes composent ce domaine : l'apprentissage supervisé et l'apprentissage non supervisé.

L'apprentissage supervisé L'apprentissage supervisé travaille avec des données auxquelles ont été associées un certain nombre de modalités qui ont pour

objectif de décrire les données. En particulier, dans un problème de classification, les données sont associées à la classe à laquelle elles appartiennent. Les algorithmes ont pour tâche d'établir des règles permettant de classer des données nouvelles.

L'acquisition automatique (l'aprentissage) de données linguistique n'est pas une tâche récente, puisqu'elle est apparue avec les premiers corpus électroniques [Andreewsky, 1973], [Fluhr, 1977].

Un exemple récent de ce type de travail, en traitement automatique des langues, est proposé dans [Brill, 1993]. À partir d'un texte où chaque mot est associé à son étiquette (texte étiqueté), le programme génère des règles contextuelles permettant l'étiquetage des mots apparaissant dans ces contextes. La taille maximale des contextes est de deux mots précédant ou suivant le mot à classer. Le tableau 1.5 donne quelques exemples de règles générées.

De *MODAL* ou *VERBE* à *NOM* si le mot précédent est *the*
De *PRÉPOSITION* à *ADVERBE* si le deuxième mot à droite est *as*
ADVERBE si le mot a pour suffixe *-ly*

TAB. 1.5 – Exemple de règles générées par le programme de E. Brill.

Ces règles sont produites grâce à des patrons comme ceux-ci :
changer l'étiquette X du mot en Y si

1. l'étiquette précédante est T
2. le mot précédent est W
3. la prochaine étiquette est T
4. le prochain mot est W

Des essais ont été menés en utilisant des textes non étiquetés, mais avec un dictionnaire associant à chaque mot la liste de ces étiquettes possibles [Brill, 1995]. On trouvera une description des algorithmes utilisés dans [Charniak, 1993]. Ces techniques s'appuyant sur une classification préétablie, ne peuvent nous convenir pour notre travail, puisque nous ne voulons utiliser ni lexique, ni corpus étiqueté.

L'apprentissage non supervisé L'apprentissage non supervisé travaille avec les données seules, sans inclure de connaissance sur celles-ci. Nous nous placons dans cette configuration. Dans le domaine des langues, il est principalement utilisé en catégorisation automatique. Les objets manipulés sont les mots d'un texte. Les techniques habituellement utilisées pour générer des catégories de mots sont décrites dans la section 3.3, ainsi que les raisons qui nous ont fait renoncer à ces techniques. Elles se basent sur un calcul de distance entre mots, distance entre les contextes des mots. Les contextes sont définis comme étant la suite de n mots encadrant le mot, n étant généralement égal à 1 ou 2 (mais pouvant aller jusqu'à 100).

Si la catégorisation des mots est une opération importante de notre travail, elle n'en reste pas moins une opération terminale. Nous pensons en effet que cette opération ne peut être menée à bien que grâce à la connaissance structurelle de la langue (section 3.4).

Découverte *et* apprentissage En fait, la réponse à la question de cette section est : découverte *et* apprentissage. Dans un premier temps, il a fallu *découvrir* les concepts nécessaires à la mise au point de la méthode, en utilisant des outils d'observation de corpus. Cette phase d'observation est totalement supervisée, l'ordinateur ayant servi d'outil d'exploration. Puis dans un deuxième temps, et en utilisant les concepts trouvés, il a fallu catégoriser les éléments de la langue et générer les structures de la langue, grâce à des algorithmes que l'on peut ranger dans le paradigme de l'apprentissage non supervisé, puisque le résultat, pour une langue donnée, n'est fourni à aucun moment du traitement. Ce deuxième travail n'a pour objectif qu'une validation des concepts linguistiques trouvés lors de la première phase.

Différence entre découverte et analyse Les travaux en TAL portent généralement sur des procédures d'analyse. Quelle différence faisons-nous entre notre travail et les travaux d'analyse ? Nous résumerons la chose en disant que, dans un processus de découverte, le but est d'identifier les objets, ici les structures de la langue, alors que, dans un processus d'analyse, le but est d'assigner à chaque objet du corpus sa catégorie. Le processus de découverte nécessite une analyse mais seulement partielle. Tout le corpus d'apprentissage n'a pas besoin d'être analysé. Un de nos objectifs est de trouver les catégories possibles d'un élément, disons un mot, dans une langue donnée. Pour cela, il n'est pas nécessaire d'assigner une catégorie à chaque occurrence du mot dans le corpus. *L'objectif de ce travail n'est donc pas la réalisation d'un analyseur syntaxique.* La plupart des systèmes d'apprentissage (tous supervisés) fusionne souvent ces deux processus [Brill, 1995], [Chanod and Tapanainen, 1995]. Le résultat final fournit une analyse, et c'est généralement cette dernière qui sert à évaluer le système. Dans le meilleur des cas, notre processus de découverte pourrait fournir des informations au processus d'analyse (prenez plutôt un locuteur de la langue). Le but de ce travail n'est pas opératoire : nous nous plaçons plutôt dans un cadre expérimental en essayant de répondre à la question : que faire avec un texte et un ordinateur ?

L'inférence grammaticale On trouve deux paradigmes très différents sous le terme d'inférence grammaticale. Si la définition est commune :

> Given a set of strings that the grammar is supposed to generate, the Grammatical Inference problem is one of inferring a grammar that satisfies these strings, and is also able to generalise to other unseen strings [Hutchens, 1994].

la différence porte sur l'objet étudié, en fait sur la nature de cet ensemble de chaînes ("set of strings"). Certains, [Miclet and de la Higuera, 1996], s'intéressent plus particulièrement à la théorie des grammaires formelles, grammaire pris dans son sens mathématique[8]. La langue n'est donc pas l'objet d'étude. Le deuxième paradigme est plus centré sur la langue : les séquences produites

[8] A grammar G, for a language L is a (computable) function, which when given as input a sequence s, outputs 1 iff $s \in L$, and 0 iff $s \notin L$. [Finch, 1993, page 65]

sont ou se veulent être des exemples d'une langue.Dans ce cas, le type de données est assez variable. Certains utilisent des données créées artificiellement à partir d'une grammaire formelle et essayent de la regénérer. Les techniques algorithmiques utilisées sont diverses : symboliques [Wolff, 1980], numériques, [Stolcke and Omohundro, 1994], à base de réseaux neuronaux [Elman, 1990], [Kohonen, 1978]. Dans les données artificielles on essaie de reproduire la structure de la langue en simple (généralement une simplification de la taille du vocabulaire). Les grammaires utilisées sont très simples (tableau 1.6). Les phrases de trois mots semblent aussi avoir droit à un traitement particulier (tableau 1.7).

S	→	NP VP
VP	→	V NP
NP	→	DET N
	→	NP RC
RC	→	REL VP
DET	→	a
	→	the
N	→	cat
	→	dog
	→	mouse
REL	→	that
V	→	heard
	→	saw

TAB. 1.6 – Exemples de grammaire utilisée par [Stolcke and Omohundro, 1994, page 115]

- Mary likes meat
- Jim speaks well
- Mary likes Jim
- Jim eats often

TAB. 1.7 – Exemples de données utilisés par [Kohonen, 1978]

On comprend que les traitements développés avec ce type de données ne produisent aucun résultat satisfaisant avec des données réelles (de l'aveu des auteurs eux-mêmes) , en particulier la polycatégorisation des éléments (ici les mots) n'est jamais prise en compte. Ce qui fait que ces données ont l'apparence de données correspondant à une langue *naturelle*, mais seulement l'apparence. Ces techniques ne peuvent donc pas servir dans notre travail.

Il existe aussi un autre type de travail, que l'on trouve parfois sous le terme d'inférence grammaticale, et qui se rapproche plus des sciences cognitives. L'objet est ici le problème de l'acquisition d'une langue par un enfant [Brent, 1996], [Cartwright and Brent, 1997]. Nous reparlerons de ce travail à la section 7.3.

Il existe de plus en plus de travaux associant langue et apprentissage (création du SIG SIGNLL[9] (SIG in Natural Language Learning) en 1992). Cette communauté s'intéresse à tous les aspects qui prennent en compte langues et apprentissage, de l'acquisition de connaissances (linguistiques) à la théorie de l'acquisition de la langue chez l'humain. On trouve dans [Daelemans and Powers, 1992] et [Powers, 1998] un exellent panorama des différents travaux effectués dans ce domaine.

Un travail de linguistique assistée par ordinateur

> *La puissance de calcul et l'augmentation de la capacité de stockage ont permis une explosion de l'utilisation de l'ordinateur dans ce domaine* [Dessen, 1995].

Un tel propos aurait pu être tenu en linguistique informatique, mais il provient, en fait, d'un article paru dans une revue de biologie, et s'applique au domaine de la bioinformatique. L'utilisation la plus connue étant les travaux portant sur le séquençage du génome. La similitude est frappante entre le travail effectué en bioinformatique et en linguistique informatique, et ces propos peuvent être appliqués parfaitement au TAL, avec la venue d'un nouveau champ baptisé *linguistique de corpus* (section 1.10). Le terme "explosion" est peut être exagéré en linguistique et concerne une partie seulement des travaux (TAL et linguistique descriptive), même si de plus en plus de domaines ont recours à une utilisation de l'ordinateur à travers la manipulation des corpus électroniques, la simulation, ou comme outil de validation. On notera l'emprunt (partiel) par la bioinformatique du vocabulaire et des outils de l'informatique linguistique, dû à la similarité (linéaire) entre les séquences de mots et séquences d'ADN[10]. Il suffit de prendre les titres d'articles comme : *Linguistics of nucleotide sequences : morphology and comparison of vocabulary* [Brendel et al., 1986] pour s'en rendre compte. De la même manière que la bioinformatique a ouvert de nouvelles perspectives en biologie, l'ordinateur joue un rôle important dans l'établissement et la validation de théorie linguistique. Mais surtout l'ordinateur a permis une exploration des données qu'il n'était pas possible (ou si fastidieuse) de réaliser "manuellement". Il nous semble que la mise au point de la méthode décrite dans cette thèse est difficilement envisageable ou réalisable sans utilisation de l'ordinateur dans la manipulation des données, celles-ci étant trop volumineuses. Si l'ordinateur a bien sûr un rôle central dans les nouveaux domaines du TAL et celui de l'informatique documentaire, son utilisation en linguistique "classique" n'est pas sans intérêt. Voilà pourquoi nous qualifions notre travail de *linguistique assistée par ordinateur*.

1.8 Le déchiffrement de langues et d'écritures

> *Much more than reading, deciphering is a genuinely linguistic task, and it is quite surprising, therefore, that linguists have taken prati-*

[9] `http://pi1093.kub.nl/~signll/`

[10] On retrouve aussi cet emprunt en musique. Dans une interview télévisée, un pianiste parlait de *phrases* pour morceaux de musique.

cally no interest at all in this most challenging activity [Coulmas, 1989, page 207].

Durant notre travail, nous nous sommes intéressé aux travaux concernant le déchiffrement de langues anciennes ou d'alphabets. Nous avons eu envie de faire un parallèle entre notre travail et celui réalisé par les linguistes qui se sont attelés au déchiffrement de langues et d'écritures. Notre travail est-il similaire à un travail de déchiffrement ? Oui et non. Non, car le but du déchiffrement est d'obtenir l'information qui est contenue dans le document. Notre but est de savoir "seulement" quelle est la structure de la langue dans laquelle le document est écrit. Oui, car connaître la structure de cette langue est un renseignement très important pour le déchiffrement. Pour aboutir au déchiffrement d'un document, des informations historiques, archéologiques, linguistiques sont nécessaires. L'on peut dire que tous les moyens sont bons et doivent être utilisés. Dans le cadre de notre travail, dont l'objectif n'est pas le même, seules les régularités formelles doivent être prises en compte. Les techniques utilisées pour déchiffrer une langue se basent essentiellement sur l'étude de textes multilingues.

langue connue	écriture connue	
+	+	lecture
+	-	déchiffrement 1
-	+	déchiffrement 2
-	-	déchiffrement 3

TAB. 1.8 – Lecture et déchiffrement [Coulmas, 1989].

Il existe en fait plusieurs types de déchiffrements, selon la connaissance que l'on a de la langue et du système d'écriture utilisé (tableau 1.8). Dans notre cas, nous pouvons dire que nous sommes dans la configuration : langue inconnue et écriture connue. Nous pourrions nous placer dans le cas : langue inconnue et écriture inconnue, mais travaillant sur des textes électroniques, nous ne pouvons considérer que le système d'écriture nous est inconnu. Nous nous plaçons donc dans le cadre du déchiffrement numéro 2.

Influence du système d'écriture sur le travail Parler du système d'écriture n'est pas sans rapport avec notre problème. Il nous est apparu que la manière utilisée pour écrire un texte pouvait compliquer ou faciliter notre travail. Un système d'écriture "parfait" ou très pratique serait un système dans lequel les mots de ce système correspondraient aux unités manipulées dans ce travail : les syntagmes simples et les propositions. Cela n'est jamais le cas. Cependant la segmentation en mots est un assez bon point de départ pour une procédure de découverte. Il faut simplement avoir conscience que les unités résultantes de cette segmentation, les mots, ne sont pas (dans la plupart des cas) l'unité de base de la structure linguistique, et qu'une opération de segmentation est alors nécessaire.

Un texte s'adresse généralement[11] à un lecteur qui comprend la langue du

[11] Sauf dans le cas de textes cryptés.

texte. Le système d'écriture peut être alors assez "pauvre" ou déficient dans certains points de la langue. Ns n lctr frnçs prrt nrmlmnt lr cs mts. Bt h wll rd ths wrds wth mr dffclts. Lasegmentationjoueaussiunerôledanslalecture. Elle peutêtre nullemaisrare mentin correcte. Lcmbnsndcsdxdffcltsstnsrmntbl.[12]

La mise au point du système d'écriture Nous allons donner quelques indices permettant la découverte du type de système d'écriture utilisé pour un texte donné. Nous avons dit plus haut que nous nous placions dans la configuration : langue inconnue et système d'écriture connu. Voyons quelles auraient été les méthodes à employer pour découvrir le système d'écriture d'un texte. Le premier travail à effectuer est un recensement des symboles utilisés, qui permet généralement de décider si l'on a affaire à un système idéographique ou phonétique (alphabet ou syllabaire). Prenons l'exemple du travail de Champollion. En travaillant sur une copie de la pierre de Rosette, il constata que le texte grec était constitué de 486 mots, et l'égyptien de 1419 signes [Février, 1948]. Sur ces 1419 signes, il en existait seulement 66 différents. Sa conclusion était simple : le texte hiéroglyphique ne pouvait être écrit dans un système idéographique, mais plutôt phonétique, alors que depuis Horapollon (390 av. J.C.), les hiéroglyphes était considérés comme représentant des idées[13]. Ce simple comptage du nombre d'éléments apparaissant dans un texte est pourtant une opération élémentaire mais essentielle. Ce recensement permet d'établir la liste des signes de la langue (une centaine pour un système phonétique). La segmentation en "mots" se fait de manière visuelle en cherchant les ruptures dans les séquences de signes. Un fait essentiel de la segmentation est qu'elle est généralement régulière, c'est-à-dire que les "coupures" entre "mots" ainsi définis se retrouvent aux mêmes endroits (une même séquence n'est généralement pas segmentée de différentes façons). La principale difficulté rencontrée concerne les systèmes d'écritures qui mixent les différentes possibilités (comme le japonais qui utilise idéogrammes et syllabaires). La segmentation en mots doit alors prendre en compte ces deux systèmes (section 6.1). Une fois le type d'écriture défini, l'étape suivante est de trouver le sens de lecture du texte. Il existe plusieurs conventions : de droite à gauche, de haut en bas, en boustrophédon (on écrit par exemple de gauche à droite, puis arrivé en fin de "ligne", l'on écrit la ligne suivante de droite à gauche en partant de la fin de la ligne précédente). Décider si l'écriture utilise un sens vertical ou horizontal est assez facile, uniquement sur des critères visuels (lorsque l'on travaille sur un texte, la chose est plus délicate pour une inscription courte). Le cas le plus délicat est une écriture qui ne se lit pas linéairement comme l'écriture maya, où les "lignes" sont des colonnes composées de deux glyphes (figure 1.4). De plus la composition des glyphes peut aussi varier (un symbole ou plusieurs, avec différents sens de lecture). Le cas est similaire pour l'écriture hiéroglyphique égyptienne [Champollion, 1997, pages 18-21]. Dans ces cas là, la procédure de construction des séquences de signes sera beaucoup plus complexe que dans le cas simple d'un texte écrit dans un système d'écriture

[12] La combinaison de ces deux difficultés rend la tâche insurmontable. Les phrases précédentes ne devraient pas poser de problèmes.

[13] Le système égyptien comptait 700 signes en 3000 av. J.C.

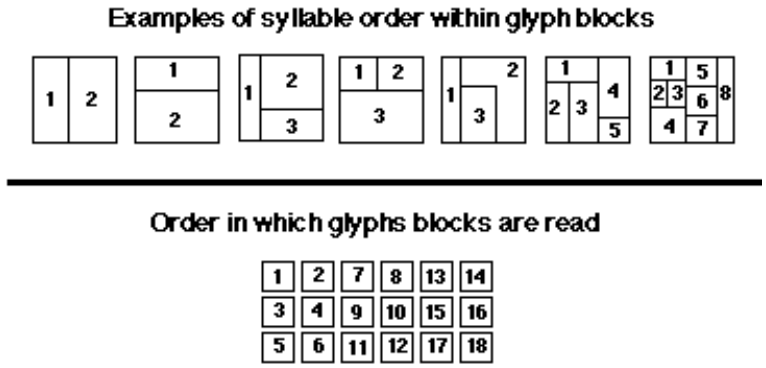


FIG. 1.4 – Ordre de lecture de glyphes mayas.

proche des systèmes européens.

La détection des signes de ponctuations (s'ils existent) est aussi une tâche importante. Ces ponctuations correspondent à des signes fréquents généralement assez simples du point de vue graphique. De plus la plupart de ces éléments se situent en fin de séquences. La ponctuation et la segmentation ne sont pas des conventions récentes (l'écriture ougaritique (1400 av. J.C.), classe I de l'écriture de Persépolis (600 av. J.C.) [Février, 1948, page 572]), même si tous les systèmes ne les utilisent pas. Nous verrons dans la section 4.1 l'importance de la ponctuation dans ce travail.

Les autres caractéristiques Certaines caractéristiques visuelles peuvent aussi être prises en compte dans le processus de découverte, car elles peuvent indiquer des relations entre éléments ou la nature des éléments. Par exemple la différence très nette (visuelle) entre certains signes du système d'écriture japonais. Certains sont assez simples (et aussi fréquents), d'autres ont une graphie plus recherchée (et un effectif plus faible). L'utilisation de deux systèmes de signes est ainsi facilement découverte, d'autant plus que l'un des systèmes est utilisé dans un emploi structurel bien spécifique (le système des kana est utilisé pour noter les suffixes, donc des marques de fin). Les cartouches égyptiens offrent aussi une petite indication (elles indiquent les noms propres). L'utilisation des majuscules fournit aussi des indices (segmentation en phrases, identification des ponctuations, et même catégorisation des mots en allemand). Un autre indice concerne le système utilisé pour noter les nombres dans le texte. Dans notre travail, nous mettrons de côté tous ces indices, très dépendant du système d'écriture ou de la langue, pour ne prendre en compte que les régularités multilingues (section 1.6).

1.9 Le minimum de connaissances

Pour réaliser ce travail, nous essayons de partir avec le moins de connaissances possibles. Mais nous ne partons pas de rien. En pratique, nous avons supposé connu la liste des signes et le système de ponctuation et de segmenta-

tion en mots (section 1.8). Nous ne considérons pas un texte comme une suite de symboles équivalents. La connaissance du système d'écriture nous permet d'obtenir deux niveaux de segmentation : la segmentation en mots et en unités que nous nommerons "entre-ponctuation" . Un mot est défini comme une suite de symboles délimitée par un espace ou une ponctuation. Nous retrouvons la définition basique du mot. Cette définition s'applique pour les langues dites alphabétiques. Pour les langues utilisant un système idéographique (chinois) le mot correspondra à un signe du système. Comme nous le verrons dans la section 4.4.2, le mot est une unité de la "strate écrite" et est utilisé comme point de départ de la découverte des structures. Rappelons, que travaillant sur une corpus électronique, la segmentation en symboles est déjà effectuée. Les unités dites "entre-ponctuation" sont définies comme étant une séquence de mots comprise entre deux ponctuations. Ces deux niveaux de segmentation vont nous offrir deux points d'accès à la structure des langues (Chapitre 4).

Tous les signes n'appartenant pas à la liste des ponctuations sont considérés comme appartenant au système d'écriture (en particulier l'apostrophe et le tiret font partie des mots). La liste des signes de ponctuation utilisés est la suivante :

```
? , . ; : !
```

Les signes considérés comme appartenant à l'alphabet de la langues sont :

abcdefghijgklmnopqrstuvwyxz
ABCDEFGHIJKLMNOPQRSTUVWXYZ
ãåáàâäéèêëíìîïòóôõøöúùûüçñ
ÃÀÂÄÉÈÊËÍÌÎÏÒÓÔÕØÖÚÙÛÜÇÁÍÓÚÑýÝÿÆæß'-

Pour les corpus qui ne sont pas écrits avec un alphabet dérivé de l'alphabet latin (coréen, chinois, japonais), la première étape consiste à trouver s'ils contiennent des ponctuations. La deuxième étape consiste à rechercher si le système admet une segmentation en mots (en utilisant un critère visuel). Si cette segmentation existe, le signe segmentant est le plus courant du texte. Sinon tous les autres caractères sont considérés comme faisant partie du système d'écriture. Toutes ces étapes se font de manière supervisée.

Une remarque importante est que *les différents systèmes d'écritures jouent un rôle dans la procédure informatique de découverte des structures, mais ne peuvent en aucun cas invalider la structure théorique des langues mise au point durant ce travail.* Si ces différences de systèmes d'écriture peuvent générer des différences dans les traitements, elles n'en restent pas moins opératoires. Par exemple, la construction des syntagmes (section 6.6) est réalisée différement si l'on traite le japonais ou le norvégien, mais dans les deux langues, cette structure existe (ainsi que toutes les autres structures décrites au chapitre 4).

1.10 Le travail sur corpus

Cette section introduit quelques remarques sur l'utilisation du corpus dans notre travail, ainsi que quelques caractéristiques des corpus utilisés. Le détail des corpus utilisés se trouve en annexe A.

1.10.1 La linguistique de corpus

Si, comme nous l'avons vu à la section 1.4, le travail sur corpus n'a pas toujours été en odeur de sainteté, son utilisation actuellement ne semble plus controversée. Pour plus de précision, nous renvoyons le lecteur à [Woodley, 1995] et à [Habert et al., 1997], en particulier à son introduction qui resitue historiquement la linguistique de corpus (le terme provient de l'anglais "corpus linguistics"). Dans ce travail, l'utilisation de corpus dans ce travail n'est pas fondée a priori sur une argumentation méthodologique mais pratique. En effet, le corpus est sans doute le meilleur moyen de travailler sur une langue étrangère, le recours au locuteur étant trop astreignant (pour tout le monde).

1.10.2 La composition des corpus

Un problème classique dans l'utilisation de corpus (et en général de données) est leur constitution. Comment obtenir des données représentatives ? Mais représentatives de quoi ? Il nous était impossible au début de ce travail de répondre à cette question. Nous avons évité de nous poser ce problème, et la sélection des textes s'est faite un peu au hasard. Notre travail de constitution a été grandement facilité par le développement du Web. Par ce médium, les textes dans des langues variées ont alors été accessibles très rapidement, sinon directement. Les corpus des langues européennes sont d'origines diverses. Pour les autres langues, le corpus est le plus souvent constitué d'une partie de la Bible, cet ouvrage étant souvent traduit (et généralement le premier traduit) dans des langues à tradition orale. De plus, ce critère de recherche dans l'hypertoile fournissait directement une quantité de textes suffisante. Nous avons essayé de prendre des langues assez variées dans leurs structures, en utilisant les critères traditionnels (langues préposées et postposées, isolantes ou synthétiques). Les corpus n'ont pas été fabriqués : ils sont composés généralement d'un seul texte ou de plusieurs textes *entiers*.

Un corpus représentatif Nous allons voir que le problème de la représentativité des corpus, dans notre étude, n'est pas un problème crucial. Le problème de la représentativité du corpus ne concerne qu'indirectement notre travail pour deux raisons. Premièrement, parce qu'un corpus de 500 000 mots contient énormement d'information sur les structures formelles d'une langue (les structures syntagmatiques et propositionnelle ont un nombre d'occurrences de plusieurs milliers). Deuxièmement, notre objectif n'est pas de donner une description complète d'une langue, mais de mettre au point une méthode de découverte des structures formelles des langues. Cette méthode est mise au point à partir de corpus, mais ne change pas d'un corpus à un autre, ni d'une langue à une autre (au moins dans ces principes généraux : les différents systèmes d'écriture nécessitent un traitement légérement différent en pratique). Plus le corpus contiendra d'information, plus le résultat sur une langue donnée sera complet, mais la méthode ne changera pas. Bien sûr, certaines structures de la langue peuvent ne pas avoir été prises en compte dans notre méthode, mais les structures trouvées dans les corpus étudiés fournissent déjà assez de grains à moudre.

De plus, l'approche multilingue nous a conduit à générer des schémas structuraux qui couvrent des nombreuses configurations (chapitre 4).

Morphèmes	Bible			Rapport technique		
	Effectif	Début	Fin	Effectif	Début	Fin
bir	2029(2)	227	5	701(2)	52	0
için	1152(4)	0	67	267(7)	29	0
ama	763 (10)	743	10	15(215)	11	0
dedi	764	0	712	0	0	0
bütün	291(58)	107	0	66(37)	19	0
tek	98(199)	23	0	26(99)	9	0
-yor	742	14	436	15	0	9
-dır	399	13	358	418	1	393

TAB. 1.9 – Effectif d'éléments dans deux types de corpus en turc. Si l'effectif peut varier d'un corpus à l'autre, le comportement positionnel des éléments est assez stable. Les nombres entre parenthèses indiquent le rang de l'élément.

Il n'est quand même pas inutile de comparer les résultats obtenus sur différents corpus. Le tableau 1.9 montre certaines différences entre deux corpus turcs, mais surtout certaines ressemblances. Les deux corpus comparés sont le nouveau testament (*turc01*) et un rapport scientifique d'une université turque datant de 1995 (*turc02*). Le premier comprend 129909 mots et signes de ponctuations et le deuxième en comprend 33001. Nous avons comparé ces deux textes selon deux critères : l'effectif des éléments et leur comportement positionnel.

Le premier critère met à jour des différences assez nettes quand à l'effectif de certains éléments. Bien sûr, ces différences sont très présentes au niveau lexical, la thématique des deux textes étant très éloignée. Comme nous n'utilisons aucun critère sémantique, ces différences ne jouent aucun rôle dans notre travail. Nous utilisons seulement le fait qu'un élément est de nature lexicale, peu importe cet élément (ou son sens). Mais l'on note aussi des différences au niveau grammatical. Par exemple, l'élément *ama* (mais) est beaucoup moins présent dans le rapport que dans la bible (il faut prendre en considération le rang et non l'effectif, puisque les deux corpus sont de tailles différentes). De même pour *dedi*, qui correspond à un élément d'une structure du discours direct, totalement absent du rapport. Il y a donc des éléments que l'on retrouvent dans certains corpus et non dans d'autres. Le résultat était attendu.

Nous allons maintenant considérer le deuxième critère : le comportement positionnel des éléments. On remarque que, lorsque deux éléments sont présents dans les deux corpus, leur comportement positionnel est identique. Reprenons l'élément *ama*. Il est catégorisé comme début absolu dans le premier corpus, ainsi que dans le deuxième, même si son effectif dans ce dernier est très faible par rapport au premier corpus. Il en est de même pour l'élément (marque du progressif) *-yor*. En fait, il nous importe peu qu'un élément comme *ama* soit fréquent ou non. L'important est que l'on retrouve bien nos marqueurs de frontière quel que soit le corpus utilisé. Nous pouvons en fait comparer les différents corpus selon deux critères : les structures qui composent ces corpus, et les éléments

utilisés dans ces structures. Les différences structurelles trouvées entre corpus concernent plus spécifiquement des structures de haut niveau, par exemple les structure liées au discours direct ou indirect. Le morphème et le syntagme sont des unités beaucoup plus stables. Mais, au niveau morphologique et syntagmatique, les différences structurelles sont très faibles (si dans un corpus d'une langue donnée, on ne trouve pas de marqueur de fin de syntagme, l'on n'en trouvera pas dans un autre corpus). Les catégories sont donc très stables d'un corpus à l'autre (on retrouve les catégories de débuts et/ou de fin), et les éléments les plus fréquents d'une classe se retrouvent aussi (les prépositions les plus courantes par exemple)

1.10.3 Analyse quantitative

Nous allons maintenant donner quelques caractéristiques chiffrées des textes. Ceci afin de prendre conscience de certaines propriétés de la langue (des textes), propriétés qui jouent un rôle prépondérant dans cette procédure de découverte.

La loi de Zipf Une caractéristique des textes écrits dans une langue est la *loi de Zipf*, du nom du linguiste George Kingsley Zipf [Zipf, 1949]. Prenez un texte, et classez tous les mots de cet échantillon dans l'ordre des effectifs décroissants (tableau 1.10). Le mot de rang 1 est le mot qui apparaît le plus souvent dans le corpus, le mot de rang 2 est celui qui apparaît le plus souvent exception faite du mot de rang 1. Ainsi de suite. La loi de Zipf énonce que l'effectif d'un mot est inversement proportionnel à son rang dans la liste. On a donc :

$$r \times f = \text{constante}$$

où r est le rang d'un mot et f sa fréquence, ceci quels que soient la langue et le corpus utilisé.

Rang	Effectif	$r \times f$
10	3807	38070
20	1759	35180
50	558	27900
100	229	22900
500	54	27000
1000	29	29000
5000	5	25000

TAB. 1.10 – La loi de Zipf : le produit *Rang*×*Effectif* est constant.

Cette loi avait déjà été observée par le sténographe J. B. Esproub. Les figures 1.5 illustrent cette loi pour les langues suivantes : français, turc, swahili et vietnamien. [Mandelbrot, 1968] a donné une deuxième approximation de ce phénomène. La formule devient :

$$(r + b)^a \times f = \text{constante}$$

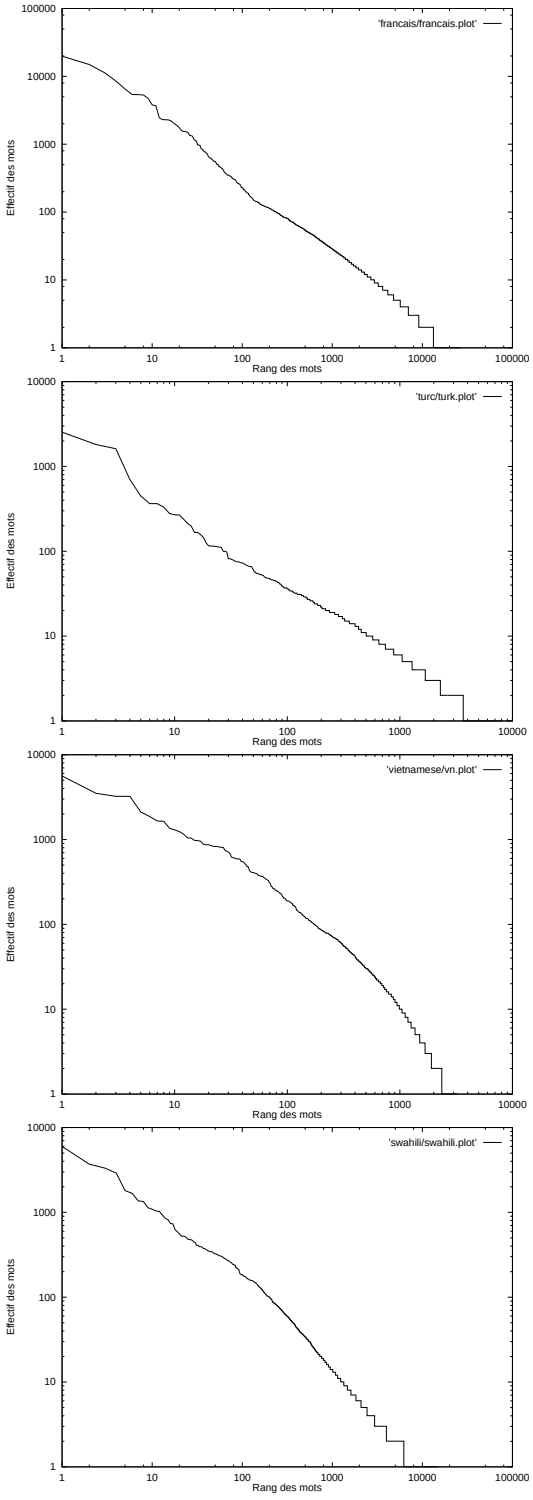


FIG. 1.5 – La loi de Zipf (échelle logarithmique)

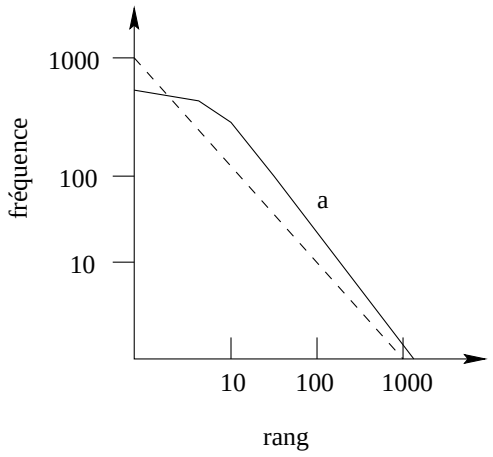


FIG. 1.6 – Nouvelle approximation [Mandelbrot, 1968].

Le facteur b est un facteur correctif pour les premier éléments qui ont un comportement déviant par rapport au reste des mots. L'exposant a est légèrement supérieur à 1 (figure 1.6).

Dans notre travail, nous ne nous servons pas directement de cette loi, mais elle met en évidence un fait très important : tous les mots de la langue ne sont pas équiprobables. Un petit nombre, une centaine, représente près de 50% des mots d'un texte. L'on peut concevoir ces mots comme le squelette structurel de la langue (chapitre 3). Ils correspondent en grande partie aux éléments grammaticaux de la langue. L'observation faite au niveau des mots est aussi valable au niveau des morphèmes (section 3.1). Cette propriété est-elle spécifique à la langue ? Loin de là. On la retrouve dans beaucoup de données : distribution des revenus, du nombre d'habitants des villes, des commerces d'après leur nombre de points de vente, etc [Guiraud, 1968, pp. 155]. L'on voit donc que des objets très divers obéissent à cette loi. [Guiraud, 1968] en conclu que :

> L'équation rang-fréquence apparaît partout où l'on définit les catégories observées comme la somme d'un certain nombre d'unités de base ; c'est une propriété de la substance discrète (discontinue et numérable). [Guiraud, 1968, pp. 156-157]

Si cette loi nous fait prendre conscience de certaines propriétés des langues, elle n'est pas directement exploitable dans notre travail.

Quelques autres caractéristiques Nous pouvons observer un certain nombre de caractéristiques à travers des mesures sur certains faits simples. En particulier le taux de "couverture" des mots les plus fréquents de la langue, c'est à dire le pourcentage que représentent ces mots dans le corpus par rapport au nombre total de mots. Le tableau 1.11 fournit ces chiffres pour quelques langues assez diverses. Ainsi, en accord avec la loi de Zipf, les dix mots les plus fréquents ont un taux de couverture d'au moins 10%. Et les cents mots (sur au moins 3000 mots minimum) les plus fréquents représentent 50% du texte. Nous pouvons

aussi voir apparaître quelques différences entre langues, la plus flagrante étant les différences entre le vietnamien et les autres langues. Différences qui reflètent la structure isolante du vietnamien par rapport aux autres langues. Le pourcentage d'hapax du turc (10,6%) peut aussi donner une indication sur le caractère agglutinant de la langue. De telles caractéristiques se sont pas pris en compte dans notre méthode, mais, étant très rapides à calculer, elles peuvent servir de guide dans une méthode supervisée.

	anglais	français	swahili	turc	vietnamien
taille du corpus	100070	100097	103580	104480	103758
nb mots différents	6655	14739	11907	15018	3270
couverture 10 mots	21.1%	20.4%	15.8%	8.89%	13.3%
couverture 100 mots	57%	52.1%	52.3%	44.4%	66.2%
hapax (relativement au corpus entier)	3,7%	5,7%	4,8%	10.6%	0.09%
longueur des mots ($\overline{x}$)	7.3	8.3	10	8.7	4.6

TAB. 1.11 – Quelques caractéristiques numériques sur les corpus.

Deuxième partie

La morphologie

Introduction

Cette partie concerne le travail que nous avons effectué au niveau morphologique, c'est à dire l'étude de la formation des mots. L'objectif de ce travail n'est pas de réaliser une analyse morphologique des mots, mais de relever les renseignements que la morphologie d'une langue peut nous apporter dans notre recherche des structures. Ceci va être réalisé en segmentant les mots du corpus afin d'obtenir des régularités mettant en jeu, non pas seulement les mots, mais des éléments "plus petits" : les morphèmes[14].

Durant ce travail, nous allons nous trouver devant deux situations très différentes. La plupart des langues admettent une morphologie, et ce travail aura alors une grande importance. Mais certaines langues (chinois, vietnamien) de par leur système d'écriture, sont considérées comme ne possédant pas de morphologie, c'est à dire que leurs mots sont indécomposables. Dans cette configuration, notre méthode de segmentation ne produit pas de résultat significatif.

Il est très important de noter que la segmentation n'est pas un but en soi et n'a d'intérêt que parce qu'elle permet d'aider à la découverte de relations entre éléments. Son intérêt est de fournir des éléments autres que les mots afin de construire les structures de la langue.

Nous verrons que certains morphèmes ont un rôle très important dans l'établissement de relation entre éléments. D'autres, par contre, ne jouent aucun rôle relationnel. Ainsi, il est sans intérêt pour nous de savoir si *délayer*, *début*, *décence* se décompose en *dé-layer*, *dé-but*, *dé-cence* ou non. Par contre, de savoir que *délayer* et *délayent* se décompose en *délay-er* et *délay-ent* est d'un grand intérêt, puisque ces morphèmes correspondent à des éléments relationnels du français. La distinction entre ces deux types de morphèmes se fait facilement puisque le premier n'exerce aucune contrainte sur son environnement, alors que le second impose certaines structures, donc certaines régularités détectables. Les deux types de morphèmes intéressants sont les morphèmes qui jouent un rôle relationnel et ceux qui, dans un contexte donné, catégorisent un élément dans une classe distributionnelle précise (comme *de N-er*[15] qui correspond à une structure infinitive à 88%[16] et *ils N-ent* à une structure verbale dans 100% des cas).

Cette partie s'organise selon le plan suivant : le chapitre 2 explique l'intérêt du processus de segmentation des mots et sa réalisation. Le chapitre 3 décrit la génération des séquences morphologiques, montre quelles sont les limites du

[14] Nous appelons cette unité un morphème, mais le terme de morphe est peut être plus adéquate (section 4.5).

[15] N-er correspond à un mot finissant par *-er* (N pour Noyau).

[16] Comptage réalisé sur 761 occurrences de la structure.

seul critère morphologique dans la découverte des structures formelles. Par séquence morphologique, nous entendons une séquence composée de mots ou de morphèmes grammaticaux[17]. Nous verrons que la génération de *couples morphologiques* (séquence morphologique composée de deux éléments) est suffisante dans notre travail.

Les programmes sont donnés en annexe B, et les résultats ainsi que leurs évaluations en annexe C.

[17] les élément de nature lexicale appelés *Noyau* dans ce travail (section sec :syntagme) sont représentés par la lettre N : la séquence [la N-ion] représente la séquence morphologique qui permet d'identifier tous les couples de mots dont le premier est *la* et le second un mot finissant par *ion*.

Chapitre 2

La découverte des morphèmes

Sommaire

2.1 L'intérêt de la segmentation

Pourquoi segmenter les mots du corpus ? Simplement parce que le "mot"[18] n'est pas l'unité de base de la structure linguistique, et qu'en s'en tenant à cette segmentation en mots, unité de l'écrit, nous ignorerions certains faits indispensables à la découverte de la structure des langues.

Selon [Harris, 1954, pages 28-29] :

> Le premier fait distributionnel est la possibilité de diviser (de segmenter) toute chaîne parlée en parties, de façon à découvrir certaines régularités d'occurrence de l'une des parties, relativement à d'autres parties de la chaîne parlée.

Cette observation est très pertinente. Il existe en effet des régularités qui échappent aux observations si nous nous contentons d'un travail au niveau des mots.

Prenons l'exemple turc illustré par le tableau 2.1. Il contient un certain nombre de couples de mots contigus. Ces couples ont tous un effectif[19] faible. Nous considérons donc qu'il n'existe pas de régularité particulière entre les mots composant ces couples. Si nous ne regardons plus ces mots comme étant unitaire, mais composés d'autres éléments, on voit apparaître une régularité entre

[18] Notre définition du mot est donnée à la section 1.9

[19] La taille et les autres caractéristiques des corpus sont données en annexe

Couple de mots	Effectif	Couples de mots	Effectif
ölümden diril*tip*	1	ölümden diril*mis*	1
ölümden diril*tirken*	1	ölümden diril*mek*	2
ölümden diril*tmeye*	1	ölümden diril*ttiine*	2
ölümden diril*mesi*	1	ölümden diril*tti*	5

TAB. 2.1 – Le couple *ölümden diril-* a un effectif total de 57 occurrences. Nous avons bien une relation entre *ölümden* et *diriltiken* bien que l'effectif de ce couple soit de 1.

le mot *ölümden* et un "morceau" du mot suivant : *diril*. Ce couple *ölümden diril-* a un effectif de 57, alors que l'effectif du mot *ölümden* est de 67. De cette observation, nous en déduisons qu'il existe une relation entre *ölümden* et tous les mots suivants qui commencent par *diril*, même si le couple formé par ces deux éléments a un effectif de 1.

La connaissance de la segmentation des mots en deux parties, radical et affixes, nous permet donc de trouver une régularité entre ces séquences de mots, qui est très difficile à observer dans le cas de la manipulation de mots. La régularité décrite ici concerne deux éléments lexicaux *ölüm(den)* et *diril*, mais elle concerne le plus souvent des éléments grammaticaux. Ainsi dans le tableau suivant (tableau 2.2), une régularité apparaît grâce à l'affixe des mots précédant le mot *için*. Quel que soit l'effectif du mot *için* avec son précédent, si ce dernier a pour suffixe *-mak*, alors les deux éléments seront en relation.

Séquences	Effectif
yazılmak için	1
bulmak için	1
katılmak için	1
sağlamak için	1
N-mak için	163

TAB. 2.2 – Régularité au niveau grammatical en turc.

Nous pourrions multiplier les exemples de ce genre. Nous voulons seulement montrer que la connaissance du niveau morphémique est essentielle pour arriver à découvrir la structure formelle d'une langue. Les contextes que nous allons construire reposent essentiellement sur des éléments grammaticaux. Ils sont donc composés des mots grammaticaux de la langue, mais aussi (et pour certaines langues surtout) des affixes de celle-ci. Les deux types d'éléments sont toujours considérés comme des marqueurs de frontières des structures de la langue. Ceci est un point important de cette partie : considérer de façon identique les mots grammaticaux et les affixes de la langue. Ils structurent de manière similaire la langue, et ils appartiennent tous les deux à ce que nous appelleront plus tard des marqueurs de frontière. Qu'ils soient libres, c'est à dire qu'ils soient considérés comme un mot de la langue, ou qu'ils soient liés, c'est à dire qu'ils soient considérés comme des affixes de la langue, n'est dû qu'aux conventions d'écriture de

la langue. Structurellement, il n'existe, pour nous, aucune différence entre ces deux éléments, si ce n'est la façon de les obtenir (obtention directe pour les mots, segmentation pour les affixes).

Comme nous l'avons dit dans l'introduction de cette partie, ce travail de segmentation ne concerne pas toutes les langues. Dans une langue dite isolante comme le vietnamien[20], la segmentation ne fournit aucun résultat. Dans une écriture idéographique comme le chinois, notre "mot" est le signe (section 4.3), et notre algorithme est totalement inadapté pour ce genre de segmentation (le codage électronique des documents chinois ne reproduit pas le côté visuel des idéogrammes). Mais dans tous les autres types de langues, cette information morphologique est très précieuse. Nous verrons à la section 4.5 les diverses définitions du morphème proposées par certains linguistes, ainsi que la nôtre.

2.2 La segmentation

Les premiers essais pour trouver les éléments morphologiques, se sont inspirés de l'algorithme décrit dans [Harris, 1955]. Puis nous avons modifié cette procédure, en la divisant en trois parties, et en mettant à profit des caractéristiques de certains morphèmes. Le travail de segmentation se fait sur la liste des mots contenus dans les corpus.

2.2.1 L'algorithme de Harris

Le principe central de l'algorithme proposé par Harris se base sur le propos suivant :

> The basic procedure is to ask how many different phonemes (in various utterances) occur after the first n phonemes of some test utterances [Harris, 1955, page 192].

En adaptant cet énoncé à un corpus écrit, la méthode consiste à compter le nombre de lettres apparaissant après une séquence donnée de n lettres et qui correspond à une séquence de début (ou de fin) de mots. Soit M_n ce nombre. Puis on compare M_n avec celui obtenu avec la séquence composée de n+1 lettres : M_{n+1}. Si M_{n+1} est supérieur ou égal à M_n et que M_{n+1} est supérieur à M_{n+2}, alors nous arrivons à une frontière entre deux morphèmes. La figure 2.3 illustre le résultat pour les mots anglais *ungraspable* et *defirmity*. Après la séquence composée de la lettre *u* et commençant les mots de la liste du corpus anglais, le nombre de lettres différentes apparaissant est de 9. Après la séquence *un* ce nombre de lettres est de 21, etc,...

Cet algorithme se base sur l'observation suivante : plus nous parcourons un mot, plus les restrictions se font grandes sur les lettres pouvant apparaître. Ces restrictions portent surtout sur les séquences correspondant au parcours du radical des mots. Elles se relâchent quand nous arrivons à un endroit où une série d'affixes peuvent apparaître. Ces affixes provoquent alors une *augmentation* du nombre de lettres pouvant apparaître à cet endroit, ce que Harris appelle un *pic* dans la courbe des successeurs. Le résultat de cette segmentation est bien

[20] Il faut aussi tenir compte des conventions de segmentation des mots.

u	n	□	g	r	a	s	p	a	b	l	e
→ 9		21	1	1	1	1	1	1	1	1	

d	e	□	f	i	r	m	i	t	y
→ 9		19	5	3	4	4	3	1	

TAB. 2.3 – Principe de la version de base de l'algorithme de segmentation proposé par Harris. Une frontière est détectée après *un* et *de*.

sûr totalement dépendant des mots de la liste utilisée. Deux listes contenant le même mot peuvent générer deux segmentations différentes pour ce mot.

Si nous reprenons le tableau 2.3, nous voyons que le mot *ungraspable* est segmenté en *un-graspable*. La segmentation attendue par un linguiste (ou locuteur) serait *un-grasp-able*. Le morphème *-able* n'est pas détecté car la "famille" du mot dans la liste est pauvre. Pour palier cela, Harris propose une amélioration : l'algorithme est appliqué en partant des débuts de mots et aussi des fins de mots. Le résultat est illustré par le tableau 2.4. La segmentation est effectivement réalisée lorsque deux pics coïncident (ou en pratique un pic et un "plateau", c'est à dire une stabilisation de *n*, sinon seuls quelques dizaines de mots sont segmentés pour une liste en comprenant plusieurs milliers.).

	1	1	1	1	1	3	**19**	8	15	24	←
u	n	□	g	r	a	s	p	□ a	b	l	e
→ 9		**21**	1	1	1	1	1	1	1	1	

	1	4	2	2	9	**19**	17	25	←
d	e	□	f	o	r	m	□ i	t	y
→ 15		**26**	9	5	4	4	3	1	

TAB. 2.4 – Segmentation avec parcours dans les deux sens.

Le parcours en avant (*forward*) est efficace pour la découverte des préfixes, et le parcours en arrière (*backward*) l'est pour la découverte des suffixes. Bien que la combinaison de ces deux parcours offre une segmentation plus complète du mot, les contraintes font que très peu de mots sont alors segmentés (moins de 5%), et la liste des morphèmes trouvés est alors très faible (moins d'une dizaine), ce qui est insuffisant pour beaucoup de langues : les pics, dans la plupart des cas ne coïncident pas (tableau 2.5).

		9	9	6	5	1	1	←
	ç	a	l	a	c	a	k	
→		3	4	**12**	4	**15**	13	

TAB. 2.5 – Le mot turc *çalacak* n'est pas segmenté : aucun pic ne coïncide avec un autre. La segmentation aurait du être *çal-acak*.

De plus la segmentation générée peut être fausse comme le montre le tableau 2.6. Ainsi le parcours "en arrière" génère en turc une segmentation avant

la séquence finale *-ak*. Donc cette séquence *-ak* est considérée comme étant un morphème de la langue, alors que cette segmentation est due au fait que la séquence *-ak* finit plusieurs morphèmes du turc : *arak*, *acak*, et *mak*. Le mot turc *çalinacak* est segmenté en *çalınac-ak* alors que la segmentation correcte est *çalın-acak*.

		9		9		6		6		3		1		**1**		1		←	
	ç		a		l		ı		n		a		c	□	a		k		
→		1		1		4		**3**		**12**		4		**15**		13			

TAB. 2.6 – Erreur de segmentation avec parcours dans les deux sens.

Le même cas se produit en français pour la séquence *-on* qui est aussi identifiée comme morphème, alors qu'elle provient du morphème *ion*. Si nous nous servons de cette liste pour segmenter le reste des mots, nous obtenons alors une segmentation générale d'assez mauvaise qualité.

Les erreurs de l'algorithme Les erreurs de segmentation se produisent majoritairement aux "frontières" entre radicaux et affixes. Prenons un exemple extrait de notre corpus anglais : à partir de la liste de mots du tableau 2.7, l'algorithme génère une segmentation incorrecte. Cela est dû au fait que la liste comporte deux familles de radicaux, semblables à une lettre près : le *l* de *startl*. La segmentation génère donc deux mauvais morphèmes : *led* et *ling*.

start
start-*ed*
start-*ing*
start-*led*
start-*ling*

TAB. 2.7 – Premier type de mauvaise segmentation

Ici la dernière lettre du radical est inclue dans l'affixe : *l*. L'inverse peut se produire lorsqu'une famille de radicaux n'est pas assez riche en variations morphologiques. En particulier, le problème se pose quand une série de suffixes commencent par la même séquence de lettres. L'algorithme "rate" alors la frontière entre radical et affixes. Le tableau 2.8 illustre ce propos. Le parcours "en avant" segmente les mots comme *puissant* et *puissance* en *puissa-nt* et *puissance*.

S'il est vrai que ce type d'erreur peut être évité grâce à une segmentation "en arrière", cette segmentation va aussi générer de mauvais morphèmes et a l'inconvénient de segmenter trop peu les mots. Il va donc falloir trouver un algorithme qui permette une segmentation d'un assez grand nombre de mots, sans générer trop de morphèmes incorrects. Le principal reproche que nous adressons à la méthode proposée par Harris est le suivant : l'algorithme segmente un trop petit nombre de mots (10% des mots du corpus *français01*), les contraintes étant trop fortes. Nous aurions pu utiliser la liste des morphèmes identifiés pour segmenter

puissa-ment
puissa-mment
puissa-nce
puissa-nces
puissa-nt
puissa-nte
puissa-ntes
puissa-nts

TAB. 2.8 – Deuxième type de mauvaise segmentation

le reste des mots du corpus (ce que nous réalisons nous mêmes dans notre étape trois), mais nous avons alors préféré utiliser une autre approche décrite dans la section suivante (identification de morphèmes très sûrs et segmentation des mots grâce à ces morphèmes). Les différentes versions présentées dans [Harris, 1946] proposent des algorithmes qui produisent de meilleurs découpages des mots, grâce à l'ajout de contraintes. Mais plus l'algorithme devient complexe, plus le nombre de mots sur lesquels il peut travailler devient faible. De plus, la complexité des algorithmes devenant très grande, augmente très fortement le temps d'exécution. Il est nécessaire de diviser cette segmentation des mots en plusieurs étapes comme nous allons le voir dans la section suivante. Nous avons préféré à la solution de Harris, une méthode plus rapide (en temps d'exécution) qui ne cherche pas à obtenir une segmentation parfaite des mots, mais qui se contente qu'une segmentation relativement correcte.

2.2.2 La découverte des morphèmes

La méthode que nous avons appliquée pour la segmentation des mots diffère quelque peu. Nous nous sommes aperçu qu'il était plus efficace de ne pas considérer tous les éléments résultant de la segmentation sur le même plan. Certains affixes, grâce à leur grand effectif ou à certaines propriétés formelles, sont très faciles à trouver. Ces éléments sont appelés les *morphèmes prototypiques* de la langue. Une fois ces affixes trouvés, nous nous en servons pour segmenter les autres morphèmes de la langue. Puis, une fois la liste des morphèmes de la langue générée, nous prenons la liste des mots et les segmentons grâce à la liste des morphèmes. La segmentation des mots se déroule donc en trois étapes :

1. La découverte des morphèmes prototypiques
2. La découverte des morphèmes restants
3. La segmentation proprement dite de tous les mots du corpus.

Notre algorithme est centré sur la découverte des affixes de la langue. Les infixes n'ont pas été pris en compte (ils ont rarement un rôle relationnel). Nous divisons en deux la recherche des affixes : préfixes et suffixes. Nous allons illustrer nos propos par la recherche de suffixes. La recherche des préfixes est totalement symétrique : il suffit d'inverser l'ordre des lettres des mots (le résultat peut être observé sur le swahili). *L'établissement des divers seuils est fait de manière*

empirique sur une douzaine de langues. Un seuil est retenu lorsqu'il permet d'obtenir un résultat convenable pour les langues sélectionnées. L'on s'aperçoit que ces seuils sont plus sensibles à la taille du corpus qu'à la langue étudiée. Leur mise au point s'est effectuée sur des corpus d'environ 100000 mots[21]. Il convient d'ajuster (empiriquement) ces seuils lorsque la taille varie fortement.

Avec l'aide de ces algorithmes, et en supervisant les résultats, on peut obtenir en moins d'une heure une bonne connaissance de la morphologie d'une langue (en particulier une liste correcte des affixes de la langue et des quelques changements morphologiques de la langue liés à la concaténation de certains affixes entre eux). Ce partage des tâches entre ordinateur et humain nous semble le meilleur compromis sur le plan du temps de travail et de la qualité des résultats. Par la suite, nous n'utilisons que les résultats obtenus automatiquement, car ils sont suffisamment bons pour passer aux autres stades de la découverte des structures.

La découverte des morphèmes prototypiques La première phase concerne la recherche des affixes *prototypiques*. Ils ne correspondent pas à des affixes ayant un rôle particulier dans la structure, et ne sont pas identifiables *a priori* pour un locuteur, mais sont appelés ainsi parce qu'ils sont obtenus grâce à un algorithme qui génère des affixes avec un grand degré de confiance. Ces affixes sont obtenus de la façon suivante : nous commençons par construire la liste des mots du corpus. C'est avec cette liste de mots que nous allons travailler. Puis nous comptons, pour une séquence donnée de lettres, le nombre de lettres différentes qui peuvent la suivre, et pour chaque lettre, son nombre d'occurrences (figure 2.1). Si ce nombre de lettres différentes est supérieur à un certain seuil (neuf en pratique), nous sommes alors à la frontière d'un morphème. Ceci reprend l'idée générale de l'algorithme de Harris. Mais un cas particulier vient s'ajouter à ce traitement. Si une des lettres de la liste représente un grand pourcentage (40% en pratique) des occurrences des lettres, nous considérons alors que nous sommes à l'intérieur d'un morphème, et nous continuons le parcours des séquences sans segmenter à cet endroit (algorithme 1).

Dans la figure 2.1, la segmentation est évitée après la séquence g[22], bien que le nombre de lettres différentes soient suffisant (9), parce qu'une lettre n, représente 95% des occurrences possibles. Nous en déduisons que la séquence ng est la fin d'une séquence morphémique. Nous continuons donc le parcours en cherchant la frontière de ce morphème. Lorsque nous arrivons à *ing*, le nombre de lettres différentes étant suffisant, et aucune lettre ne représentant un pourcentage significatif, nous considérons que nous sommes arrivés à la limite du morphème, et nous ajoutons à la liste des morphèmes la séquence obtenue.

Nous ne travaillons que sur des morphèmes occurrant plus de 20 fois (le test ($M >$ SEUIL) dans l'algorithme). Il arrive en effet que certaines séquences dont l'effectif est très faible (ici moins de 20 occurrences sur une liste généralement composée de plus de 5000 mots) soient identifiées comme morphème selon nos critères. Le fait de fixer un seuil minimal à l'effectif d'un morphème permet

[21] Ce qui génére une liste d'environ une dizaine de milliers de mots.

[22] L'algorithme de Harris fournit souvent comme affixes les premières et dernières lettres des mots, donc la plupart des lettres de l'alphabet utilisé.

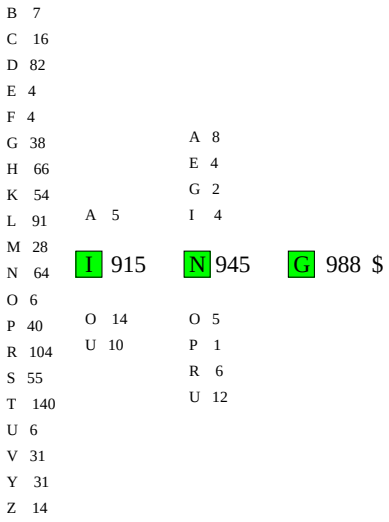


FIG. 2.1 – Recherche des affixes caractéristiques à partir d'une liste de mots extraits d'un corpus. Les nombres après les lettres correspondent à leur nombre d'occurrences.

Algorithme 1 Découverte des morphèmes prototypiques

pré-requis S : une séquence de lettres finissant les mots.
  Soit M le nombre de mots finissant par S.
  Soit L l'ensemble des lettres occurrant avant S.
  Soit n le cardinal de L
  Soit l_i le nombre d'occurrences de la lettre l, $l \in L$.
  si ($M >$ SEUIL) **alors**
    pour tout l $\in L$ **faire**
      si ($l_\mathrm{i} > 0.4{*}M$) **alors**
        on continue le parcours avec la séquence $l_i{+}M$.
      sinon si $n > MAX$ **alors**
        la séquence S est un morphème.
      fin si
    fin pour
  fin si

d'augmenter le degré de confiance dans des morphèmes obtenus. De plus, si un bon morphème est éliminé par ce critère, les conséquences sont limitées puisque son effectif est très faible.

Nous avons introduit une heuristique dans notre recherche des morphèmes. Nous identifions un morphème par le fait qu'une lettre représente plus de 40% des lettres possibles après une séquence (figure 1). Ce seuil n'est pas toujours respecté. Nous prenons en compte les morphèmes légèrement moins fréquents (morphèmes dont la fréquence est supérieure à 20%) si la somme de ces derniers morphèmes est supérieure à un certain seuil (60% des séquences). Dans ce cas, nous considérons que nous parcourons une séquence qui correspond simultanément à plusieurs morphèmes. Pourquoi ajouter cette modification ? Cette heuristique est surtout intéressante pour des familles des morphèmes qui partagent une fin (pour les suffixes) ou un début (pour les préfixes) similaire. Ce cas se produit fréquemment dans une langue comme le turc ou les voyelles de certains affixes dépendent du radical[23] comme pour *-mak* et *-mek*. Il y a une répartition entre les deux voyelles *e* et *a*, ce qui fait que la valeur l_{i} pour chaque élément est inférieure à $0.4{*}M$ (algorithme 1). La valeur de l_{ak} est de 38% et celle de l_{ek} est de 35%, donc tous les deux en dessous du seuil établi de 40%. Si l'on ajoute le score des deux morphèmes, on obtient un score 73%. Cette opération peut sembler ad hoc au turc, mais cette situation peut se rencontrer aussi dans les autres langues, et la modification est généralement bénéfique. Le cas est illustré par le tableau 2.9. Sans cette heuristique, la séquence finale *che* serait identifiée comme morphème de la langue. Avec celle-ci, la séquence *iche* est considérée comme morphème potentiel et le parcours continue pour trouver la séquence *-iche* comme morphème. La séquence *sche* n'aboutit à aucun morphème. Cette heuristique n'apporte pas de grandes modifications à la liste des morphèmes prototypiques, mais elle améliore légèrement celle-ci pour plusieurs langues.

La liste des morphèmes prototypiques est plus ou moins longue selon les langues. La liste française est composée de 101 éléments (70 suffixes et 31 préfixes), la liste allemande de 27 éléments (11 suffixes et 16 préfixes) morphèmes, contre 65 (54+11) pour le turc et 54 (17+37) pour le swahili. Cette longueur dépend de la langue (de sa morphologie), mais aussi de la taille du corpus. Dans les langues où les préfixes ne jouent pas de rôle relationnel (français, turc), la découverte des préfixes prototypiques est très mauvaise (parfois plus de de 75% d'erreur). Mais puisque ces affixes n'interviennent pas dans la construction des relations, ce bruit ne génère aucune gêne pour la suite du travail, en particulier dans la génération des couples morphologiques (l'environnement de ces préfixes ne possède aucune régularité formelle). De plus, le fait d'avoir segmenté de mauvais éléments est identifiable grâce à l'opération suivante : la découverte des morphèmes restants qui ne produit alors aucun résultat significatif (aucun autre morphème n'est découvert). *Puisque notre méthode (décrite au chapitre 6) permet de sélectionner les bonnes séquences morphologiques des mauvaises, la segmentation en affixes des mots est systématique.*

[23] Le phénomène d'harmonie vocalique.

Séquence	Effectif
ache	12
eche	4
iche	29 (33.7%)
lche	3
oche	3
rche	2
sche	23 (26.7%)
uche	8
äche	1
üche	1
total	86

TAB. 2.9 – Parcours de plusieurs morphèmes. La séquence *che* peut correspondre à plusieurs morphèmes (ici un morphème (*-iche* et la séquence *sche*), d'où une répartition entre les lettres précédentes possibles (*i* et *s*).

La découverte des morphèmes restants Une fois la liste de ces morphèmes obtenue, il nous reste à compléter celle-ci par la méthode suivante : nous parcourons les mots du textes, et pour une séquence donnée (*consider* dans le tableau 2.10) nous regardons si les séquences restantes (*able*, *ably*, *ation*, *ed*, *ing*) correspondent à des morphèmes déjà trouvés. Si la moitié des éléments correspondent, nous considérons que les éléments restants (*able*, *ably*) correspondent aussi à des morphèmes.

Morphèmes trouvés	Mots	Nouveaux morphèmes
	considerable	able
	considerably	ably
-ation	consideration	
-ed	considered	
-ing	considering	

TAB. 2.10 – Recherche de nouveaux morphèmes

Seuls les nouveaux éléments apparaissant plus de quatre fois sont conservés. Cela évite d'inclure dans cette liste des morphèmes incorrects comme *-son* dans le tableau 2.11. De tels morphèmes étant souvent liés à la collision entre deux familles de radicaux, leurs effectifs sont très faibles, ce qui explique le seuil assez bas permettant leur élimination.

L'application de cet algorithme fournit une nouvelle liste de morphèmes. Ils sont ajoutés à la liste des morphèmes prototypiques, et l'algorithme est une nouvelle fois appliqué avec ces nouveaux morphèmes. Ceci jusqu'à ne plus obtenir de nouveaux morphèmes. La plupart des langues se stabilisent après une demi douzaine de tours.

Algorithme 2 Découverte des suffixes restants

pré-requis S : une séquence de lettres commençant des mots.
  Soit M l'ensemble des séquences constituée des fins de mots.
  Soit m le cardinal de cet ensemble
  Soit MC le nombre de morphèmes appartenant à M
  si $MC > 0.5*$ M **alors**
    Les éléments de M sont ajoutés à une liste L.
  sinon
    On continue le parcours de S en ajoutant les lettres suivantes.
  fin si
  On enlève de la liste L les éléments ayant un effectif inférieur à 5.

Morphèmes trouvés	Mots	Nouveau morphème
-ie	garnie	
-er	garnier	
-es	garnies	
	garnison	-son
-ture	garniture	

TAB. 2.11 – Erreur dans la segmentation : la séquence *-son* est considérée comme un morphème français.

2.2.3 La segmentation des mots

Une fois la liste des morphèmes générée, il suffit pour segmenter tous les mots du corpus, de rechercher quels sont les morphèmes les plus longs qui correspondent au début et à la fin des mots (algorithme 3).

Algorithme 3 Segmentation des mots

pré-requis M : La liste des morphèmes
pré-requis *Mots* : La liste des mots
  pour tout m_{i} dans *Mots* **faire**
    $D \leftarrow$ rechercher le plus long morphème matchant le début du mot
    $F \leftarrow$ rechercher le plus long morphème matchant la fin du mot
    décomposer le mot m_{i} en $D+$ R $+ F$
  fin pour

Tous les mots ne sont pas segmentés. Il existe généralement dans la liste des morphèmes, des éléments composés d'une seule lettre. Ces éléments ont un "pouvoir de segmentation" très grand : ils peuvent segmenter beaucoup de mots, y compris les mots grammaticaux. Nous verrons plus tard (section 3.2) pourquoi il n'est pas souhaitable de segmenter ces mots là. Pour éviter cela, les mots fréquents (une caractéristique de beaucoup de mots grammaticaux) ne sont pas segmentés. Le seuil est fixé pour ne pas segmenter les cinq premiers pourcents des mots les plus fréquents. Dans cette liste, sont compris certains mots grammaticaux mais pas tous, et certains mots lexicaux. Ces derniers ne

sont donc pas segmentés. Le reste des mots est segmenté, et permet la génération des séquences morphologiques (chapitre 3).

2.3 Analyse des résultats

Nous donnons en annexeC les listes de morphèmes obtenus pour différentes langues. Les éléments obtenus correspondent aux suffixes de la langue dans plus de 90% des cas (tableau 2.12). Le résultat de la segmentation sur les préfixes dépend beaucoup plus des langues. Si la segmentation est très bonne pour une langue comme le swahili où les préfixes jouent un rôle fonctionnel, elle est relativement muavaise pour les langues où les préfixes ne jouent aucun rôle fonctionnel, comme le français où l'anglais. Les séquences correspondant à des morphèmes incorrects sont écrites en italique dans les annexes.

Langues	Suffixes corrects	Préfixes corrects
français	92%	49%
anglais	98%	19.5%
allemand	97%	62.5%

TAB. 2.12 – Évaluation de la liste des préfixes et des suffixes.

Langues	Segmentation correcte
français	94,8%
anglais	96%
allemand	93%

TAB. 2.13 – Évaluation manuelle de la segmentation des mots (seuls les suffixes sont pris en compte).

Les estimations du tableau 2.13 ont été réalisées sur 1000 mots de la liste pris au hasard. Une segmentation est jugée bonne si elle identifie même partiellement un affixe du mot. Il était parfois très difficile de juger de la justesse d'un morphème. Les langues utilisées pour cette estimation sont le français, l'anglais, l'allemand, langues où nous pouvions aisément vérifier la segmentation. Une évaluation plus systématique a été faite pour l'anglais, en comparant les résultats de notre segmentation avec les résultats de l'analyseur morphologique *PC-KIMMO* [Antworth, 1990]. Voici le protocole d'évaluation :

- Les mots du corpus anglais sont segmentés avec PC-KIMMO.
- Seuls les mots admettant une seule segmentation sont retenus (PC-KIMMO n'assure pas une segmentation bonne à 100%[24]).
- les mots segmentés par PC-KIMMO sont comparés à notre segmentation.

La comparaison n'est pas immédiate puisque PC-KIMMO donne parfois des résultats irréguliers[25]. Ainsi la décomposition de *seriouly* est *serious+ly*, mais

[24] par exemple *parisian* donne *pare+ise+ian*.

[25] de notre point de vue.

celle de *vigorously* est *vigor+ous+ly*[26]. Dans notre segmentation, l'algorithme segmentera tous les mots finissant par *ously* de la même manière (sauf les plus fréquents, qui eux ne seront pas segmentés).

Type d'erreurs	Exemples			Taux
	Mot	PC-kimmo	Notre	
Morphèmes marquants	perceptible	ible	e	1,5%
Partie de morphèmes	genial	ial	al	6%
Morphèmes trop grands	seriously	ly	ously	10%
Mots non segmentés	that's	's	that's	6.5%
Correspondance stricte	stability	ity	ity	76%

TAB. 2.14 – Comparaison entre notre segmenteur et PC-KIMMO

-ent -ant -ish -ite -ible

TAB. 2.15 – Liste des morphèmes manquants en anglais : ils concernent 1% des mots du corpus

Le tableau 2.14 illustre les différents cas de figure rencontrés. La segmentation réalisée manuellement considère les points "morphèmes trop grands" et "partie de morphèmes" comme correcte. On retrouve alors une estimation similaire à celle du tableau 2.13 (92.5% contre 95%). Le fait que les erreurs "morphèmes trop grands" soit plus grand que les erreurs "partie de morphèmes" s'explique logiquement par l'algorithme de plus long matching utilisé. Les résultats obtenus avec PC-KIMMO nous montre qu'il est très difficile d'une part de décomposer les mots en morphèmes (PC-KIMMO offre plusieurs solutions généralement et parfois de fausses segmentation), d'autre part que l'évaluation d'une telle opération est très délicate et nécessite des connaissances étymologiques sur la langue. En particulier, la segmentation des préfixes dans les langues où ils ne jouent aucun rôle relationnel s'est révélée très délicate, c'est pourquoi ils n'ont pas été pris en considération dans l'estimation de la segmentation. En fait, notre critère de validité de la segmentation est tout autre. *Pour considérer une segmentation correcte, il suffit que cette dernière permette une découverte des structures de la langue, dans les étapes ultérieurs.* L'évaluation ne se fait donc pas au niveau du mot mais sur les résultats obtenus par la suite, le but de ce travail étant la construction des séquences morphologiques de la langue.

Les différents types de morphèmes obtenus La définition du mot étant une séquence de lettres comprise entre une ponctuation ou un blanc, des éléments un peu atypiques sont rencontrés dans cette liste. Ainsi on trouve des séquences telles que *-a-t-il*, *s'*, *n'*, *d'* dans la liste des affixes en français et *n't* et *'s* dans la liste anglaise. De même que *'in*, *'ten*, *'dan* en turc, qui correspondent à la désinence utilisée pour les nom propres (*Mesih'in*, *Apolonya'dan*, *Milet'ten*).

[26] L'adjectif *vigor* existe mais pas *sery* ou *seri*

Cette segmentation peut aussi fournir des résultats intéressants sur la segmentation des mots composés. Ce cas arrive assez souvent en allemand et en anglais. L'identification de ces éléments peut se faire en vérifiant s'ils existent

jung	
junger	-er
jungen	-en
jungfrau	-frau
jungfrauen	-frauen

TAB. 2.16 – Segmentation des mots composés.

dans la liste des mots (comme frau et frauen, si on ne tient pas compte de la majuscule initiale des substantifs), et permet ainsi de les différencier des morphèmes incorrects.

Les éléments obtenus peuvent être soit des morphèmes soit des séquences composées d'une suite de morphèmes. Si nous observons la liste de morphèmes turcs, nous voyons que beaucoup des séquences obtenues sont composées de séquences correspondant elles aussi à des morphèmes. En fait ce phénomène apparaît dans toutes les langues, même celles considérées morphologiquement pauvres comme l'anglais. Nous n'avons pas cherché à resegmenter ces séquences morphologiques, puisqu'elles ont un comportement distributionnel identique au morphème "principal" de la séquence, c'est à dire celui qui joue un rôle relationnel (le dernier généralement).

Langues	Séquences	Décomposition
Français	ances	ance-s
	ionelle	ion -elle
	ation	at-ion (?)
Turc	mektir	mek-tir
	lerinden	ler-in-den
	malarini	ma-lar-in-i

TAB. 2.17 – Exemple de séquences composées de plusieurs morphèmes unitaires.

Ainsi la séquence française *-ances* va apparaître (majoritairement) dans les mêmes contextes que le morphème *-s*, c'est à dire un syntagme nominal pluriel. De même pour *ation* et *ion*. Comme ces séquences ne gênent pas particulièrement la découverte des structures, nous ne cherchons pas à les segmenter en une séquence de morphèmes unitaires. Mais si cette resegmentation était nécessaire, la règle 2.18 peut être utilisée pour segmenter automatiquement la liste de morphèmes.

Par exemple, la séquence *ionelle* est segmentée en *ion+elle*, puisque ces deux morphèmes appartiennent à la liste. Les morphèmes d'une lettre peuvent poser quelques problèmes, et il est préférable, dans un premier temps de ne pas les prendre en compte. Cette opération ne s'applique pas à toutes les séquence

Si A et B appartiennent à la liste des morphèmes et qu'un morphème C soit composé des séquences A+B, alors décomposer le morphème C en A+B.

TAB. 2.18 – Règle de segmentation des séquences de morphèmes.

de morphèmes, dans ce cas où l'agglutination des deux morphèmes s'accompagne d'un changement de forme. Ainsi en turc, le morphème *acak*, lorsqu'il est suivi d'un morphème commençant par une voyelle, devient *acağ* (-*acağım*, -*acağını*). De telles transformations se détectent assez vite manuellement, puisqu'elles sont très régulières. Parfois, elles permettent même de retrouver les distinctions entre voyelles et consommes (comme en turc où l'harmonie vocalique génère des contraintes fortes sur les voyelles des morphèmes).

Analyse des erreurs Chacune des trois étapes de la segmentation génère des erreurs spécifiques. La première étape peut générer une liste de morphèmes dont certains sont incorrects. Par exemple la liste française contient deux de ces morphèmes : -*che* et -*resse*. Mais ces morphèmes incorrects n'ont souvent aucune incidence sur la deuxième phase. En effet cette phase nécessite l'utilisation de plusieurs morphèmes pour générer de nouveaux éléments. Pour que cette phase génère de mauvais éléments, il faudrait que tous les morphèmes utilisés soit incorrects, cas qui ne se produit jamais. Ainsi ces deux morphèmes français ne provoquent aucune génération de morphèmes incorrects. Il en est de même pour les autres langues.

La deuxième étape peut fournir aussi de mauvais morphèmes (tableau 2.11), mais ces erreurs sont assez faibles. Elles concernent surtout les morphèmes composés d'une lettre.

La troisième étape est celle qui génère le plus d'erreurs, puisqu'elle porte sur l'ensemble des mots du corpus. Ceci est dû au fait que l'algorithme utilisé est assez "rudimentaire" (algorithme 3). Le tableau 2.19 donne quelques exemples d'erreurs. Il est parfois très délicat de juger de la justesse d'une segmentation. Une étude étymologique du mot peut parfois être nécessaire.

Mot	morphème	segmentation
Mantoue	-ue	Manto+ue
indique	-ique	ind+ique
d'arrêt	-t	d'arrê+t
réciproques	-ues	réciproq+ues
esprit	-it	espr+it
continûment	-ent	continûm-ent
reçoive	-ive	reço-ive
bassin	-in	bass+in (?)
hideux	eux	hid+eux (?)
propos	pro-	pro-pos

TAB. 2.19 – Erreur de segmentation de la troisième étape

Certaines erreurs sont dues à un morphème manquant (par exemple, il manque le morphème *-ment* pour segmenter *continûment* correctement, la liste ne contenant que *-ement*). Beaucoup d'erreurs de segmentation (20% des erreurs en français) sont dues à des mots étrangers de la langue (noms propres pour la plupart). Ces mots peuvent aussi générer des morphèmes (ainsi on trouve *-ing* dans la liste des mots français, et *-ath*, *-oth* dans la liste des morphèmes latins alors qu'ils correspondent aux terminaisons de noms propres hébraïques). Les morphèmes d'une lettre peuvent aussi conduire à de mauvaises segmentations. Nous avons considéré qu'un mot finissant par un *e* muet en français ne correspondait pas à une erreur, ce qui, à l'écrit, est parfaitement justifiable.

Le faible taux d'erreur peut surprendre, mais il est dû au principe de segmentation. Nous considérons le plus long morphème pouvant segmenter un mot donné. Et ces longs morphèmes sont souvent corrects et correspondent généralement bien à un affixe de la langue. Pour corriger ces erreurs, il faudrait tenir compte des familles de radicaux (section 5.2). Le résultat de la segmentation étant jugé suffisamment bon, cette amélioration n'a pas été prise en compte.

2.4 La segmentation de textes phonétisés

Durant notre mise au point de la méthode de segmentation, nous avons remarqué que certains résultats (en fait certaines erreurs) étaient spécifiques à l'écrit, et qu'ils ne se produiraient pas si nous traitions une forme phonétique des mots. Par exemple, une segmentation comme *réciproq-ues* est impossible puisque la transcription phonétique est */resiprok/*. Nous avons phonétisé[27] les mots de notre corpus en français, et segmenté ces nouveaux mots. Le résultat a été sans surprise. Là où on obtenait plusieurs morphèmes à l'écrit, la forme phonétisée n'en génère qu'un (*ance*, *ence*). À l'inverse, là où la forme écrite n'avait qu'un seul morphème, la forme phonétisée peut en générer plusieurs (cas du *s* anglais qui donne trois phonèmes différents : /s/ /z/ /iz/). Il n'est apparu aucun comportement fondamentalement différent entre forme écrite et phonétique. Ceci n'a rien de bien surprenant puisque la forme phonétique a été générée automatiquement à partir de la forme écrite. En fait, nous n'avons fait qu'utiliser un autre alphabet pour écrire le texte. Travailler sur un corpus phonétique n'est donc pas bien différent du travail sur un texte écrit.

2.5 La segmentation à partir des entre-ponctuations

Si on reprend les articles de Harris, on s'aperçoit que ses premiers exemples de segmentation portent non pas sur les mots mais sur des suites de mots. [Chatman, 1955] proposait de réaliser une telle opération en utilisant l'algorithme de Harris. Nous avons alors tenté de faire de même en éliminant les blancs des corpus, et en prenant la ponctuation comme séparateurs de ces nouveaux segments, les *entre-ponctuations*. Le résultat a été décevant. Une grande partie des "morphèmes" ainsi trouvés correspondent aux morphèmes les plus fréquents

[27] En utilisant le phonétiseur décrit dans [Morel and Lacheret-Dujour, 1998].

trouvés au niveau des mots. Nous trouvons aussi quelques mots grammaticaux fréquents de la langue (prépositions, déterminants, adverbes de phrase). Les autres types de segmentations sont rares et très disparates.

Nous espérions que ce type de segmentation donnerait des éléments de niveaux intermédiaires entre le morphème et la séquence de morphème. Il n'en est rien. Nous accédons directement au niveau morphologique de la langue, le plus bas, sans passer par des structures différentes. Il est donc plus profitable d'utiliser le mot comme point de départ. Cette expérience cependant montre que les morphèmes d'une langue sont accessibles même si le système d'écriture n'admet pas de segmentation en mots. Ainsi notre algorithme de segmentation appliqué au japonais permet la génération des séquences morphologiques.

2.6 Les travaux similaires

On pourra trouver d'autres algorithmes effectuant la même tâche de segmentation des mots. L'algorithme présenté dans [Brent et al., 1995] se base sur le concept de *longueur de description minimale* (MDL : miminal description length). [Kazakov, 1997] utilise une solution mixte en intégrant la notion de MDL dans un algorithme génétique. Le principe est de minimiser le nombre de lettres N contenu dans le lexique. Le lexique est divisé en deux parties : un lexique contenant les radicaux, et un lexique contenant les affixes (en fait seulement les préfixes dans les cas traités). Comme la liste des mots peut être stockée par une liste de couples <radical-préfixes>, la minimisation de N permet un stockage, ou une description minimale de la liste des mots. Les données présentées concernent uniquement l'anglais et le français. [Brent et al., 1995] intègrent la catégorie des mots afin d'améliorer les résultats (un mot finissant par *-ed* est plus souvent un verbe qu'un substantif). La liste des morphèmes est moins fournie que la nôtre.

[de Marcken, 1995] présente un travail concernant la segmentation d'énoncés. Son travail porte, à l'origine, sur l'acquisition de lexique à partir de chaînes sonores. En pratique, il utilise des textes (phonétisés ou non) où la séparation entre mots a été enlevée. Le principe algorithmique est toujours le MDL. Les éléments trouvés sont surtout les morphèmes de la langue (le travail porte uniquement sur l'anglais.).

On trouvera dans [Brent and Cartwright, 1996] un travail similaire (segmentation de textes), mais il introduit la notion de marqueurs de frontière de mots. Il note en effet que certaines séquences, pour une langue donnée, ne peuvent apparaître en début ou en fin de mots. L'anglais n'admettant pas qu'un mot commence par *gd*, le mot *gdog* ne peut résulter de la segmentation de la séquence *thebigdog*. Mais il ne dit pas comment obtenir ces éléments (il se pose en fait cette question). Cette idée de travailler sur les marqueurs de frontières (possible ou non) nous semble très intéressante, puisque c'est sur cette notion que se base tout notre travail. Nous reviendrons sur les propriétés de début ou fin de mots dans la section 4.6.

[Hutchens and Alder, 1998] propose aussi une méthode pour segmenter un texte en *chunks*, en se basant sur l'entropie d'une séquence. Le résultat, là aussi,

n'est pas surprenant : les chunks les plus fréquents correspondent aux mots grammaticaux de la langue (l'anglais en l'occurrence).

[Wolff, 1977] propose un système aussi simple qu'original de découverte des segments. Son corpus est constitué de textes où la segmentation entre mots a été éliminée. Ses unités de départ sont la lettre. Puis il calcule les couples d'unités contiguës, et fusionne en une nouvelle unité le couple le plus fréquent. Ceci correspond à un passage (*scan*). Il réitère ceci un certain nombre de fois (environ 500), et obtient une segmentation du texte qui est finalement assez bonne :

((IT)(IS))(SUMMER)(TIME)(SCHOOL)(IS)(OVER)(AND)(THE)
((LONG)(SUMMER))(HOLIDAY)(IS)(HERE)(JANE)((AND)PETER)
(**T**)(**AL**)(**K**)(ABOUT)(THEIR)((LONG)(SUMMER))(HOLIDAY)
(AND)(WHAT)(THEY)(ARE)(GOING)(TO)(DO) ...

Les premiers éléments à apparaître sont les éléments grammaticaux de la langue (affixes et mots grammaticaux). La segmentation n'est bien sûr pas parfaite (par exemple la segmentation de **TALK**) mais ce travail montre qu'une segmentation assez correcte en mots peut être réalisée avec assez peu de moyens.

Nous reviendrons sur le travail de Gerry Wolff dans le chapitre sur la découverte des structures, puisqu'il propose aussi un système générant de telles structures.

Les autres approches utilisent des techniques probabilistes. Un modèle $n$grammes est utilisé par [Stolcke and Shriberg, 1996] afin d'apprendre la détection de limites des *segmentations linguistiques*, en particulier les fins de phrases. Un échantillon d'apprentissage (contenant une segmentation manuelle) est nécessaire.

Quelle que soit la méthode utilisée, il semble difficile de ne pas produire de bons résultats. En effet, il existe toujours dans la langue, une série de morphèmes très fréquents, qu'il est difficile de ne pas trouver. Ces éléments peuvent alors servir d'amorce à la segmentation.

Chapitre 3

Les séquences morphologiques

Sommaire

Dans ce chapitre, nous allons montrer l'importance des séquences morphologiques des langues dans un travail de découverte des structures formelles, mais aussi en quoi leur génération ne peut suffire dans un processus de découverte des structures linguistiques. L'idée à l'origine de ce travail était que la découverte de ces séquences morphologiques rendrait possible la catégorisation des éléments des langues. Nous verrons à la section 3.3.1 qu'il n'en est rien. Mais la suite (chapitre 6) nous montrera que ce travail n'a pas été inutile et que les éléments construits dans cette partie serviront de point de départ à la "vraie" découverte des structures.

3.1 La schtroumpfance des séquences schtroumpfologiques

Pour illustrer l'importance des séquences morphologiques d'une langue dans un processus de découverte, nous trouvons dans la littérature un certain nombre d'exemples. Le premier est le poème du Jabberwocky de [Carroll, 1994] que [Fries, 1952, page 70] donne en exemple afin d'illustrer l'importance structurelle des éléments morphologiques :

> **Twas** brillig **and the** slith**y** tove**s**
> **Did** gyre **and** gimble **in the** wabe ;
> **All** mims**y were the** borogove**s**,

> **And the** mome raths outgrabe
> Somehow [Alice said], it seems to fill my head with ideas
> -only I don't know exactly what they are !

En voici une version allemande :

> **Es** sunn**te** Gold, **und** Molch **und** Lurch
> krawallten 'rum **im** grün**en** Kreis,
> **den** Flattrings **ging es durch und durch**,
> **sie** quieps**ten wie die** Quiekedeis.

D'autres versions sont consultables à l'adresse suivante :

> `http://www.pair.com/keithlim/jabberwocky/`.

Ce poème est construit en utilisant comme ossature structurelle des morphèmes et mots grammaticaux de la langue (anglais, allemand,...), et en inventant certains éléments lexicaux. Comme le remarque Alice, le texte semble familier à un locuteur de ces langues, mais il est difficile de préciser davantage le sens du poème (Pour les curieux, une explication est donnée par notre ami Humpty Dumpty [Carroll, 1994, pp. 102-104]). Comme le note Fries,

> If we assumes that these utterances are using the structural signals of English, then at once we know a great deal about these sequences. [Fries, 1952, page 71]

Le deuxième exemple, plus surprenant, se trouve dans la bande dessinée des Schtroumpfs [Peyo, 1959]. Ces petits bonshommes parlent une langue où certains éléments (les radicaux) sont remplacés par la séquence *schtroumpf*, ou l'équivalent pour les autres langues que le français (*smurf* dans les pays anglo-saxons.). Cela donne des phrases comme :

- Inspiration hasn't *smurf*ed yet.
- Lazy smurf have You *smurf*ed that play for our village fair ?
- *Smurff*atje, heb jij de Brilsmurf en de Loismurf soms gezien ?

FIG. 3.1 – La langue des schtroumpfs (hollandais et anglais).

Les textes sont plus compréhensibles que le Jabberwocky, puisque seulement quelques radicaux sont remplacés. Nous pouvons nous aussi facilement générer des textes de ce style, en remplaçant les radicaux des mots segmentés de notre corpus par un élément quelconque (prenons schtroumpf). À partir de la phrase :

> *Les erreurs des spécialistes de la planification urbaine au cours des dernières décennies ont été nombreuses.*

l'opération de segmentation génère la phrase suivante :

> *Les err-eurs des spéci-alistes de la planifi-cation urbaine au cours des dernières dé-cenn-ies ont été nombr-euses.*

En remplaçant les radicaux des mots segmentés par *schtroumpf*, nous obtenons finalement la phrase :

> *Les schtroumpfeurs des schtroumpfalistes de la schtroumpfication urbaine au cours des dernières schtroumpfies ont été schtroumpfeuses.*

Mise à part le côté ludique, il est important de constater que tous les mots de cette phrase peuvent être catégorisés par un locuteur français en nom, adjectif, verbe, préposition ou déterminant. Cette catégorisation est rendue possible grâce à la présence des éléments morphologiques de la langue. Ils joueront donc un rôle important dans le processus de catégorisation. Ces éléments sont composés des mots grammaticaux, mais aussi des affixes de la langue. Ces deux types d'éléments forment le squelette structurel de la langue. On remarque que les affixes grammaticaux suivent aussi la loi de Zipf, du moins pour les éléments les plus fréquents. Ainsi, un petit nombre de ces affixes vont être très fréquents. Les éléments qui ont un rôle fonctionnel dans la structure (les marques casuelles et les affixes verbaux par exemple) en font généralement partie.

3.2 Les couples morphologiques

Une fois l'importance des éléments grammaticaux notée, nous allons voir comment les utiliser. La génération des séquences morphologiques de la langue se fait de manière très simple. L'algorithme est le suivant :

Algorithme 4 Génération des couples morphologiques

pré-requis C : un corpus segmenté en morphèmes
  pour tout couple de mots contigus m_i et m_{i+1} de C **faire**
    pour tout morphème mf_k de m_i **faire**
      pour tout morphème mf_l de m_{i+1} **faire**
        incrémenter l'effectif du couple (mf_k, mf_l)
      fin pour
    fin pour
  fin pour

Pour tous les couples de mots d'un corpus dont les mots ont été segmentés par la méthode décrite à la section 2.2.3, on forme tous les couples morphologiques possibles. Par exemple, à partir du couple de mots segmentés (*in-formation, judici-aire*), les couples suivants sont formés :

information	judiciaire
in-	judiciaire
in-	judici-
in-	-aire
-ation	judiciaire
-ation	judici-
-ation	-aire
form-	judici-
form-	-aire
form-	judiciaire

Les couples résultants peuvent être composés de deux mots, de deux morphèmes, ou d'une combinaison d'un morphème et d'un mot. Le tableau 3.1 montre les couples les plus fréquents obtenus en allemand. La segmentation obtenue au chapitre 2 est suffisamment correcte pour générer les couples intéressants. De manière similaire, les séquences de trois, quatre éléments peuvent être générées, mais nous verrons qu'elles sont inutiles dans la démarche finale. Les séquences morphologiques utilisées sont donc des séquences composées de deux éléments : les couples morphologiques.

Couple	Effectif
zu N-en	645
N-en und	387
N-en N-en	372
die N-e	369
den N-en	302
daß ich	278
und N-en	236
wir N-en	220
die N-en	219

TAB. 3.1 – Les couples morphologiques les plus fréquents en allemand.

Ces structures composées de séquences de mots grammaticaux et d'affixes sont donc assez faciles à construire. Il est à noter que les mots apparaissant dans ces couples sont pour une grande majorité les mots grammaticaux de la langue.

Comme la segmentation des mots ne produit pas une liste parfaite de morphèmes, la liste des couples morphologiques contient nécessairement des couples non pertinents du point de vue Notre processus de catégorisation expliqué au chapitre 6 nous montrera comment ces couples sont utilisées. **En résumé, la segmentation génére des morphèmes, bons et mauvais, qui nous servent à construire des séquences morphologiques, et notre processus de génération des structures linguistiques utilise des philtres (positionnels) permettant une sélection des séquences intéressantes qui permettent la catégorisation des mots et morphèmes du corpus.**

3.3 Les limites intrinsèques du critère morphologique

3.3.1 Les problèmes de catégorisation

Nous allons maintenant nous intéresser aux problèmes rencontrés dans la suite de notre travail. Suivant les préceptes développés dans [Harris, 1951], nous avons essayé de catégoriser les éléments de la langue grâce à des contextes distributionnels. Pour mener cette tâche à bien, les morphèmes, pensions-nous, allaient nous offrir des contextes beaucoup plus adéquats que les mots. En fait, les difficultés décrites dans la section 1.3 s'appliquent aussi bien aux contextes composés de mots qu'aux contextes composés de morphèmes. S'il est vrai que les contextes morphologiques font apparaître des régularités très intéressantes, et que les morphèmes offrent un meilleur élément de base à leur construction, nous ignorons toujours quels contextes retenir dans la liste des contextes possibles (tableau 3.2). Retour au point de départ ! Le problème de définition du contexte est toujours présent.

Séquence	Effectif
N-e [N-e] de	636
N-e [N-ion] de	96
N-e [N-s] de	73
N-e [N-ent] de	25

Séquence	Effectif
les N-s [N-s]	1391
les N-s [N-ent]	253
les N-s [N-e]	99
les N-s [N-aux]	55

TAB. 3.2 – Les contextes, même morphologiques, n'offrent pas de contraintes suffisantes pour permettre une catégorisation. Comment savoir que le contexte *N-e [ ] de* est inadapté pour le français. Ou que la séquence *les N-s* n'offre pas suffisamment de contraintes pour catégoriser les séquences suivantes (adjectifs ou verbes) ?

Les tableaux 3.2 nous montrent bien que l'effectif seule n'est pas un critère suffisant pour discriminer les bons des mauvais contextes (à supposer que l'on puisse définir a priori un bon contexte d'un mauvais). Nous verrons que le critère de validité d'un contexte ne peut se faire sans recours à la structure des langues (section 3.4) et que le fait de ne pas savoir quelles catégories construire est vraiment un frein au développement de la méthode. Bien sûr, il est vrai que, dans certaines langues, les régularités morphologiques sont telles qu'elles offrent un guide très efficace dans la découverte manuelle des structures. Par exemple, il est difficile de ne pas remarquer la structure française *les N-s*. Mais il en était déjà de même au niveau du mot, et les exemples de la section 1.3 peuvent être identiquement repris dans cette section.

3.3.2 Un essai de catégorisation avec les structures d'accord

Armé de nos séquences morphologiques, nous avons essayé de mettre au point un algorithme de catégorisation. Durant ces essais, une construction particulière est apparue, construction que nous avons appelée *structure d'accord*. Cette structure est construite comme suit : nous prenons la liste des couples obtenue grâce à la méthode décrite à la section précédente. Puis, pour chaque

couple, nous recensons les *mots* qui peuvent venir s'intercaler entre ces deux éléments. Il arrive qu'un même affixe apparaisse dans la plupart de ces mots intercalés. Si cet élément apparaît dans une majorité de cas (plus de 50% des mots), nous considérons que la séquence générée correspond à une structure d'accord de la langue et que les éléments de ces structures sont en relation (tableau 3.3). Ainsi, en allemand, à partir du couple *des N-es*, nous recherchons les mots qui peuvent s'intercaler entre *des* et *N-es*. Puis nous recherchons un affixe qui se rencontre dans la liste des mots intercalés, et nous trouvons le préfixe *-en* Nous obtenons donc la structure *des N-en N-es*. Ces relations où les marques morphologiques surabondent sont assez faciles à découvrir mais n'existent pas dans toutes les langues. Cet algorithme, comme le montre le tableau 3.3[28] ne donne aucun résultat (ou très peu) sur certaines langues.

Allemand	Français	Italien
des N-en **N-es**	**les** N-s **N-s**	**la** N-a **N-ione**
die N-e **N-ung**	**la** N-e **N-ion**	**la** N-a **N-a**
eine N-e **N-e**	**des** N-s **N-s**	**del** N-o **N-io**
den N-en **N-ern**	**les** N-s **N-s**	**dei** N-i **N-i**
einen N-en **N-es**	**aux** N-s **N-es**	**della N-a N-ia**
Anglais	Swahili	Turc
was N-ly **N-ed**	**kile** ki-N **ki-N**	AUCUN
	wale wa-N **wa-N**	
	kila ki-N **ki-N**	
	vile u-N **u-N**	
	ule u-N **u-N**	

TAB. 3.3 – Les structures d'accord internes. Si certaines langues semblent posséder ce type de structures, d'autres ne s'en servent pas ou très peu.

Le même algorithme peut être appliqué à la recherche de régularités sur les éléments précédant ou suivant la structure. Le tableau 3.4 montre le résultat de la recherche de régularités morphologiques à droite de couples. Cette variante de l'algorithme produit assez peu de résultat, ou alors assez similaires à ceux déjà obtenus.

Allemand	Français	Italien	Anglais	Swahili
AUCUN	**les N-s** N-s	**la N-a** N-a	**was N-ly** N-ed	AUCUN
	la N-e N-e	**la N-a** N-a		
	des N-s N-s	**dei N-i** N-i		
	les N-s N-s			

TAB. 3.4 – Les structures d'accord externes à droite.

Nous avons alors eu l'idée de nous servir de ces propriétés pour catégoriser

[28] Dans tous les tableaux suivants, les éléments permettant la construction de la structure sont en gras.

certaines séquences. L'algorithme est simple :

- Pour tout couple, nous recherchons l'élément intercalé le plus fréquent
- Les couples ayant un même élément intercalé sont regroupés.

Nous avons généralisé la nature de l'élément intercalé. Il peut être un affixe (le cas traité jusqu'alors) ou bien un mot. Par exemple, les couples *il N-ait*, *on N-ait* admettent tous deux le mot *ne* comme élément intercalé le plus fréquent. Ils sont donc regroupés, ainsi que tous les couples partageant cette spécificité (tableau 3.5).

il		-ait
nous		-ons
on	ne	-ait
je		-ais
on		-e

les		-s
des		-s
de	-s	-s
ses		-s
aux		-s

TAB. 3.5 – Catégorisation de couples morphologiques grâce à l'élément intercalé le plus fréquent

De manière plus générale, alors que les techniques classiques catégorisent grâce aux contextes "extérieurs" droit et gauche, il nous semble que le contexte "intérieur" est beaucoup plus fiable. En effet, *les éléments qui viennent s'intercaler entre deux éléments d'une structure sont très caractéristiques de cette dernière*. Nous pouvons aussi utiliser cette technique pour catégoriser les éléments intercalés. Nous réalisons l'opération inverse : pour chaque couple, nous recherchons les éléments intercalés, qui sont alors regroupés dans une même catégorie. Le tableau 3.6 montre une catégorisation obtenue grâce au couple :*il N-ait*.

	ne		
	en		
il	se	-ait	
	lui		
	y		

	faire		
	le		
de	se	-er	
	leur		
	nous		

TAB. 3.6 – Le contexte des intercalés produit généralement une bonne catégorisation ...

Nous pensions avoir alors notre algorithme de catégorisation. Mais la mise au point de cette méthode s'est faite sur le français. Nous avons alors essayé d'appliquer la méthode à l'allemand. Les résultats furent catastrophiques ! Ainsi, le couple allemand *zu N-en* est caractéristique d'une structure verbale (à plus de 90%). Mais le fait de "casser" cette structure par certains éléments "dénature" totalement la structure et la transforme en groupe nominal (tableau 3.7). Les mots intercalés sont donc très hétérogènes (pronoms ou déterminants). Nous voyons aussi que si l'élément intercalé est un mot possédant une régularité morphologique, le résultat n'est pas meilleur. La structure *die N-te* correspond à une structure *Déterminant Substantif* ou *Déterminant Adjectif antéposé*, alors

que la structure *die N-e N-te* correspond à 70% à une structure *Déterminant Substantif Verbe*. Les éléments ainsi regroupés sont alors très divers.

	dem			die		-e
	den			eine		-e
zu	ihm	-en		eine	-e	-ung
	uns			eine		-te
	mir			die		-te

TAB. 3.7 – . . .et parfois ne produit rien de bon !

Avec une langue comme l'allemand (ce n'est pas la seule dans ce cas), une connaissance de la structure est réellement indispensable pour mener à bien une catégorisation des éléments. Nous voyons ici l'intérêt (ou l'inconvénient !) de travailler sur plusieurs langues. De plus, il faut se souvenir que ces séquences morphologiques ne sont construites que pour certaines langues, mais sont inexistantes pour des langues comme le chinois ou le vietnamien (la segmentation ne donne aucun affixe).

Un tel travail sur ces séquences morphologiques d'une langue est intéressant et a totalement sa place dans une méthode supervisée, mais si le but est d'automatiser le processus de découverte, alors ce critère là est insuffisant.

3.3.3 Les algorithmes de clustering

Cherchant à catégoriser des mots, nous nous sommes intéressé aux travaux déjà existants. La littérature sur ce sujet est assez abondante, et englobe différentes variantes ([Redington et al., 1996], [Finch and Chater, 1992], [Mahon and Smith, 1996], [Pereira et al., 1993], [Schütze, 1995], [Kohonen, 1978]). On trouvera un panorama de ces méthodes dans [Zhang, 1996]. Elles se réclament toutes du courant distributionnel. Dans ces approches, le but est de catégoriser les mots grâce à des contextes générés automatiquement, objectif similaire au nôtre. L'algorithme généralement utilisé est celui décrit par [Sokal and Sneath, 1963] (algorithme 5). Pour chaque mot, nous construisons sa distribution. Puis, nous agrégeons les mots qui ont une distribution similaire (grâce à un calcul de distance entre deux distributions). Au début les mots sont agrégés deux à deux puis aux classes déjà constituées. Ceci jusqu'à obtenir une seule classe. D'autres techniques numériques [Ploux and Victorri, 1998], [Honkela, 1997] [Elman, 1990], sont parfois utilisées, en particulier lorsque le but est de catégoriser uniquement les éléments *lexicaux*.

Cet algorithme pose plusieurs problèmes. Le premier concerne la construction des contextes des mots. Nous avons vu combien il était difficile de construire de tels contextes. Dans ces algorithmes, ce problème est tranché en considérant le contexte d'un élément comme une suite de mots environnant le mot à catégoriser. Cet environnement varie selon les auteurs. Il est généralement composé d'une séquence de n mots encadrant l'élément à catégoriser, n pouvant aller de un à cent. Mais [Brown et al., 1992] utilise seulement le contexte droit. Il semble que le contexte le plus usité soit celui composé de deux mots à gauche

Algorithme 5 catégorisation des mots

pré-requis C : un corpus
  Créer un cluster par mot de C
  tant que Il y a plus d'un cluster **faire**
    trouver les deux clusters les plus proches
    créer un nouveau cluster contenant les deux clusters
    éliminer les deux clusters de la liste de clusters.
  fin tant que

et à droite du mot à catégoriser. À noter que ces contextes sont généralement constitués des mots les plus fréquents du corpus. Ainsi apparaîtront seulement les mille mots les plus fréquents sur un corpus comprenant plusieurs millions de mots. De même, tous les mots du corpus ne seront pas catégorisés. Là encore, seuls les plus fréquents le seront.

Un deuxième problème est celui du calcul de la distance entre éléments, ceci afin de déterminer si deux éléments partagent une distribution similaire. Là, une demi douzaine de distances, très diverses sont utilisées : la distance euclidienne dans [Huckle, 1995], la distance *kullback-leibler* dans [Pereira et al., 1993], l'ACMI (Average Class Mutual Information dans [Mahon and Smith, 1996]). On trouve dans [Finch, 1993, pages 94-95] une description de certaines mesures citées.

Mais quels sont donc les résultat de tels algorithmes ? La figure 3.2 et celles qui se trouvent en annexe D illustrent différents essais sur les mêmes mots d'un texte. Nous avons pris les vingt mots les plus fréquents de notre corpus *français01*, et les avons classés selon différents contextes. Le résultat de cette catégorisation se présente sous forme d'un dendrogramme. Nous avons essayé plusieurs contextes : un mot ou deux avant et/ou après l'élément à catégoriser. Nous pouvons voir deux classes[29] majeures qui resortent : la classe des déterminants et celle des prépositions. Une troisième classe composée des éléments *il* et *qui* apparaît parfois. Les meilleurs contextes semblent être ceux des figures D.1, D.2, D.3 et D.6. Le fait de passer d'un élément à deux peut dégrader considérablement la classification. Ainsi les pires contextes sont ceux construits avec deux mots avant (figure D.4) ou deux mots après (figure D.5).

La qualité du résultat ne dépend donc pas du nombre d'éléments qui constituent les contextes. Les résultats obtenus ne sont pas mauvais en soi puisqu'on retrouve bien les classes attendues : prépositions et déterminants. Mais le problème n'est pas là. Cette technique offre divers inconvénients que nous allons détailler dans la section suivante.

Nous pouvons appliquer ces algorithmes non pas en utilisant les mots mais les séquences morphologiques mises à jour grâce à l'opération de segmentation. Le résultat est similaire et les problèmes restent les mêmes.

[29] La création de ces classes est faite de manière supervisée : nous avons utilisé un critère visuel.

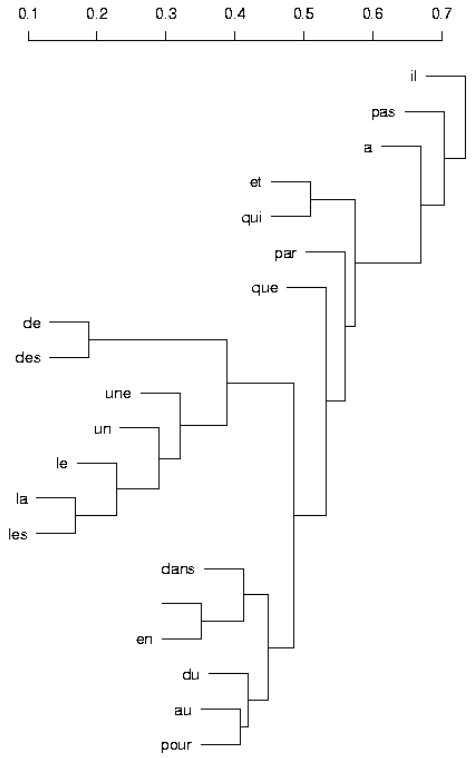


FIG. 3.2 – Catégorisation de mots : le contexte est constitué du mot précédent.

3.4 La nécessité de la connaissance structurelle

Comme les figures de l'annexe D le montrent, les résultats obtenus avec une telle méthode de catégorisation peuvent être considérés comme (assez) bons, bien que ces catégorisations n'aient jamais été utilisées à notre connaissance dans une application quelconque (mais l'objectif n'est pas là). Nous allons expliquer pourquoi nous ne nous sommes pas orienté vers une telle approche. Premièrement, *elle ne permet pas de catégoriser un élément dans différentes classes*. En effet, à chaque élément est associée sa distribution totale, et les catégories sont formées à partir de cette distribution. Il est donc impossible de catégoriser un élément dans plusieurs catégories. La catégorie de l'élément va généralement correspondre à son comportement le plus fréquent. Le deuxième problème est celui de la définition arbitraire de la distribution. Enfin, cette méthode ne traite que les éléments fréquents d'un corpus. La méthode que nous proposons (chapitre 6) palie tous ces inconvénients. Elle permet :

- une construction des contextes adéquats
- une polycatégorisation des éléments
- la prise en compte des hapax

Elle offre de plus une solution multilingue. Ce résultat est obtenu en considérant en premier la structure de la langue. La polycatégorisation est obtenue en effectuant un traitement inverse de celui des algorithmes décrits ci-dessus. Au lieu de prendre chaque mot et de lui construire son contexte total, nous construisons d'abord un contexte pour chaque catégorie de la langue. Si un élément apparaît dans plusieurs de ces contextes, alors il est polycatégorisé. Cette technique demande bien sûr de connaître a priori les catégories possibles de la langue, et donc une connaissance de la structure linguistique. Nous voyons qu'il est difficile, à notre avis de parler de catégorisation sans parler de structures. Les catégories obtenues ne sont que le résultat de la connaissance que nous avons de la langue. De plus, la connaissance de la structure seule permet une catégorisation efficace des éléments. Notre principe de catégorisation rejoint celui décrit dans [Halliday, 1961] :

> A class is always defined with reference to the structure of the unit next above, and structure with reference to classes of the unit next below. A class is *not* a grouping of members of a given unit *which are alike in their own structure.*[Halliday, 1961, page 261]

Le critère retenu est assez opposé à ce que l'on peut dans les ouvrages de la communauté de l'apprentissage :

> Clustering and segmentation is the problem of creating a partition of the data base so that all members of each set of the partition *are similar according to some metric* [Decker and Focardi, 1995].

Ainsi, pour Halliday, il est nécessaire de connaître le niveau supérieur à un élément donné pour pouvoir catégoriser cet élément. Voilà pourquoi, comme nous allons le voir, les morphèmes sont catégorisés grâce à la connaissance des syntagmes simples, et ces derniers grâce aux couples de syntagmes (section 4.8). Si nous avons dit qu'il était difficile de parler de catégorie sans idée de structure,

le contraire est vrai. Nous reprenons pleinement à notre compte les remarques suivantes de M.A.K. Halliday :

> The relation between stucture and class is a two-way relation, and there is no question of "discovering" one "before" the other. In any given instance there may be descriptive reasons for stating the one without the other ; but all structures presuppose classes and all classes presuppose structures.[. . .]

Le fait de ne pas associer structure et catégorie rend très difficile la validation d'une catégorisation. Seule la structure offre un critère de validation des catégories obtenues. Inversement, la structure se construit grâce aux catégories mises à jour. Au début de ce travail, nos différents essais ont produit plusieurs catégorisations. Nous avons alors été obligé de juger de ces catégorisations. Là, plusieurs attitudes sont possibles. Soit on se laisse guider par les catégories classiques de la langue. Le résultat est souvent la génération de contextes ad hoc pour une catégorie donnée dans une langue donnée. En effet, les contextes pour une catégorie donnée varient aussi d'une langue à une autre. Ainsi le contexte permettant la catégorisation des prépositions en français (la prise en compte du mot suivant, majoritairement les déterminants, offre un très bon contexte), est inadapté en russe où les articles définis n'existent pas et les prépositions imposent un cas à leur substantif (il faudra au moins considérer un contexte comprenant les marques casuelles). La deuxième solution consiste à avoir le moins d'a priori possibles. Par exemple ne pas rejeter une classe constituée de substantifs et de verbes. [Hughes and Atwell, 1994] dénomme cette méthode par l'expression : "looks good to me". Cette approche ne peut se faire qu'en construisant parallèlement une structure de la langue. En utilisant le critère de catégorisation de Halliday (recourir à un niveau supérieur pour catégoriser un élément d'un niveau donné), nous voyons que les catégories obtenues sont très fonctionnelles, puisque la discrimination entre éléments se fait généralement grâce à une différence de fonction dans la structure supérieure, différence de fonction qui se traduit par une différence dans la distribution des éléments.

Conclusion

Pour terminer, nous rappelons que la segmentation des mots n'est pas une finalité en soi, et n'est intéressante que parce qu'elle fournit des marques de mise en relation d'éléments et qu'elle permet la génération des couples morphologiques de la langue, qui serviront de point de départ à notre algorithme de catégorisation (section 6.3). L'ensemble de mots fréquents et d'affixes va servir d'élément de base à la construction des contextes, opération réalisée grâce aux structures définies dans la partie suivante.

Nous n'avons pas voulu essayer d'améliorer les résultats obtenus lors de l'opération de segmentation, d'une part parce qu'ils sont suffisants pour passer aux étapes suivantes, d'autre part parce que l'amélioration nécessite assez souvent une connaissance de la structure de la langue. Il nous semble aussi qu'il était important de ne pas rester à ce premier niveau de la structure, en négligeant les niveaux supérieurs, beaucoup plus intéressants nous semble-t-il.

Il est à noter qu'il n'est pas nécessaire de trouver tous les morphèmes de la langue. Comme les mots, ils obéissent à la loi de Zipf (section 1.10.3), et donc seuls les plus fréquents suffisent à amorcer la découverte de structures. Le cas typique est l'anglais, où les morphèmes suivants suffisent : *-ed, -ly, -ing,-s, -ion.* De plus, le grand nombre de morphèmes d'une langue est généralement dû à la combinaison de plusieurs morphèmes "basiques" et non pas à une plus grande diversité dans la morphologie (turc, swahili).

Les morphèmes les plus importants pour nous, c'est à dire les morphèmes qui marquent une relation entre éléments sont les plus faciles à trouver, car ils sont généralement très fréquents. La segmentation n'a pour but que la découverte des éléments qui peuvent nous aider dans la découverte des structures, comme nous le verrons dans le chapitre suivant.

Une étude manuelle de la morphologie des mots, accompagnée de ces algorithmes, permet en quelques heures (deux ou trois) d'avoir une très bonne connaissance morphologique de la langue. La morphophonologie des langues n'a pas du tout été prise en compte, puisque le travail se base sur des textes écrits. Cette lacune ne semble pas avoir eu de conséquence. Les éléments recueillis avec les algorithmes présentés ici (en particulier les seuils) suffisent à lister les éléments importants de la structure de la langue (en particulier les morphèmes relationnels).

Les résultats bruts de ces algorithmes (sans aucune supervision) donnent déjà un très bon aperçu de la morphologie de la langue. L'ordinateur est un outil très performant dans ce cadre de travail qui consiste à manipuler des chaînes de caractères. Des algorithmes très simples peuvent détecter des séquences mor-

phologiques de la langue. Ainsi le simple fait d'observer quelles sont les lettres qui peuvent apparaître en début ou en fin de mots, donne déjà des indications intéressantes sur la morphologie des langues. Nous voyons là un exemple simple de la puissance de l'ordinateur : cet algorithme prend quelques secondes de temps d'exécution, alors qu'il prendrait plusieurs dizaines d'heures pour un humain.

Troisième partie

Les structures

Introduction

Nous avons vu dans la partie précédente l'utilité et les limites de ce que l'on peut appeler le critère morphologique. La difficulté que nous avons rencontrée à mettre au point une technique de catégorisation des éléments, nous a amené à nous poser la question suivante : n'existe-t-il pas une propriété formelle de la structure des langues que nous n'utilisons pas. Nous avons alors recherché dans les travaux des structuralistes quelles étaient les marques formelles qu'ils utilisaient. La littérature comme [Sapir, 1921] ou [Vendryes, 1923] nous en offre plusieurs :

- l'affixe (la morphologie)
- la position
- l'accent
- le morphème zéro

Le premier critère, la morphologie, a déjà été pris en compte. Travaillant sur l'écrit, nous avons éliminé le troisième : l'accent[30]. La suite du travail, en particulier sur le vietnamien, nous a montré que cette option était la bonne. D'ailleurs nous ne considérons pas ce critère comme étant un critère structurel des langues, mais nous le classerions plutôt comme élément phonologique (une différence de ton n'est-elle pas équivalente à une différence phonologique ?). Reste le deuxième critère : la position. Que faut-il entendre par position ? L'illustration classique[31] consiste à permuter les mots *Pierre* et *Paul* dans *Pierre frappe Paul* qui produit *Paul frappe Pierre*. Le sens de ces deux énoncés n'est pas le même[32]. Si ce fait est facilement admissible, il n'en reste pas moins qu'un problème se pose : comment mettre à profit un tel indice, comment l'exploiter ? Faut-il recenser toutes les positions d'un élément dans une phrase, toutes les positions où l'élément n'apparaît pas, toutes les permutations entre éléments ? Ce problème rejoint en fait le problème de la définition du contexte pour un élément. La réponse est apportée dans la section 4.2 : nous verrons que l'étude de seulement deux positions particulières : la première position et la dernière, a suffi à guider notre recherche des structures des langues. Ce critère positionnel a conduit à la construction d'une structure de la langue, avec différents niveaux d'éléments (chapitre 4.4 et 4.8).

[30] Vendryes précise :

> Par accent il faut ici entendre d'ordinaire l'accent de hauteur, le ton. [Vendryes, 1923, page 95]

[31] repris de [Vendryes, 1923, page 99]

[32] surtout pour Pierre.

L'intérêt de ce travail n'est pas d'avoir découvert de nouvelles structures, celles manipulées ici sont bien connues, mais de présenter une méthode formelle et automatique afin de les découvrir à partir d'un simple texte d'une langue donnée. Savoir qu'il existe telle ou telle structure ne permet pas d'identifier celle-ci. Il a donc fallu découvrir non pas les structures mais mettre au point un moyen permettant d'identifier automatiquement *les traces formelles de ces structures.*

Chapitre 4

La découverte des structures

Sommaire

La structure d'une langue est caractérisée par la régularité des faits d'une langue, l'existence de classes, la primauté de l'ensemble (= système) sur l'unité et enfin les différences et les ressemblances de la structure d'une langue à l'autre. [Mahmoudian, 1981].

Nous allons donc parler de structures dans ce chapitre. Mais qu'entendons nous par structure ? Pour préciser la chose, nous allons citer les premières lignes de [Harris, 1954] :

> Dans le cadre de cet exposé, nous donnerons au terme *structure* le sens large suivant : un ensemble de données est structuré au regard d'une certaine caractéristique dans la mesure où nous pouvons constituer à partir de cette caractéristique un système organisé de règles qui décrit les membres de l'ensemble et leur interrelation. [Harris, 1954, page 14]

Une structure $\mathcal{S}$ est donc un couple ($\{\mathcal{E}\}$, $\{\mathcal{R}\}$), où $\{\mathcal{E}\}$ est l'ensemble des éléments composant la structure, et $\{\mathcal{R}\}$ l'ensemble des règles de construction régissant les relations entre éléments. Autrement dit, une séquence d'éléments de $\{\mathcal{E}\}$ ne suffit pas à former une structure $\mathcal{S}$: ils doivent obéir à des règles de composition. Les structures que nous utilisons sont qualifiées de *structures formelles* car leur construction n'utilise que des critères de *forme*. De plus, nous ajoutons une autre contrainte : ses critères de forme ne peuvent être extraits que du corpus que nous étudions.

Ce chapitre s'articule selon les points suivants : nous allons d'abord voir quelles sont les indices qui nous ont permis de mettre à jour la hiérarchie grammaticale, en particulier grâce à une réflexion sur les propriétés d'un objet linéaire (sections 4.1 à 4.3). Puis nous décrirons les structures mises à jour grâce à ces indices (sections 4.4 à 4.8). Enfin, nous finirons en précisant certaines méthodes et notions (section 4.9 à 4.12).

4.1 La segmentation en "entre-ponctuations"

Dès les premiers mois de la thèse, la ponctuation s'est révélée importante. Elle est apparue lors des premières expériences de catégorisation (ces essais furent réalisés sur le français). En essayant de construire des contextes appropriés, nous avons plutôt trouvé un contexte inapproprié : celui incluant des ponctuations. Les mots de part et d'autre d'une ponctuation ne sont pas en relation[33]. Nous avons donc supposé que les ponctuations étaient des délimiteurs de séquences. Quelles séquences ? Nous n'avions pas de réponse alors. Nous avons donc segmenté le corpus en utilisant toutes les ponctuations comme séparateur. Cette segmentation nous fournissait des séquences de mots que nous appellerons des *entre-ponctuations*. Les séquences ainsi construites étaient majoritairement des séquences de trois, quatre, ou cinq mots (tableau 4.1).

Nous pensions pouvoir utiliser les entre-ponctuations courtes (de longueur trois ou quatre) comme définition de la distribution d'un élément. Ces séquences étaient en effet assez courtes pour être fréquentes, et donc utilisables dans un algorithme de catégorisation. Nous nous sommes vite aperçu (en travaillant sur le français) que ces entre-ponctuations n'offraient en fait aucune régularité structurelle (elles correspondent à toutes sortes de structures), et n'étaient donc pas des contextes beaucoup plus intéressants qu'un contexte arbitrairement choisi de

[33] Ou très rarement.

Longueur (en mots)	Effectif
1	2182
2	3230
3	3541
4	2846
5	2607
6	2330
10	1364
20	327
50	6

TAB. 4.1 – Effectif des séquences entre-ponctuations dans le corpus *français01*.

Mot	Effectif en position		
	un	deux	trois
mr	280	52	0
le	163	105	0
en	139	67	0
de	112	152	0
dans	88	3	0
les	75	77	0
et	65	97	**2**
il	64	8	0
la	60	131	0
Somme	1046	692	0
Effectif total		3541	

TAB. 4.2 – Répartition des débuts des entre-ponctuations de trois éléments.

longueur similaire. Mais en travaillant sur ces séquences, en particulier sur la position relative des éléments, nous avons remarqué une caractéristique : certains mots apparaissaient très fréquemment en début de ces séquences, et n'apparaissaient jamais en fin de celles-ci. Le tableau 4.2 montre ces résultats pour les séquences de trois mots. Dix mots représentent près de 30% des débuts de séquences, mais un seul de ces mots n'apparaît que deux fois (le mot *et*) en fin de séquences. À partir de cette observation, nous avons sélectionné[34] une première représentation de la structure des langues : une séquence d'éléments (syntagme ou proposition), ces derniers possédant des marqueurs caractéristiques de début . En travaillant sur d'autres langues, la notion de début a été généralisée à celle de début et de fin de séquences . Ainsi la structure des langues est considérée comme une structure linéaire où les différents éléments structurés possèdent des marqueurs de frontière (figure 4.1).

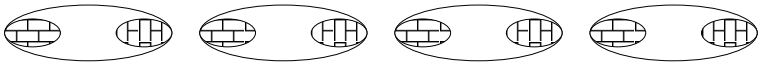


FIG. 4.1 – Une première idée de la structure de la langue : une séquence d'unités dont les débuts et les fins sont marqués par des éléments caractéristiques.

La méthode présentée ici repose entierement sur cette notion de marqueur de frontière. L'utilisation de cette notion n'est nouvelle ni en linguistique ni en TAL, comme le montre pour la linguistique le passage suivant trouvé dans [Tesnière, 1959, page 25], même si nous appliquons cette définition non pas au mot mais au syntagme et à la proposition.

> Or un segment linéaire est une portion de ligne comprise entre deux points. De même un mot est une portion de la chaîne parlée comprise entre deux coupures. En d'autres termes, on ne saurait définir le mot par lui-même, mais seulement par les **coupures** qui en marquent le commencement et la fin.

Nous trouvons aussi cette notion en TAL, en particulier dans des travaux portant sur l'extraction terminologique, [Bourigault, 1993], [Debili, 1982], basée sur la reconnaissance de groupes nominaux, où les prépositions, déterminants et groupes verbaux servent de délimiteurs à ses groupes. On la retrouve aussi dans des travaux en apprentissage de structures comme ceux de [Marcus, 1991] ou [Magerman, 1991], mais ces travaux utilisent d'une part un lexique, et d'autre part les données fournies au système sont triées et seules les phrases simples sont utilisées. Un travail très intéressant est celui de [Ramshaw and Marcus, 1995] où l'apprentissage (à partir de corpus étiqueté et paranthésé) des structures nominales et verbales (appelées *chunk* verbal ou nominal) est basée sur cette notion de frontière. Il utilise en particulier trois marques : I, O, et B[35], où un mot marqué par I est à l'intérieur d'un groupe nominal, un mot marqué par

[34] Au détriment d'autres représentations comme le schéma X-barre [Chomsky, 1970], bien que dans ce schéma, les spécifieurs (spec) peuvent être interprétés comme des marqueurs de frontières, même si cette terminologie n'est pas utilisée par Chomsky.

[35] probablement pout *Inside*, *Outside*, et *Boundary*

O est à l'extérieur, et un mot marqué par B correspond au premier mot le plus à gauche d'un groupe nominal (donc le marqueur de début du groupe). Nous voyons que la notion de frontière n'est pas généralisée aux marqueurs de fin ni au niveau propositionnel. La technique d'apprentissage est celle décrite dans [Brill, 1993]. Nous voyons donc que ces notions de marqueurs de frontières semblent être très utiles dans un travail de segmentation, mais tous ces travaux ont une connaissance *a priori* des mots qui peuvent jouer le rôle de maruqeur de frontière. Comme nos données se résument à un simple texte, *notre problème est différent : nous devons mettre au point une méthode qui nous permette d'extraire automatiquement la liste de ces marqueurs.* Ce travail constitue la première phase de la méthode décrite au chapitre 6. Une fois certains de ces marqueurs identifiés, la génération des structures syntagmatique et propositionnelle est possible.

4.2 Des propriétés d'un objet linéaire

Les marqueurs de frontière : Mais, si ces notions de début et de fin sont assez simples en soi, il nous a fallu près de deux ans pour les exploiter correctement. Nous allons présenter les différentes caractéristiques d'un objet linéaire que nous allons utiliser dans notre méthode. Le premier stade, assez facile, a été la généralisation des débuts aux fins. En effet, travaillant sur des langues privilégiant les marqueurs de début, seul le concept de début a d'abord été exploité. Puis, la nécessité d'introduire des marqueurs de fin s'est très vite fait sentir pour des raisons pratiques et théoriques. La raison pratique provient des langues postposées (comme le turc) qui utilisent des mots pour le marquage des fins de séquences. La raison théorique est la suivante : pour segmenter une séquence d'objets linéaires, on peut utiliser deux méthodes : soit le marquage des débuts de séquences, soit le marquage des fins de séquences. Il y a donc, en théorie, symétrie parfaite entre ces deux notions (figure 4.2).

Une combinaison des deux est bien sûr possible (elle se rencontre même assez souvent dans les langues). Le problème majeur auquel nous nous sommes confronté est que tous les segments d'un corpus ne sont pas toujours marqués par un début ou une fin. Cette lacune ne gène pas trop le processus de découverte des structures si suffisamment de segments dans le corpus sont marqués (ce qui est toujours le cas), mais elle représente un inconvénient majeur dans un processus d'analyse (figure 4.3).

Comme le montre le tableau 4.3, il existe des éléments caractéristiques de ces marqueurs qui sont facilement identifiables. Certains éléments (ici des mots mais cela peut aussi être des morphèmes) ont un comportement particulier. Le tableau est construit comme suit : pour chaque mot, nous recensons le nombre de fois où il apparaît après une ponctuation (colonne Début) et avant une ponctuation (colonne Fin). Nous voyons alors que certains éléments n'apparaissent pratiquement jamais après une ponctuation (comme les mots allemands *als*, *in*), et d'autres jamais avant une ponctuation (comme *de*, *il* en français). Ces éléments sont des éléments caractéristiques des marqueurs de début ou de fin de séquences. La construction de ce tableau ne permet pas de mettre à jour le

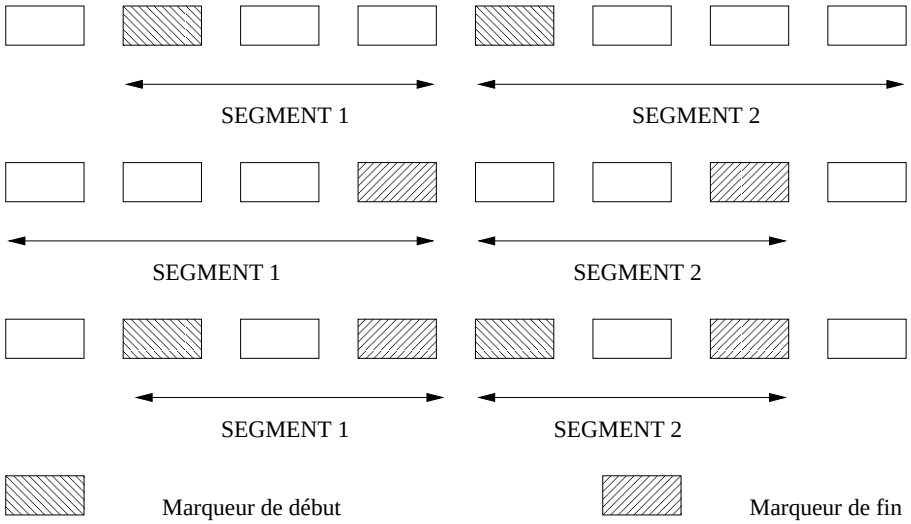


FIG. 4.2 – Comment construire des structures dans une séquence linéaire ? En marquant leur début ou leur fin, ou les deux à la fois.

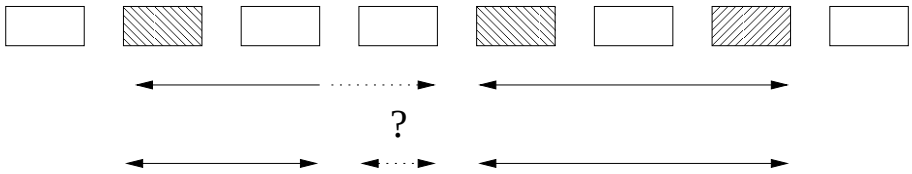


FIG. 4.3 – Toutes les séquences ne sont pas toujours marquées à leur frontière. Se pose alors le problème de trouver la segmentation correcte. A-t-on deux segments ou trois ?

comportement de tous les mots du corpus : il ne donne aucune indication en ce qui concerne les éléments polycatégoriels. Par exemple, dans notre méthode, l'élément *pas* appartient à trois catégories : début de groupe nominal, fin de groupe verbal et noyau de groupe nominal[36]. Si l'on regarde la ligne du tableau concernant cet élément, ces deux comportements ne sont pas identifiables aisément car ils sont opposés. Les 54 occurrences de début sont dues à la catégorie de début de groupe nominal, et les 88 occurrences de fin à la catégorie de fin de groupe verbal (le *pas* substantif ne correspond qu'à une occurrence des débuts et six occurrences des fins.)

	Mot	Effectif	Début	Fin	
français	de	14943	648	3	1er groupe
	la	8427	1300	0	
	il	1605	1195	0	
	mais	845	694	69	
					2ème groupe
	et	5311	760	115	3ème groupe
	pas	1523	54	88	
	avons	54	0	0	4ème groupe
	grandes	41	0	0	

	Mots	Effectif	Début	Fin	
Allemand	daß	1251	1169	0	1er groupe
	als	653	362	1	
	in	1566	241	0	
	die	2943	702	4	
	her	65	0	40	2ème groupe
	zurück	168	4	139	
	ich	4313	1725	264	3ème groupe
	an	755	79	159	
	des	712	6	0	4ème groupe
	meinem	89	0	0	

TAB. 4.3 – Position de certains mots en français et en allemand. On voit apparaître pour certains mots une caractéristique : ils ne finissent jamais une séquence (premier groupe), ou ne la commencent jamais (deuxième groupe). Certains mots (troisième groupe) ont un comportement apparemment neutre par rapport aux ponctuations : ils peuvent commencer ou finir une séquence. Enfin, il existe des mots qui n'apparaissent jamais avant ou après une ponctuation.

La figure 4.4 explique l'interprétation qui est faite des marqueurs de début :

1. Ils n'apparaissent pas avant une ponctuation
2. Ils peuvent apparaître après une ponctuation

[36] Nous utilisons pour l'instant la terminologie classique pour dénommer les catégories.

3. Ils sont en relation[37] avec l'élément suivant

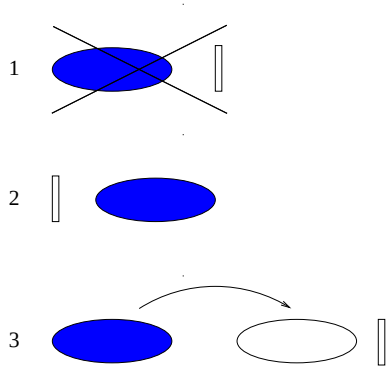


FIG. 4.4 – Propriété d'un marqueur de début. La barre symbolise le début ou la fin d'une séquence.

Les marqueurs de fin sont considérés de façon symétrique. Selon cette conception structurelle, un élément peut appartenir à trois catégories :

1. la catégorie des marqueurs de début
2. la catégorie des marqueurs de fin
3. la catégorie des noyaux

Un *mot* (qui, comme nous allons le voir, ne fait pas partie de notre hiérarchie structurelle) peut appartenir à ces trois catégories ou à une combinaison de celles-ci. La catégorie des noyaux correspond aux éléments qui ne sont ni marqueur de début, ni marqueur de fin. Ils se trouvent entourés par des marqueurs de début ou fins. Ils correspondent, pour le niveau syntagmatique par exemple, à un élément radical (section 4.6).

Le recours à certains éléments pour segmenter un texte en unités n'est pas innovant comme l'indique le commentaire suivant :

> [...] ; d'autre part, les déterminatifs égyptiens, plus nombreux, plus aisés à identifier que leurs correspondants cunéiformes, lui [Champollion] permettait de séparer les mots, [...] [Février, 1948].

De même, [Aristote, 1990] défini les articles comme :

> L'article est un mot dépourvu de signification qui indique le commencement, la fin ou la division de la phrase [...][38]. [Aristote, 1990, 1457a].

Nous retrouvons donc bien le fait que certains éléments délimitent une séquence (une phrase pour Aristote) en segments.

[37] Dans le reste de ce chapitre, le terme *relation* signifie relation de dépendance (de subordination). Tout autre type de relation sera noté explicitement.

[38] Nous encourageons vivement les lecteurs intéressés à lire le texte original, les différentes traductions lues offrant de grandes différences terminologiques.

La détection des niveaux hiérarchiques : Mais ce premier modèle est insuffisant pour représenter la structure des langues. Nous avons pour l'instant supposé qu'un seul type de segment existait dans ces séquences. Mais il peut en exister plusieurs, comme nous allons le voir dans les sections suivantes. Comment faire alors pour pouvoir les différencier ? Pour cela, il suffit d'utiliser différents types de marqueurs de début et de fin. Chacun de ces types de marqueurs va caractériser un type de segment particulier (figure 4.5). Par ce moyen, nous

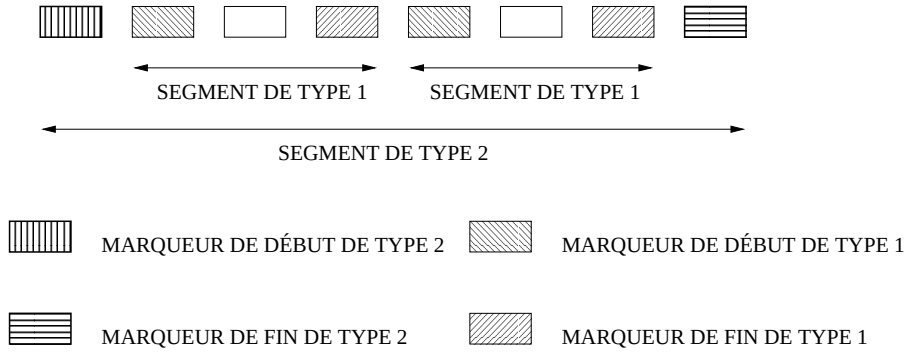


FIG. 4.5 – Plusieurs segments peuvent être définis en utilisant différents types de marqueurs de début et de fin.

avons identifié plusieurs types de structures dans les langues qui sont détaillés dans les sections 4.5 à 4.8. Notons que les structures définies par ce procédé sont *hiérarchiques*[39] : une structure d'un niveau donné est construite avec les éléments d'une structure inférieure. Nous verrons dans la section 4.7 comment les différents types de marqueurs peuvent être identifiés facilement. La nature structurelle de ces marqueurs peut varier selon le niveau hiérarchique de la structure. D'une manière générale, *les marqueurs de frontière peuvent utiliser toutes les structures inférieures à la structure dont ils marquent les frontières.* Ainsi la structure de premier niveau utilise des marqueurs dont la structure est l'élément de base. La structure de deuxième niveau peut utiliser des marqueurs de structure basique, mais aussi des éléments de premier niveau (figure 4.6). *Une structure de niveau n peut utiliser comme marqueur de frontière les éléments des niveaux 0 à n-1, le niveau 0 étant le niveau de base indécomposable.*

Les problèmes rencontrés : À partir de cette conception de la structure de la langue, les questions auxquelles nous devrons répondre sont les suivantes :

1. Comment identifier les éléments qui marquent les débuts et fins de structure ?
2. Une fois un marqueur de frontière identifié, quelle(s) structure(s) délimite-t-il ?
3. Comment gérer la polycatégorisation des éléments ?
4. Comment gérer les structures non délimitées ?

[39] le terme de hiérarchie est définie à la section 4.4.1

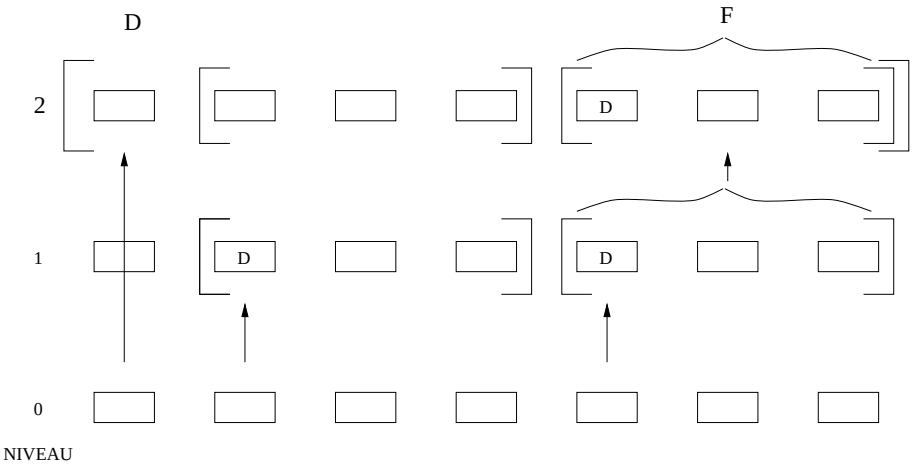


FIG. 4.6 – Une structure d'un niveau hiérarchique donné peut utiliser tous les niveaux inférieurs comme marqueurs de frontière. Le début de la structure de niveau 2 est marqué par un élément de niveau 0 , et sa fin par une structure de niveau 1.

La réponse au premier problème, l'identification des marqueurs de frontière, est partiellement donnée par le tableau 4.3 : certains éléments de la langue sont très caractéristiques de la catégorie à laquelle ils appartiennent, et sont assez facilement identifiables. En utilisant ces éléments comme amorce, nous avons mis au point des algorithmes permettant l'identification des autres éléments de la catégorie concernée, élément qui eux ne sont pas aussi facilement identifiables (car souvent polycatégoriels). Ces algorithmes sont expliqués dans le chapitre 6.

Pour répondre à la deuxième question, il est nécessaire d'identifier toutes les structures des langues. La liste de ces structures est présentée dans les sections 4.6 à 4.8. Une fois ces structures identifiées, il suffit de trouver les éléments caractéristiques qui marquent les frontières de celles-ci.

La troisième question concerne une des caractéristiques de la langue : la polycatégorisation des éléments : un élément peut appartenir à plusieurs catégories de la structure. Comme nous l'avons vu, pour chaque type de segments de la structure, il existe trois catégories au maximum (début, noyau, fin). Si le nombre de niveaux dans la hiérarchie (le nombre de types de segments différents) est n, le nombre maximal de catégories de la structure est $3n$. Un élément peut, en théorie, appartenir à ces $3n$ catégories. La liste des catégories identifiées dans ce travail est donnée à la section 4.12.

La polycatégorisation la plus délicate à traiter est celle qui concerne les marqueurs de frontière : les éléments qui peuvent être à la fois marqueurs de début et marqueurs de fin (figure 4.7). Cela a une répercussion directe sur la construction des structures élémentaires que sont le syntagme et la proposition. Comme nous le verrons dans la section 4.6, les contextes dans lesquels ils sont marqueur de début sont très différents des contextes dans lesquels ils jouent le rôle de marqueur de fin.

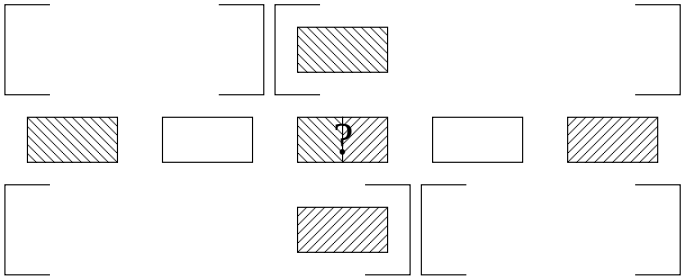


FIG. 4.7 – Un élément peut appartenir à plusieurs catégories. Se pose alors le problème de l'analyse de la séquence, c'est-à-dire reconnaître la bonne structure.

La quatrième réponse est généralement réglée lors de la construction des différentes structures composées (section 4.8). Ce problème relève plus de l'analyse que de la découverte des structures. Une analyse (c'est-à-dire l'identification (la catégorisation) des éléments d'une séquence) est nécessaire dans le processus de découverte, mais cette analyse peut ne porter que sur une certaine partie du corpus. Notre expérience sur les langues étudiées nous montre que toutes les structures de la langue possèdent des éléments caractéristiques de marqueurs de frontière qui permettent leurs identifications.

4.3 Le rôle de la ponctuation

Comme nous l'avons expliqué à la section 4.1, nous utilisons la ponctuation pour construire des séquences de mots. Les signes de ponctuation utilisés sont les suivants :

`. , : ; ? !`

Le fait de considérer tous les signes de ponctuation sur le même plan (points et virgule par exemple) étonne souvent. La raison en est simple. Si les points définissent une unité de segmentation classiquement appelé *phrase*, la segmentation produite par les autres signes (la virgule en particulier qui est souvent le signe le plus fréquent du corpus) n'est généralement pas retenue. Or les séquences générées par cette segmentation sont toutes aussi intéressantes dans une procédure de découverte. Nous verrons dans la section 4.4.2 que les différentes segmentations considérées de l'écrit ne sont vues que comme des points d'entrée qui permettent la génération des "vraies" unités de la structure grammaticale. Ainsi, si la segmentation en phrases correspond généralement (du point de vue de la taille des unités segmentées) à une segmentation du niveau de la proposition et des couples de propositions, la segmentation produite en utilisant les virgules peut correspondre à tous les niveaux de la structure grammaticale : syntagme, couple de syntagmes, proposition et couple de propositions. Si cette segmentation produite peut sembler irrégulière, elle possède une caractéristique essentielle : *elle segmente rarement un syntagme en deux.* Nous pouvons donc considérer que la segmentation en entre-ponctuations nous fournit des séquences de syntagmes (séquences qui peuvent correspondre au non à des propositions).

En fait, ce travail de segmentation a uniquement pour but la construction de segments qui vont permettre la génération des structures de la langue. Et ces segments peuvent être obtenus avec ou sans ponctuation. Il existe des textes qui ne possèdent pas de ponctuation, par exemple notre corpus latin. Nous utilisons alors un autre critère pour obtenir ces segments. Dans le cas du corpus latin, une segmentation alternative est celle en verset, qui sont "visuellement" faciles à délimiter comme le montrer l'extrait suivant :

> 1 :1 in principio creavit Deus caelum et terram
>
> 1 :2 terra autem erat inanis et vacua et tenebrae super faciem abyssi et spiritus Dei ferebatur super aquas
>
> 1 :3 dixitque Deus fiat lux et facta est lux
>
> 1 :4 et vidit Deus lucem quod esset bona et divisit lucem ac tenebras
>
> 1 :5 appellavitque lucem diem et tenebras noctem factumque est vespere et mane dies unus

De même, un poème offre une segmentation visuelle en vers[40]. Ce critère de segmentation en blocs visuels peut être appliqué à tous les textes, même très anciens. Ainsi les segments obtenus sur le texte de la figure 4.8 seraient tout simplement la ligne de hiéroglyphes.

Nous ne disons pas que ces segments obtenus correspondent à des structures de la langue, mais ils sont utilisés pour découvrir ces structures. Dans la suite de ce travail, les segments obtenus en utilisant cette méthode de segmentation seront toujours appelés entre-ponctuations, même s'ils n'ont pas été obtenus grâce à la ponctuation (comme avec le corpus latin).

4.4 Les structures

Lorsque nous nous sommes intéressé à la structure des langues, nous avons consulté la littérature existante sur ce point. Un écueil est apparu. Si les linguistes utilisent bien des unités structurelles, il n'existe pas de consensus sur leur définition. De plus, la plupart offre des définitions inopérantes dans le cadre d'un traitement formel. Par inopérantes, nous entendons qu'à partir de la définition d'une unité, nous ne pouvons générer d'algorithme qui permette une segmentation *systématique* et *régulière* d'une séquence en cette unité grâce à des ressources formelles.

Un intérêt du travail présenté ici est de fournir des définitions formelles[41] et opératoires en vue d'un traitement automatique. Ainsi, à partir de notre définition du syntagme (section 4.6), un algorithme de segmentation est réalisable (et réalisé). Le fait que la définition se base sur des critères purement formels (aucun recours au sens des énoncés) facilite grandement l'écriture de tels algorithmes.

Cette section veut aussi illustrer le problème qu'a posé et pose la définition des concepts (les niveaux de la hiérarchie) en linguistique structurale. On peut se poser légitimement la question de savoir s'il existe réellement une hiérarchie dans la structure des langues. Comme beaucoup, nous répondons par l'affirmative et

[40] Il y a toujours des exceptions.

[41] C'est-à-dire qui n'utilisent pas d'autres critères comme ceux sémantiques ou discursifs

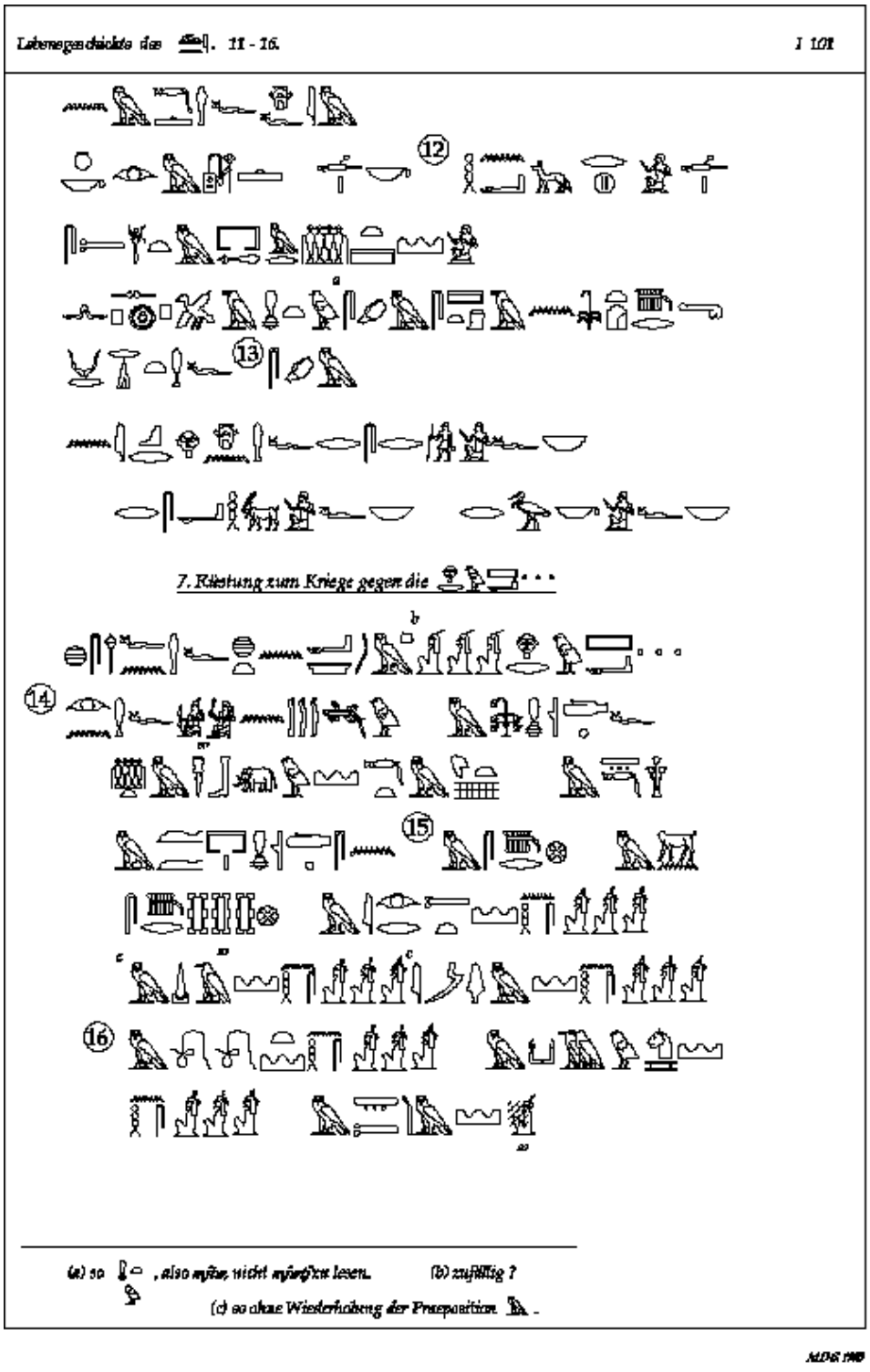


FIG. 4.8 – Même lorsque les ponctuations ne sont pas présentes, la construction des "entre-ponctuations" est réalisable grâce à l'aide de la mise en page. Les unités ainsi définies sont tout simplement les lignes du texte.

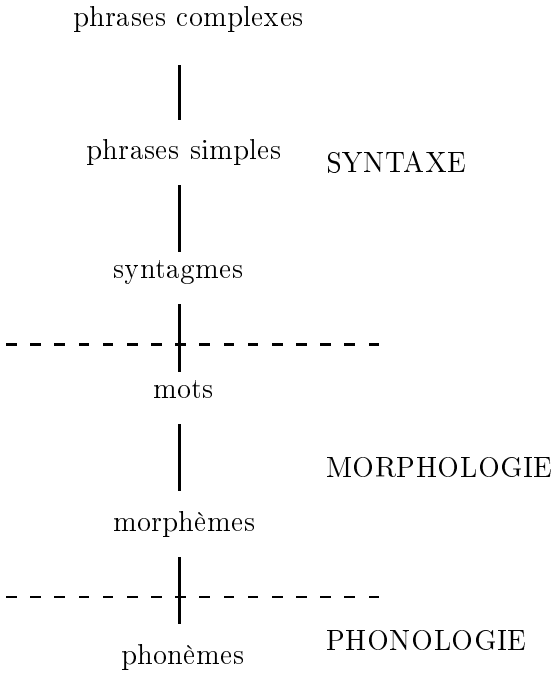
offrons ici notre propre hiérarchie. Comme nous l'avons dit, elle résulte d'une étude multilingue sur corpus. Les critères qui nous ont servi à retenir et à définir les niveaux hiérarchiques sont les suivants :

- les unités ainsi définies sont multilingues, c'est-à-dire que la hiérarchie est applicables à toutes les langues[42].
- Les unités sont définies selon des critères formels.

4.4.1 La hiérarchie classique

Nous avons (et allons) beaucoup utilisé(er) le terme de hiérarchie. Nous définissons une hiérarchie comme étant une organisation de la structure comprenant plusieurs niveaux. Et chaque élément d'un niveau est constitué d'éléments des niveaux inférieurs.

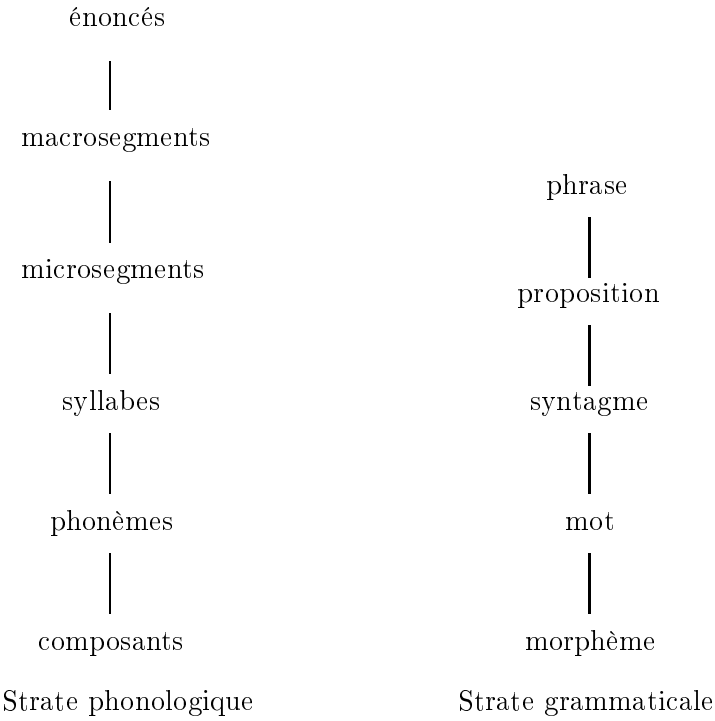
Quelles sont donc ces hiérarchies structurelles manipulées par les linguistes ? La figure 4.4 représente les différents niveaux que l'on trouve dans [Grevisse, 1986]. On y trouve trois grands domaines : la phonologie (qui étudie les phonèmes), la morphologie ([qui] est la science des mots), et la syntaxe ([qui] étudie les relations entre les mots dans la phrase). Les unités les plus communes sont le morphème, le mot, le syntagme (ou groupe), la proposition et la phrase ([Lyons, 1969]).



TAB. 4.4 – La structure classique avec les trois niveaux : phonologique, morphologique, et syntaxique.

[42]Comme d'habitude lire : à toutes les langues que nous avons étudiées.

L'unité de base : le morphème ou le phonème ? Tous les auteurs s'accordent pour définir l'existence d'un élément de base indécomposable[43] à la hiérarchie. Certains ([Harris, 1955]) font commencer la hiérarchie par l'unité appelée le *phonème*. L'unité supérieure, le *morphème*, est donc composée de phonèmes. Il existe donc un procédé pour construire les morphèmes à partir de phonèmes. Pour d'autres ([Hockett, 1961]), ces deux éléments sont des éléments n'appartenant pas à la même hiérarchie. Il est donc impossible de construire les morphèmes à partir des phonèmes. Pour eux, le morphème est l'unité de base de la structure grammaticale, et à ce titre indécomposable.



TAB. 4.5 – Les deux strates structurales proposées par [Hockett, 1961]

Hockett explique cette "erreur" de chercher à décomposer les morphèmes en phonèmes par le fait qu'il existe plusieurs (au moins deux) *strates* ("stratum") dans la langue, et chaque strate possède plusieurs niveaux (figure 4.5). Les deux strates centrales ("inner strata") sont la strate grammaticale ("grammatical stratum")et la strate phonologique ("phonological stratum"). Le morphème est l'unité de base de la strate grammaticale, et le phonème appartient à la strate phonologique. Le fait que l'on cherche une relation de composition entre phonème et morphème provient, selon Hockett, de ce que le phonème soit une unité de taille inférieure[44] au morphème. Or, la relation C de composition[45]

[43] c'est-à-dire qui ne peut s'analyser en terme d'unités plus petites.

[44] Pour s'en rendre compte, il suffit de compter le nombre de phonèmes et de morphèmes dans un énoncé. Le nombre de phonèmes est généralement supérieur au nombre de morphèmes.

[45] The relation C. 'is composed of (an arrangement of)' is the relation that holds between a whole and its part.[Hockett, 1961]

existe entre niveaux d'une même strate et non entre niveaux de deux strates différentes.

Comme nous l'avons vu dans le chapitre 2, selon Harris, une génération de ces morphèmes peut être réalisée automatiquement sans recours au sens. Cependant pour Hockett, le seul résultat possible d'un tel processus est le suivant :

> Beyond this, the procedure will also excise and reveal some, though not necessary all, of the specific clusterings and clumpings of phonemes that constitute part of the evidence for some of the morphemes. [Hockett, 1961, page 46]

Les éléments tels que les morphophonèmes, les morphes (a morph is composed of phonemes, or at least of an arrangement of phonemic material. [Hockett, 1961]), les phones seraient des artéfacts[46] créés pour permettre une correspondance entre strates, et non des éléments du langage (des langues ?). Il est donc clair qu'un titre comme "from phonemes to morphemes" [Harris, 1955] serait revu par Hockett en "from phonemes to morphs", les morphes étant la réalisation concrète (ici écrite) de taille similaire aux morphèmes.

On retrouve aussi chez [Halliday, 1985], [Longacre, 1964] et [Pike, 1967] cette distinction entre grammaire et phonologie[47] (ils ajoutent en plus un troisième élément qui complète la structure des langues : le lexique). Nous partageons ce point de vue, mais nous verrons que la strate grammaticale que nous avons construite est un peu différente des leurs (section 4.4.2).

4.4.2 La hiérarchie construite

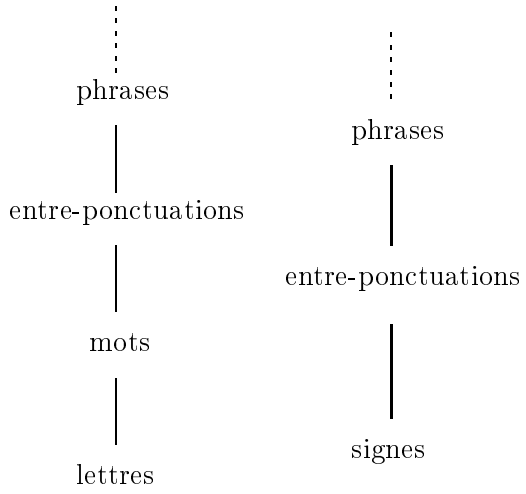
La hiérarchie que nous proposons s'est formée petit à petit, en commençant par l'unité de base. De la même manière que Hockett voyait dans les travaux de ces contemporains une confusion entre la strate phonologique et la strate grammaticale, nous pensons que Hockett a commis aussi une confusion entre deux strates. Sa strate grammaticale est composée des éléments suivants : morphèmes, mots, syntagmes, propositions, phrases (figure 4.5). Nous conservons volontiers les deux strates de Hockett, mais nous en ajoutons une : la strate "écrite", et modifions sa strate grammaticale en lui enlevant le niveau du mot et de la phrase. Hockett semble avoir ignoré la structure écrite, et privilégié la structure phonologique (sonore). Or la strate écrite est similaire à la strate phonologique, même si cette dernière est très antérieure à la première : elles sont toutes les deux un *support physique* de l'information. Selon [Halliday, 1985, page 12], la strate écrite est une *reconstruction* de la strate phonologique, mais les deux sont des modes d'expression des langues :

> Thirdly, however, both writing and speaking are modes of EXPRESSION in language. Writing is in a sense parasitic of speaking; but both function as the REALIZATION of linguistic patterns of a higher level, namely those of GRAMMAR. [Halliday, 1985, page 14-15]

[46] artefact of analysis or convenience for description.

[47] Similarly, attempts to combine grammar and phonology in one complex ste of rules must inevitably result in continued neglect of such units as the syllable and stress group - in spite of the fact that the former is so basic to linguitic structure that most writing systems devised in the ancient Near East were syllabaries. [Longacre, 1964, page 9]

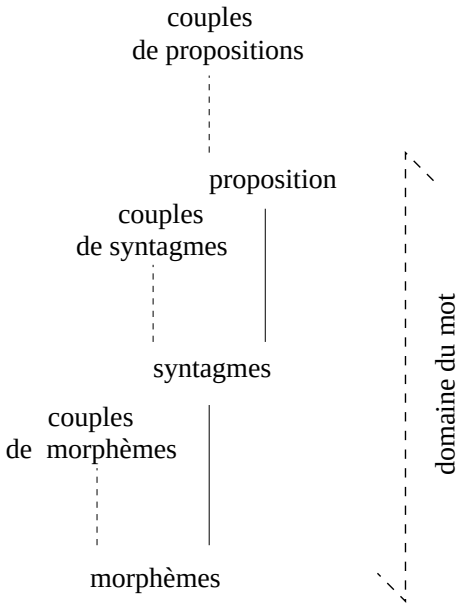
On peut donc voir notre travail comme une méthode permettant de découvrir la strate grammaticale en partant de la strate écrite. Un travail similaire très intéressant serait de partir de la strate phonologique et de construire la strate grammaticale, puis de comparer les deux strates grammaticales obtenues par ces deux chemins en espérant qu'elles coïncident. Cette strate écrite se compose des niveaux suivants : lettres, mots, entre-ponctuations, phrases (figure 4.6). En fait, dans ce travail, les niveaux entre-ponctuations et phrases n'ont pas été distingués. Ce choix remonte au début de ce travail. Nous avions bien vu (nous savions) qu'il existait des différences formelles (utilisation des majuscules) entre les signes de ponctuation. La différenciation entre ces signes nous apporte de l'information supplémentaire (par exemple, cela permet de distinguer immédiatement les fins de propositions en turc, japonais). Mais nous ne savions pas comment utiliser cette information à ce moment du travail. Nous n'avons donc pas pris en compte ces différences. Cela offrait l'avantage de simplifier et d'unifier les traitements informatiques (les corpus étaient réécrits en transformant les majuscules en minuscules). [Halliday, 1985, page 3-6] propose différentes strates écrites en prenant en compte les différents signes de ponctuation et en les hiérarchisant (virgule, point virgule, point). Dans l'écrit, deux unités sont particulièrement utiles pour découvrir la strate grammaticale : le mot et l'entre-ponctuation. Le mot permet un accès au niveau du morphème et du syntagme. L'entre-ponctuation permet un accès au niveau du syntagme et de la proposition. La confusion entre les différentes strates provient du fait que la strate



TAB. 4.6 – La hiérarchie de la strate écrite utilisée pour construire la strate grammaticale pour un système alphabétique et un système idéographique. Les strates écrites sont dépendantes du système d'écriture. Elles peuvent donc être assez nombreuses.

grammaticale n'est pas observable directement et doit être construite en passant par l'intermédiaire des strates observables (écrite, phonologique). Les unités de ces dernières strates sont alors souvent confondues avec les unités de la strate

grammaticale. Le cas caractéristique est celui du mot : unité de la strate écrite et non unité grammaticale. Le mot reflète un niveau de la strate grammaticale qui correspond le plus souvent au niveau du syntagme, mais peut correspondre aussi à bien d'autres niveaux hiérarchiques. La figure 4.7 qui présente notre strate grammaticale, montre la couverture possible d'un mot. Nous voyons qu'il peut aller de l'unité de base, le morphème, jusqu'au niveau propositionnel en passant par les différentes structures syntagmatiques.



TAB. 4.7 – Notre strate grammaticale.

Nous allons maintenant présenter notre strate grammaticale. Comme nous pouvons le voir sur la figure 4.7, notre hiérarchie se compose de trois niveaux de base : le morphème, le syntagme, et la proposition. Puis le syntagme et la proposition peuvent se composer pour former des couples de structures (couples de syntagmes et couples de propositions). Nous aurions pu utiliser le terme *séquence* au lieu de *couple*, mais ce dernier terme semble suffisant pour *décrire* les différentes structures (une séquence de n éléments se décomposent en *n-1* couples) (section 4.8). Le morphème, le syntagme et la proposition sont appelés les *structures élémentaires* de la hiérarchie. Ils sont décrits dans les sections suivantes. Nous allons plutôt nous intéresser aux rapports qui existent entre ces éléments. La hiérarchie est composée de trois niveaux : le niveau morphologique, le niveau syntagmatique et le niveau propositionnel. Dans la suite, nous utiliserons l'adjectif *morphologique* pour désigner le premier niveau, *syntagmatique* pour désigner le deuxième niveau, *propositionnel* pour le troisième niveau et *grammaticale* pour désigner cette hiérarchie. Nous rappelons que les éléments d'un niveau sont construits avec les éléments des niveaux inférieurs. Voyons quels sont les rapports qui existent entre éléments de cette hiérarchie. Un élément X est dit inférieur à un élément Y s'il appartient à un niveau inférieur de la hiérarchie. Le morphème est inférieur au syntagme qui est lui même in-

férieur à la proposition. Nous avons une *relation d'ordre total* entre ces trois éléments, donc le morphème est inférieur à la proposition. On peut utiliser symétriquement le terme supérieur. Ceci est notre premier type de relation entre éléments. Il en existe un deuxième. Nous voyons sur la figure 4.7 qu'il existe aussi d'autres éléments dans la hiérarchie. Ce sont les couples de morphèmes, de syntagmes et couples de propositions. La question qui se pose est de savoir si un couple de syntagmes est supérieur à un syntagme et inférieur à une proposition ? Si cela est le cas, alors la figure serait fausse, car la proposition devrait être reliée au couple de syntagmes et non au syntagme. Pourquoi n'avons nous pas considéré la proposition comme supérieure au couple de syntagmes ? Parce qu'il existe plusieurs critères pour comparer des séquences d'éléments. Pour l'instant, nous avons utilisé le critère que nous appellerons *hiérarchique*. Il existe (au moins) un deuxième critère : le *critère de taille*. On dira alors qu'une séquence est plus *petite* ou plus *grande* qu'une autre. Ce critère ordonne deux séquences d'éléments en comparant le nombre d'éléments de ces deux séquences, les éléments pouvant être le morphème, le syntagme, ou la proposition. Nous pouvons même construire un système où le morphème serait la première unité, le syntagme l'unité des "dizaines" et la proposition l'unité des "centaines". Chaque séquence serait composée d'un certain nombre de morphèmes, de syntagmes, et de propositions. Le problème est de savoir combien de morphèmes font un syntagme et combien de syntagmes font une proposition. La question ne doit pas se poser en ces termes. En fait, l'on possède trois unités de compte : le morphème, le syntagme et la proposition. Et la taille d'une séquence peut être calculée en fonction de ces trois unités de mesure. Il est important de noter qu'une séquence d'éléments d'un niveau hiérarchique donné ne forme pas nécessairement un élément du niveau hiérarchique supérieur. Ainsi, une séquence de morphèmes ne forme pas obligatoirement un syntagme. De même une séquence de syntagmes ne forme pas obligatoirement une proposition (tableau 4.8). *Il existe des règles de construction pour qu'une séquence d'éléments forme une structure supérieure.* Dans notre structure, il existe deux moyens d'organiser une séquence d'éléments : soit l'on organise les éléments pour qu'ils constituent une unité supérieure, soit on organise pour qu'ils constituent une unité plus grande. Ceci explique pourquoi la segmentation systématique en morphèmes n'est pas nécessaire (indispensable) dans notre travail : seule une identification entre marqueurs de frontière et noyau est importante car elle permet de savoir si une séquence de morphèmes forme un syntagme ou non, peu importe le nombre de morphèmes composant le noyau. De plus, la segmentation des morphèmes formant le noyau du syntagme est beaucoup plus délicate que celle des marqueurs de frontière pour une raison majeure : le faible effectif de ceux-ci relativement aux effectifs des morphèmes grammaticaux (marqueurs de frontière) rend leur étude beaucoup plus délicate. L'utilisation des marqueurs étant suffisante pour construire les structures, l'étude des morphèmes nucléaires[48] a été délaissée. Nous reviendrons plus longuement sur ce propos dans la section concernant le lexique (section 5.1). Il existe quand même une corrélation assez forte entre la longueur d'une séquence et son niveau hiérarchique. Un syntagme

[48] appartenant au noyau du syntagme.

Séquences	Morphème	Syntagme	Proposition
de	1	0	0
isationnellement	≈3-4	0	0
anticonstitutionnellement	≈5-7	1	0
dans la banque	≈3-4	1	0
la banque du Japon	≈4-5	2	0
je viens	≈2	1	1
si le mark faiblit	≈5	2	1

TAB. 4.8 – Taille des séquences dans le système MSP (morphème, syntagme, proposition). Une séquence de morphèmes peut être plus longue qu'une proposition (en terme de morphèmes). Le nombre de morphèmes est assez difficile à déterminer (d'où les approximations).

est *en moyenne* plus court qu'une proposition.

Lorsque l'on étudie un niveau de cette hiérarchie, il est très important de se souvenir que ce niveau sert à construire le niveau supérieur de la hiérarchie. Il est bien sûr nécessaire d'étudier particulièrement les règles qui structurent chaque niveau (comme par exemple les règles de construction des syntagmes), mais sans perdre de vue la totalité de la hiérarchie. *Ainsi toutes les sous-classes qui peuvent exister à l'intérieur d'un niveau donné n'existent que parce qu'elles sont pertinentes au niveau supérieur.* Le meilleur exemple est celui du syntagme. Nous n'avons pour l'instant parlé que DU syntagme. Nous allons en fait voir qu'il en existe trois sortes : le syntagme absolu , le syntagme relatif et le syntagme subordonné . Cette distinction ne peut se faire qu'en ayant connaissance des deux niveaux supérieurs au syntagme : la proposition (pour le syntagme absolu) et le couple de syntagmes (pour le syntagme subordonné).

4.5 Le morphème

Le morphème est donc l'unité de base de notre structure grammaticale. Essayons de le définir. Voici quelques définitions :

> [Bloomfield, 1933] : le morphème est une forme linguistique qui ne possède pas de ressemblance phonétique et sémantique partielle avec une autre forme.
>
> [Vendryes, 1923] : [Le morphème est un] élément phonétique qui indique les rapports grammaticaux qui relient les idées entre-elles. (il existe aussi les sémantèmes qui sont les éléments lexicaux)
>
> [Hockett, 1961] : We can easely define 'morpheme' to specify the not-futher-decomposable elements out of which all larger grammatical elements, up to whole sentences (and beyond), are built.

Comme cet élément est l'unité de base de la structure, on ne peut le définir (comme le syntagme ou la proposition) en donnant sa structure puisqu'il n'en

possède pas[49]. La plupart de ces définitions utilisent des critères phonologiques et sémantiques. Cela nous est impossible, et seul un critère formel peut être retenu. Notre point de départ est une liste de mots. Nous avons vu au chapitre 2 comment les morphèmes étaient obtenus grâce à une segmentation de ces mots. Il est donc difficile de donner une définition du type : un morphème est un élément composé, formé par ...Il semble qu'il soit nécessaire d'utiliser le syntagme pour le définir. La définition serait donc : un morphème est un élément qui compose un syntagme[50]. Selon Hockett, l'opération de segmentation ne peut conduire à la génération de la liste des morphèmes, mais à celles des morphes. Le seul critère formel ne peut suffire pour cette génération : il faut lui ajouter un critère sémantique, qui seul permet le passage de la strate phonologique à la strate grammaticale. Nous admettons ce propos, en arguant simplement que la segmentation en morphes est suffisante pour permettre la découverte du reste de la structure grammaticale et que nous faisons un abus de langage en utilisant le terme morphème pour morphe. Mais cela ne nous dit pas quelle est la définition du morphème. Pour définir le morphème, il nous faut revenir au syntagme (section suivante). Le syntagme est composé de deux types de morphèmes : les marqueurs de frontière et les éléments du noyau. Cette dichotomie reprend la dichotomie classique des morphèmes : éléments grammaticaux et lexicaux[51]. Notre segmentation des mots nous permet d'identifier les marqueurs de frontière qui sont liés au noyau, en d'autres termes, les affixes des langues. Notre algorithme de segmentation nous permet d'en identifier certains, mais pas tous. Il semble réellement que le critère formel ne suffise pas dans le cadre d'un recensement exhaustif de ces éléments. Une information sémantique, et étymologique semble nécessaire. Nous tombons ici sur le problème de l'analyse morphologique. Nous sommes donc incapable de donner une définition du morphème autre que :

> un *morphème* est l'élément de base de la structure grammaticale. A ce titre il est indécomposable. Il existe deux types de morphèmes : les marqueurs de frontières (de syntagme et de proposition), et les morphèmes nucléaires qui composent le noyau du syntagme.

4.6 Le syntagme

Notre définition du syntagme est la suivante :

> un *syntagme* est une structure constituée de deux parties : un noyau formé d'un ou d'une séquence de morphèmes, et de marqueurs antéposés et postposés à ce noyau qui sont constitués d'un ou d'une séquence de morphèmes (figure 4.9).

En d'autres mots, *Un syntagme est constitué d'un élément de nature lexicale et de tous les éléments grammaticaux contigus qui dépendent de ce noyau.* Ce noyau est souvent appelé le radical. Les éléments qui sont antéposés au noyau sont

[49] S'il en possédait une, il ne serait pas l'unité de base de la structure.

[50] Il n'y avait donc pas de quoi se moquer des définitions données par les autres auteurs !

[51] La terminologie est assez variée d'un auteur à l'autre pour désigner ces deux types de morphèmes : lexèmes et morphèmes [Vendryes, 1923], sémantèmes et morphèmes [Martinet, 1970],...

FIG. 4.9 – La structure canonique d'un syntagme : un noyau (le radical) auquel sont rajoutés tous les éléments grammaticaux contigus qui dépendent de lui. Les éléments préposés sont considérés comme des marqueurs de début, et les éléments postposés comme des marqueurs de fin du syntagme.

considérés comme des marqueurs de début du syntagme. Les éléments qui sont postposés au noyau sont considérés comme des marqueurs de fin du syntagme. Les affixes sont considérés de la même manière : les préfixes sont considérés comme des marqueurs de début du syntagme, les suffixes comme des marqueurs de fin du syntagme. *La présence des marqueurs de frontière est facultative : un syntagme peut être composé de son seul noyau.* Cette définition est très stable d'une langue à une autre et répond à nos critères : elle ne prend en compte que des critères formels, et est opératoire, c'est-à-dire qu'elle offre un algorithme de segmentation en syntagmes d'un texte (décrit dans la section 6.4.8). Nous appelons cette structure la *structure canonique* d'un syntagme, car, comme allons le voir dans la section 4.8, elle peut subir des modifications. Nous rapprocherons cette définition de celle du *chunk* de [Abney, 1995] :

> We can define a chunk as the parse tree fragments that are left intact after we have unattached problematic elements. It is difficult to define precisely which elements are "problematic".

Les segments ainsi produits sont le plus souvent très proches de nos syntagmes (ou l'inverse), le rattachement des éléments grammaticaux étant assez peu problèmatique. On trouvera aussi dans [Giguet and Vergne, 1997] un analyseur produisant une segmentation en unités qui sont très proches de notre définition. La première référence à une *analyse* d'une séquence en syntagmes (ou chunks) se trouve dans [Longacre, 1960], qui désapprouve la structure des constituants immédiats, très à la mode à cette époque, pour proposer une structure en constituant en chaine ("String constituent") :

> [...] that some linguistic structures are layered while others are ordered like beads on a string.

La composition d'un syntagme Nous avons vu la définition théorique du syntagme. À quoi correspond-elle en pratique ? Le tableau 4.9 offre quelques exemples dans différentes langues. Les langues qui privilégient les marqueurs de début sont généralement appelées langues préposées, et les langues qui privilégient les marqueurs de fin sont appelées langues postposées.

Nous allons maintenant regarder en détail les deux parties qui composent un syntagme : le noyau et les marqueurs de frontière. Nous dirons peu de choses du noyau, car au début de ce travail son étude a été considérée comme inutile pour nos besoins. Cette vision des choses a été revue, et une étude plus approfondie des informations lexicales est développée au chapitre 5. La deuxième partie du syntagme est composée des marqueurs de frontière. Ce sont ces marqueurs qui

Langues	début	*noyau*	fin
Français	dans toutes les	*opér*	-ations
Anglais	I	*let*	him off
Allemand	in die	*Grenz*	-en
Swahili	na kile ki-	*tamba*	-a
Turc	bir	*süre*	için
Vietnamien	trong moät	*hoaøn*	

TAB. 4.9 – Exemple de syntagmes dans différentes langues. Les affixes (indiqués par un tiret) sont aussi vus comme des marqueurs de frontière.

nous ont permis de retenir et de définir formellement cette notion de syntagme. La sélection de cette structure a été facilitée par le fait que certains mots de la langue ont la particularité de n'être (pratiquement) que des marqueurs de frontière. Ils sont donc facilement identifiables grâce à leur comportement positionnel. Le tableau 4.10 en montre quelques uns.

Langues :

Français

	Effectif	Début	Fin
de	14943	648	0
la	8427	1300	0
les	5382	562	0

Allemand

	Effectif	Début	Fin
die	2944	701	4
in	1566	241	0
von	1242	122	0

Swahili

	Effectif	Début	Fin
ya	3704	27	0
kwa	3318	601	0
ni	1370	200	0

TAB. 4.10 – Marqueurs de début caractéristiques de syntagme dans plusieurs langues.

Comme expliqué à la section 4.2, ces éléments nous ont servi à segmenter le texte. Cette notion de marqueurs de début et de fin a été introduite parce que l'effectif n'était pas un critère suffisant pour permettre une mise en relation (section 1.6). Ils ont la particularité d'être toujours en relation avec un élément donné (suivant pour les marqueurs de début et précédant pour les marqueurs de fin) quel que soit l'effectif des autres éléments environnants. Ces marqueurs de frontière correspondent généralement aux traditionnels déterminants des langues (article, adjectif possessif, démonstratif, ...) et aux prépositions ou postpositions ainsi qu'aux différents affixes.

Si l'on étudie la structure interne d'un syntagme, c'est-à-dire savoir quelles sont les règles auxquelles les éléments du syntagme obéissent, nous en trouvons trois qui sont particulièrement intéressantes. Nous parlons de *règle*, mais le terme *tendance* serait peut être plus adéquat, puisqu'il existe toujours des exceptions à celles-ci. Premièrement, les noyaux sont rarement coupés par les marqueurs[52]. La deuxième règle concerne l'ordre linéaire des marqueurs de frontière. On peut les catégoriser en deux : les éléments qui ont un rôle fonctionnel (c'est-à-dire qui jouent un rôle dans la structure supérieure à laquelle appartient le syntagme) et les éléments non fonctionnels. L'on peut diviser la zone des marqueurs de frontière en deux : contiguë au noyau nous trouvons la *zone interne*, puis la *zone relationnelle* qui contient les éléments fonctionnels du syntagme (figure 4.10). Les éléments que l'on trouve dans la zone relationnelle correspondent typiquement aux pré(post)positions, les éléments de la zone interne aux déterminants. La générationde la liste des morphèmes appartenant à la zone relationnelle est en général plus facile. Ainsi les éléments qui jouent une fonction dans la mise en relation entre syntagmes sont plus facilement "disponibles", "accessibles" pour les autres syntagmes qui en auraient besoin. La troisième règle concerne les

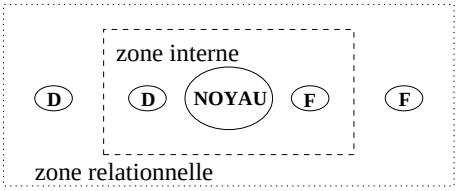


FIG. 4.10 – Les marqueurs de frontière de syntagmes qui marquent les relations entre syntagmes se rencontrent dans la zone périphérique du syntagme.

éléments qui composent ces marqueurs de frontière. Ils semblent être majoritairement utilisés pour ce rôle. Si nous observons leur répartition dans le corpus (tableau 4.11), leur utilisation principale est bien celle de marqueurs de frontière. Ils peuvent bien sûr se retrouver dans un noyau de syntagme, mais cela arrive beaucoup plus rarement. Cette observation est valable pour les marqueurs de frontière correspondant aux mots mais aussi pour les affixes (par exemple *ion* n'est utilisé que 7 fois dans notre corpus comme début de mots (*ion*, *ionas*, et *ionesco*) contre 5733 utilisations en fin de mots). Comme nous le voyons avec le mot *le*, cette caractéristique n'est pas valable pour tous les marqueurs de début. Cette règle est encore plus pertinente lorsqu'il s'agit de séquences de marqueurs (ou de marqueurs bissyllabiques). Par exemple, le couple de marqueurs de frontière le plus fréquent en français est *de la* (2423 occurrences). Il existe seulement 15 occurrences des séquences *dela* où ni *de*, ni *la* ne sont marqueurs de début, 944 où *de* est marqueur de début (*la* est une partie commençant le noyau), et 5 où *la* est marqueur de début (*de* est une fin de noyau). Nous voyons donc que la séquence *de la* (avec ou sans espace) correspond à un début de syntagme à plus de 99%. La prise en compte de considérations phonologiques serait intéressante (par exemple, le mot *les* /*lɛ*/ ne se prononce pas comme la séquence finale *-les*

[52]Les infixes sont assez rares même s'ils existent. Ils ne semblent pas jouer de rôle relationnel.

	Effectif dans le corpus			
	de la chaîne de caractères	du mot	en début de mot	en fin de mot
de	25748	14943	7375	1350
des	5278	4750	174	285
la	12450	8427	702	203
le	23550	6504	6580	5068
les	7384	5882	52	1820
et	9091	5311	207	904
ion	8729	7	2	5733
ique	2827	0	0	1895
ment	4642	0	29	3755

TAB. 4.11 – Peu de mots dans un corpus finissent par des séquences correspondant aux marqueurs de début fréquents. Il en est de même pour les marqueurs de fin : peu de mots commencent par les préfixes les plus courants.

dans tab*les.*), mais notre travail portant sur l'écrit, nous laissons ce travail à d'autres (ou à plus tard).

Le syntagme étant une structure assez simple, la couverture des structures syntagmatiques de la langue étudiée est très grande[53]. c'est-à-dire qu'un corpus de 50,000 mots permet une très bonne connaissance des structures syntagmatiques[54].

Les différents types de syntagmes Nous avons pour l'instant parlé simplement du syntagme. Existe-t-il un seul type de syntagme ? La réponse à cette question est donnée en considérant les structures composées de syntagmes. Ces structures sont décrites dans les sections suivantes. Nous allons voir qu'il existe trois types de syntagmes. La structure propositionnelle met en évidence le Syntagme Absolu (SA) . Les structures de syntagmes mettent en évidence deux autres syntagmes : le Syntagme Relatif (SR) et le Syntagme Subordonné (SSub). La partition du syntagme en trois types ne peut se faire au niveau syntagmatique. Elle nécessite la connaissance des structures supérieures. Les caractéristiques de ces trois syntagmes sont expliquées aux sections 4.7.2 pour le SA, et 4.8.2 pour le SR et le SSub. C'est essentiellement leur différence fonctionnelle qui permet cette catégorisation, et non pas une différence morphologique, même si cette différence fonctionnelle s'accompagne de différences morphologiques. On notera que les différentes catégories de morphèmes (marqueur de frontière et noyau) sont obtenues de la même manière : en observant la fonction de ceux-ci dans l'unité supérieure qu'est le syntagme.

Marqueur de début et de fin La construction des syntagmes est facilitée par un fait : les marqueurs de début (de fin) ne jouent généralement pas en même

[53] Cette affirmation est difficilement quantifiable, puisque il n'existe pas de recensement de ces structures (au moins sur corpus), et que ce recensement n'a pas été effectué durant ce travail.

[54] voir l'évaluation du travail dans la section 6.5.

temps le rôle de marqueurs de fin (de début). Ainsi, en français, une préposition n'indique jamais[55] la fin de son syntagme. Ce propos n'est généralement pas vrai pour toutes les constructions, en particulier pour les syntagmes absolus (vous pouvez lire pour l'instant syntagmes verbaux). Les marqueurs de frontière de ce type de syntagme peuvent assez souvent indiquer le début ou la fin du syntagme (tableau 4.12).

Langues	Début	Noyau	Fin
Vietnamien	*hoï*	hoûi	oâng
	oâng	hoûi	*hoï*

	Début	Noyau	Fin
Allemand	ich	kann	*es*
	es	kann	dir

TAB. 4.12 – Dans un syntagme absolu, un marqueur de début (*hoï*, *es*) peut se trouver marqueur de fin.

Ce cas peut s'expliquer par le fait que les syntagmes absolus jouent un rôle particulier dans la structure propositionnelle, ce qui les différencie nettement des syntagmes relatifs. En ce qui concerne les syntagme relatifs (lisez syntagmes nominaux pour l'instant), le cas existe aussi mais est beaucoup plus rare. Il concerne généralement un marqueur de début d'un certain type de syntagme et un marqueur de fin d'un autre type de syntagme (ou de proposition). L'anglais illustre parfaitement ce cas avec certains éléments (comme *in*) qui jouent le rôle de marqueur de début de syntagme relatif et de marqueur de fin de syntagme absolu :

- *even when his aunt came in*,
- *In the course of it aunt polly said :*
- *But an unforeseen phenomenon came in to subject the public impatience to a severe trial.*

Dans la première séquence, le mot *in* est un marqueur de fin de syntagme absolu, dans la deuxième, un marqueur de début de syntagme relatif (la ponctuation nous offre un bon critère de décision). Dans le troisième cas, le problème se pose. Est ce que *in* appartient au syntagme *came* ou au syntagme *to the subject* ? Deux segmentations sont alors en concurrence. Notre méthode permet de mettre à jour de telle situation conflictuelle, puisque *in* se trouve catégorisé dans deux catégories au comportement opposé (section 3.3.1). Nous pouvons identifier ce double emploi, mais il est plus difficile d'assigner une catégorie à toutes ces occurrences de *in*.

[55] Tellement peu souvent.

4.7 La proposition

Nous allons maintenant décrire le deuxième niveau de notre hiérarchie. Il s'agit de la *proposition*. Avant de donner notre définition de cette structure, il nous faut d'abord introduire quelques considérations, ce niveau étant plus complexe que le niveau syntagmatique.

Pourquoi un niveau propositionnel ? Pourquoi introduire un niveau supplémentaire au dessus du niveau syntagmatique ? Pendant assez longtemps, nous avons travaillé avec le niveau syntagmatique, croyant que cela était suffisant. Mais nous avons été confronté à plusieurs problèmes. Une fois ces syntagmes construits (plus ou moins bien), nous avons essayé de les mettre en relation. Sans succès. Par exemple, il était très difficile de différencier, en français, une relation entre un substantif et son adjectif et entre un substantif en fonction sujet et son verbe. En fait une question se posait : fallait-il essayer de trouver une différence entre ces deux relations ? Bien sûr, il est facile de mettre au point une méthode qui permette une telle différenciation, mais ad hoc pour le français, et qui ne s'appliquait donc pas (ou très mal) aux autres langues. C'est en fait en travaillant sur ces autres langues que nous avons introduit le niveau de la proposition. En particulier, en travaillant sur l'allemand et le turc où le niveau propositionnel est très fortement marqué. Nous voyons là un exemple des bienfaits de l'étude multilingue.

En travaillant sur des langues où une structure est très fortement marquée, et dont, en général, la manipulation est indispensable pour bien traiter la langue en question, nous intégrons cette structure dans notre hiérarchie, avec généralement de très bonnes retombées sur les autres langues.

Ainsi la compréhension du niveau propositionnel en allemand est indispensable pour un traitement correct de cette langue. Qui plus est, cette structure est très bien marquée. La structure propositionnelle du français étant moins marquée, sa mise en évidence a été plus difficile. Mais le transfert des concepts formels de l'allemand vers le français a été très fructueux. Il en est de même pour le niveau du syntagme (section 7.4).

Pourquoi disons nous que le niveau propositionnel est indispensable ? Prenons un exemple en allemand. Soit la séquence suivante :

Du gibst also die Waffen ab.

Si nous restons au niveau syntagmatique, le mot *ab* est analysé comme un marqueur de fin caractéristique (effectif :94, début :0, fin : 69). Il correspond donc à un marqueur de fin de syntagme (puisque c'est la seule structure connue). La construction des syntagmes de la phrase produit donc :

[*Du gibst also*] [*die Waffen ab*].

Mais le mot *ab* ne partage pas les caractéristiques des autres marqueurs syntagmatiques. D'une part, ces syntagmes apparaissent très souvent avant une ponctuation (trois fois sur quatre). D'autres part, les seuls syntagmes pouvant apparaître après un syntagme finissant par *ab* possèdent une caractéristique singulière : 96% commencent par *und* comme :

Sie gingen ab ***und*** *ich folgte ihnen.*

Nous voyons donc que cet élément n'est pas distributionnellement similaire aux autres marqueurs de frontière de syntagme (qui imposent peu de contrainte sur le syntagme suivant). L'introduction d'un niveau supérieur qui est la proposition est une réponse qui permet de réinterpréter le comportement de cet élément. D'autres solutions auraient pu être envisagées (en particulier définir d'autres classes de marqueurs de frontière), mais celle-ci semblait la plus intéressante. Le fait principal qui nous a conduit à introduire la proposition est le suivant : tous les éléments qui partagent ces caractéristiques étaient de nature propositionnelle (conjonctions, morphèmes verbaux, particules verbales).

Nous allons maintenant voir quelles sont les marques formelles qui caractérisent la proposition. La proposition possède des marqueurs de frontière qui sont de deux types : des éléments du niveau morphologique et des éléments du niveau syntagmatique. Comme il a été dit à la section 4.2, la proposition étant d'un niveau supérieur aux morphèmes et aux syntagmes, ces deux derniers niveaux peuvent être utilisés pour marquer les frontières de la proposition.

4.7.1 Les marqueurs morphologiques

Nous allons d'abord nous intéresser aux marqueurs de frontière morphologiques. Le principe est identique aux marqueurs de frontière de syntagmes : certains éléments, mots ou morphèmes liés, indiquent le début ou la fin d'une proposition. Leur caractéristique est assez similaire aux marqueurs de frontière de syntagmes. Les marqueurs de début ne se rencontrent pas avant une ponctuation (et vice versa pour les fins), comme le montre le tableau 4.13. Mais ils

Langues	Morphèmes	Effectif	Début	Fin
français	mais	845	694 (82%)	9
	car	127	125 (98%)	5
allemand	daß	1251	1169 (93%)	0
	sondern	127	125 (98%)	0
	her	65	0	40 (61%)
	zurück	168	4	139 (82%)
turc	ama	763	743 (97%)	4
	çünkü	659	648 (98%)	1
	-dı	445	7	414 (93%)
	-im	570	76	303 (53%)
swahili	lakini	1133	1027(90%)	73
	bali	223	201(90%)	3
vietnamien	thì	809	516 (63%)	4
	nhöng	409	387 (94%)	1
latin	tunc	35	19 (54%)	0
	at	84	53 (63%)	0

TAB. 4.13 – Des marqueurs morphologiques caractéristiques de début et fin de proposition.

possèdent une caractéristique supplémentaire. Prenons les marqueurs de début :

non seulement ils n'apparaissent pas à la fin des entre-ponctuations (caractéristique des débuts syntagmatiques), mais ils apparaissent essentiellement en début de ces séquences. Ceci est simplement une conséquence de la taille des propositions. Les entre-ponctuations sont le plus souvent composées de séquences de syntagmes. Les propositions étant composées de syntagmes, les débuts de syntagmes se rencontrent le plus souvent à "l'intérieur" des entre-ponctuations (figure 4.11). Par contre, les entre-ponctuations étant plus rarement composées de séquences de propositions, les marqueurs de frontière de proposition se rencontrent plus rarement à l'intérieur des entre-ponctuations, donc plus souvent en début et fin de ces séquences.

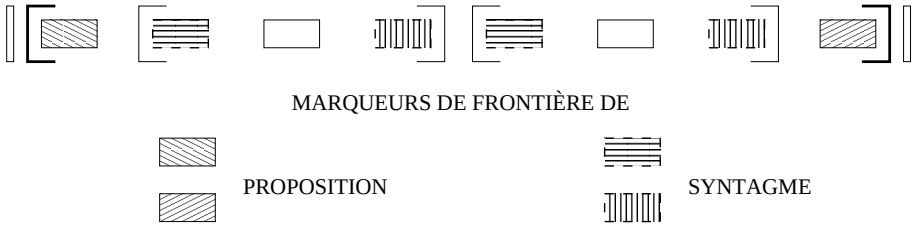


FIG. 4.11 – Les marqueurs de frontière de syntagmes se rencontrent plus souvent à l'intérieur des entre-ponctuations que les marqueurs de frontière de proposition.

4.7.2 Les marqueurs syntagmatiques : le Syntagme Absolu

Le deuxième type de marqueurs de frontière propositionnels correspond à ce que nous avons appelé les *Syntagmes Absolus* (SA). Ce sont des syntagmes qui partagent la même propriété que les marqueurs morphologiques : ils apparaissent très souvent en début (ou en fin) d'entre-ponctuations. Ils possèdent donc une caractéristique positionnelle très caractéristique qui ne se retrouvent pas dans les autres types de syntagmes. Les tableaux 4.14 donnent quelques exemples de structures caractéristiques. Nous voyons bien que ces structures se recontrent essentiellement en début ou en fin d'entre-ponctuations, d'où leur nom de syntagme *absolu*, leur position étant très contrainte. Cette terminologie provient de l'étude du turc où le groupe verbal est un élément postposé de la proposition, et donc apparaît en fin de *phrase*. Nous avons étendu cette terminologie aux autres langues puisque l'on y trouve aussi de telles structures. Elles correspondent le plus souvent à un modèle *pronom sujet + verbe* ou *conjonction + verbe*. De même que pour les marqueurs morphologiques, tous les syntagmes absolus ne sont pas marqués positionnellement, il existe des structures caractéristiques qui vont permettre l'amorçage de la génération de tous les SA (section 6.3). Nous appelerons pour l'instant syntagmes relatifs les syntagmes qui ne sont pas des SA.

Ces traces du niveau propositionnel sont donc accessibles directement, sans qu'une construction du niveau syntagmatique soit nécessaire. Alors qu'il est couramment admis ([Powers and Daelemans, 1992, page 143]) que dans une classification du type "bottom' up", le niveau n doit être construit avant de passer au

	Syntagme	Effectif	Début	Fin	Début et fin
Français	il N-ait	249	171 (68%)	11	6
	nous N-ons	191	109 (57%)	3	2
	je N-e	134	96 (72%)	3	3
	elle N-ait	61	42 (69%)	3	2

	Syntagme	Effectif	Début	Fin	Début et fin
Swahili	mimi ni-N	120	73 (60%)	14	7
	yeye a-N	167	81 (48%)	42	15
	ninyi m-N	179	74 (41%)	34	10
	wewe u-N	63	36 (57%)	19	7

TAB. 4.14 – Position de Syntagmes Absolus (SA) en français et swahili. Ils apparaissent majoritairement en début (ou en fin) d'entre-ponctuations.

niveau *n+1*, la construction du niveau propositionnel peut et doit se faire, si ce n'est avant, au moins en même temps que la construction du niveau syntagmatique.

4.7.3 La définition de la proposition

Après avoir décrit les marqueurs de frontière de la proposition, nous allons en donner une définition :

> Une *proposition* est composée d'un syntagme absolu ou d'une séquence de syntagmes comprenant un seul syntagme absolu ou une séquence de syntagmes absolus entretenant une relation de dépendance. Ses débuts et Ses fins sont marqués par des éléments de nature morphémique ou syntagmatique.

Voici quelques exemples d'entre-ponctuations (cela aurait pu être des séquences de mots) extraites des corpus *français01* et *allemand01* qui forment des propositions : (les syntagmes absolus sont en gras et délimités par un rectangle)

1. L'unité **employait alors** cent dix salariés.
2. **Qui lit** dans un texte ?
3. , **qui a gardé** des traits d'adolescent en dépit d'une taille de géant,
4. **Il n'empêche**.
5. **Er hat dir** einen Mund **gegeben**.
6. , daß er von euch auch eine Vergütung der Überraschung **verlangt**,
7. **Ich weiß es nicht**.

Les exemples 1 et 2 sont canoniques (mais rares !) : nous avons une entre-ponctuation (une phrase) qui possède un seul syntagme absolu. Les exemples 3 (français) et 6 (allemand) contiennent un seul syntagme absolu qui ne dépend d'aucun autre syntagme de la séquence. Les exemple 4 (français) et 7 (allemand) correspondent à une proposition composée d'un seul syntagme absolu. L'exemple 5 est déjà plus complexe : il possède deux SA, avec le premier (*Er hat dir*) en relation avec le dernier (*gegeben*). Cette dépendance provient du fait que le dernier SA nécessite la présence du premier. Les exemples suivants ne sont pas des propositions simples :

8 , sous l'influence parfois décisive de la majorité elle-même.

9 , après l'incendie **qui a détruit** 3 800 hectares de forêt entre le Porge et Lacanau,

L'exemple 8 ne possédant pas de syntagme absolu, il ne forme pas une proposition, bien qu'étant une séquence de syntagmes. L'exemple 9 possède bien un syntagme absolu, mais qui dépend d'un syntagme relatif (*après l'incendie*). Cette séquence n'est donc pas une proposition mais elle en contient une. Par contre la séquence ***qui a détruit** 3 800 hectares de forêt entre le Porge et Lacanau,* en est une (similaire à l'exemple 3). En termes "classiques", nous pouvons donc voir la proposition comme étant composée d'un verbe et de tous les syntagmes qui dépendent de lui (on retrouve la définition classique).

De même que la segmentation en syntagmes présente parfois certains problèmes, il en est de même pour la proposition. Dans un énoncé comme :

J'entends les oiseaux **chanter**[56].

si le syntagme *chanter* dépend de *les oiseaux*, nous avons deux propositions (les séquences *j'entends* et *les oiseaux chanter*). S'il est dépendant de *J'entends*, alors nous avons une seule proposition avec deux syntagmes absolus en relation. Nous retrouvons le même problème que celui décrit au paragraphe *Marqueur de début et de fin* de la section 4.6 au niveau du syntagme. Nous reviendrons plus longuement sur ce problème dans la section 4.9.2. Dans une langue comme le français, la segmentation en propositions est plus délicate que dans une langue comme l'allemand, où le niveau propositionnel est assez fortement marqué.

Il est clair que la mise au point de la définition de la proposition ne s'est pas uniquement basée sur des critères formels. Notre connaissance du français et des autres langues ainsi que nos *a priori* comme la notion classique de la proposition ont grandement participé à l'élaboration de la définition. Il n'en reste pas moins que l'introduction de cette structure nous a semblé nécessaire afin de pouvoir réaliser une segmentation des textes en syntagmes, en particulier pour la gestion des marqueurs de frontière propositionnels.

La structure d'une proposition Nous allons maintenant observer en détail la structure d'une proposition. Nous avons vu qu'elle était composée d'au moins un syntagme absolu . Mais elle comprend aussi des syntagmes relatifs. Nous allons étudier les différentes constructions possibles entre ces syntagmes relatifs et

[56] Énoncé extrait de [Grevisse, 1969]. Nous n'avons trouvé aucune structure similaire dans notre corpus *français01*.

le syntagme absolu. Pour cela, nous allons revenir au schéma théorique complet de la proposition (figure 4.12). Dans cette figure, la structure est dite complète

FIG. 4.12 – Le schéma complet des marqueurs de proposition. Les éléments grisés marquent les éléments caractéristiques d'une proposition.

car les deux types de marqueurs de frontière sont représentés : morphologique et syntagmatique. Le début ou la fin peuvent être marqués par un syntagme absolu. On remarque dans ce cas que la morphologie de ces deux syntagmes, s'ils existent dans une même langue, est assez différente. Nous parlerons de Syntagme Absolu de Début (SAD) , et de Syntagme Absolu de Fin (SAF) pour distinguer ces deux types de SA. Il existe en fait peu de langues qui utilisent un tel schéma complet. Le cas le plus complet rencontré est celui de la proposition allemande où les marqueurs de fin morphologiques et syntagmatiques sont mutuellement exclusifs[57]. Il existe une typologie des langues qui utilisent la structure propositionnelle comme critère de classification. Cette classification utilise la position de trois constituants de la proposition : le verbe (notre SA), et les deux actants principaux de la proposition désignés par le terme de Sujet (S) et Objet (O). On trouvera dans [Hagège, 1982] une typologie des langues qui utilisent ces différentes structures. Comme nous pouvons le voir, ces structures peuvent être vues comme la manière d'ajouter des éléments (les actants décrits plus bas) au squelette de la proposition que sont les marqueurs de frontière et le Syntagme Absolu. Pour ajouter les autres éléments de la proposition (les SR), il existe plusieurs possibilités. La première consiste à ajouter ces éléments à gauche et à droite du SA. Cela donne la structure XVX[58] (SVO ou OVS). Pour cette structure, nous voyons que les Syntagmes Absolus n'occupent plus

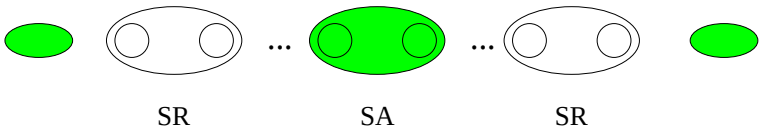


FIG. 4.13 – La structure dite SVO ou OVS, rencontré en français, anglais.

une position "absolue" dans la proposition. Elle va dépendre du nombre de SR utilisés entre le début et la fin de la proposition et le SA. Mais la réalisation de la position absolue du SA est obtenue pour certaines structures, même dans ces langues. La principale étant la structure *Pronom sujet* + *Verbe*. Dans ce

[57] On a :
- **und ich habe dich** seit gestern **nicht gesehen**.
- **wir nahmen ihm** dabei die Waffen **ab**.

Mais on ne peut avoir une combinaison du type : *[. . .] gesehen ab.*

[58] X=S|0.

cas là, le SA devient un SAD (nous n'avons pas rencontré de langues où un SA devenait SAF). Comme nous l'avons dit, ce sont ces structures sur lesquelles nous allons nous appuyer pour découvrir les SA dans ces langues.

Une deuxième solution consiste à ne jamais pouvoir intercaler de SR entre les marqueurs de début et le SA. On a alors la structure VXX (VSO, VOS). La

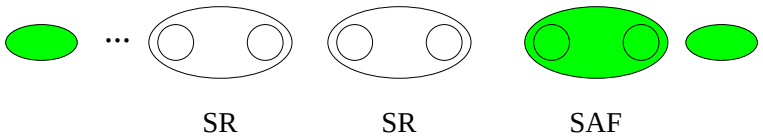


FIG. 4.14 – La structure dite SOV ou OSV, rencontrée en turc et japonais.

troisième solution, symétrique à la deuxième, consiste à ne jamais intercaler de SR entre la SA et les marqueurs de fin (structure XXV). Tous ces types peuvent

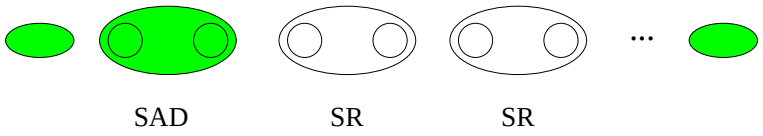


FIG. 4.15 – La structure dite VSO ou VOS, rencontrée dans les langues sémitiques.

se trouver dans une même langue. Il existe en fait une dissymétrie entre les structures VXX et XVX d'une part et XXV d'autre part. La structure XXV admet toujours des marqueurs de début morphologiques (comme toutes les constructions propositionnelles), alors que les structures VXX et XVX n'admettent que très rarement des marqueurs de fin. Il semble donc que le marquage des débuts de propositions soit privilégié par rapport au marquage des fins.

les différents types de propositions De même qu'il existe plusieurs types de syntagmes, il existe aussi plusieurs types de propositions. Le critère retenu est celui de la dépendance de la proposition. Les trois types de propositions retenus sont :

- la proposition indépendante
- la proposition subordonnée à une proposition
- la proposition subordonnée à un syntagme

La proposition indépendante ne dépend d'aucune autre proposition ni syntagme. Les deux autres types de proposition dépendent soit d'une proposition soit d'un syntagme. Les propositions dépendant d'un syntagme peuvent être à leur tour discriminées selon la nature du syntagme (absolu, relatif ou subordonné). Nous rappelons que la catégorisation d'une unité ne se base pas sur des considérations formelles intrinsèques, mais sur le rôle (la fonction) qu'elle joue dans des structures l'incluant. Il existe bien une relation entre le type d'une unité et sa forme (sa composition formelle), mais ce critère n'est pas assez fiable à lui seul.

Les actants Parlons maintenant des Syntagmes Relatifs que nous rencontrons dans une proposition : ce sont les *actants*. Un actant est un syntagme (ou une séquence de syntagmes) qui dépend du Syntagme Absolu de la proposition. Nous reprenons ici la terminologie utilisée dans [Tesnière, 1959] bien que sa définition ne soit pas formelle[59]. La notion d'actant permet de se débarrasser de la notion de sujet, d'objet,..., qui porte une connotation "sémantique". La séquence de SR formant un actant peut elle même constituer une proposition. La nature des actants peut donc être syntagmatique ou propositionnelle. Les actants sont caractérisés par un numéro d'ordre (prime, second, tiers actant) qui correspond simplement à la fréquence de ces structures dans une proposition (le prime actant est plus fréquent que le second, le second que le tiers). Ces différents actants possèdent généralement des marques formelles (positionnelles ou morphologiques) qui permettent de les différencier. Ils correspondent formellement à une séquence de syntagmes en relation. Les langues possèdent des marqueurs plus ou moins spécifiques pour indiquer le rôle actanciel d'un SR. Dans certaines langues, certains actants vont être très faciles à identifier (le second actant en turc, le prime en japonais) car ils sont marqués par des marques (dites casuelles) très spécifiques à cette relation. L'identification de ces actants se fait en construisant les couples de syntagmes dont un syntagme est un Syntagme Absolu (section 4.8.3). La recherche de ces structures actancielles est aidée par le fait qu'une proposition ne peut posséder qu'un seul prime actant, second actant,.... Ainsi deux séquences de syntagmes d'une proposition ne peuvent correspondre à deux primes actants d'une proposition[60]. De plus, il semble que les actants ne peuvent être constitués de syntagmes discontigus (hypothèse à vérifier). Voici donc un ajout à notre définition de la proposition canonique :

> Le SA d'une proposition possède des actants constitués de SR. Ces SR peuvent former eux-mêmes une proposition. Une proposition ne peut avoir plus d'un actant de même type.

La structure du syntagme a été beaucoup mieux étudiée que celle de la proposition, car "d'accès" plus immédiat. Il reste beaucoup à faire au niveau de la proposition. L'étude du niveau syntagmatique (en particulier des différents types de syntagmes) a été possible grâce à la connaissance du niveau supérieur (la proposition). Si l'on veut suivre la même démarche (mettre à jour les différents types de propositions), il est alors nécessaire de trouver le niveau supérieur à la proposition pour pouvoir appréhender complètement cette dernière. On trouvera des descriptions du niveau de la proposition (entendre souvent la phrase simple) dans de nombreux ouvrages [Benveniste, 1966], [Chomsky, 1969a], [Lyons, 1969]. On notera que la structure de la proposition décrite ici ne reprend pas le découpage de la proposition en sujet et prédicat décrite dans [Arnauld and Lancelot, 1660] et (donc) plus récemment dans [Chomsky, 1969a] (le fameux $S \rightarrow NP + VP$).

[59] Les actants sont les êtres ou les choses qui, à un titre quelconque et de quelque façon que ce soit, même à un titre de simples figurants et de la façon la plus passive, participent au procès. [Tesnière, 1959, page 102]

[60] Mise à part le cas de la coordination.

4.8 Les structures composées

Nous avons pour l'instant décrit les structures "canoniques" ou "simples" de la hiérarchie que sont le syntagme et la proposition. Ces deux structures ne suffisent pas pour décrire tous les énoncés trouvés dans un corpus. Nous allons voir comment elles peuvent se combiner entre elles pour former des structures composées. Dans les exemple suivants, le type de relation considéré entre les deux éléments est celui de la *relation de dépendance*. Les autres types de relation sont expliqués à la section 4.10.

4.8.1 Les opérations de composition

Nous allons d'abord voir quelles sont les manières de composer deux éléments (morphème, syntagme, proposition) linéaires, c'est-à-dire la façon dont deux éléments en relation se positionnent l'un par rapport à l'autre. Il en existe deux principales : la composition externe et la composition interne. La composition externe correspond simplement à une juxtaposition des deux éléments (exemple 1 de la figure 4.16). Il conserve la contiguïté des éléments de chacun des deux éléments. Comme nous le verrons (section 4.10), ces deux éléments peuvent être eux-mêmes contigus ou discontigus (d'autres éléments peuvent s'intercaler entre eux). Ce type de composition correspond, par exemple, à la structure française *substantif-complément du nom*.

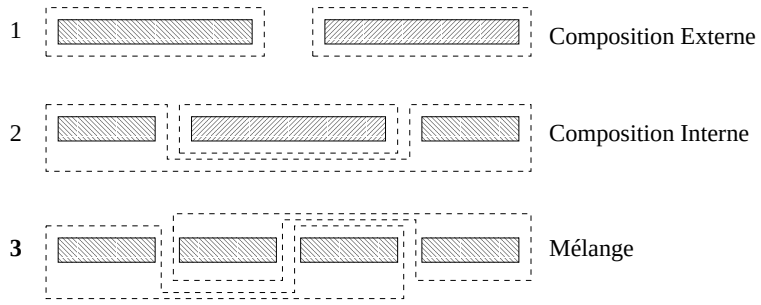


FIG. 4.16 – Les deux compositions : la composition externe (1) et la composition interne (2).

Le deuxième type de composition est la composition interne (exemple 2 de la figure 4.16). Dans ce cas, un élément est "cassé" pour permettre l'insertion du deuxième élément. Nous voyons donc que l'élément cassé n'est plus une suite d'éléments contigus. Un exemple de ces compositions a déjà été donné dans la section 3.3.2 : ce que nous avons appelé les structures d'accord interne. Ce type de composition correspond à la structure française *substantif-adjectif antéposé*. Par exemple la séquence :*la* $\boxed{\textit{petite}}$ *maison*, où le syntagme *la maison* est devenu discontigu. Une caractéristique importante concerne l'élément qui est inséré : il ne peut être d'un niveau hiérarchique supérieur au deuxième élément[61]. Dans une composition interne, l'élément inséré est toujours en relation

[61] La composition interne est à différencier de la construction appelée *incise* que l'on trouve

avec l'élément receveur, alors que deux syntagmes contigus (même configuration qu'une composition externe) ne le sont pas systématiquement. Il existe en théorie une troisième sorte de composition : les deux éléments sont cassés (cas 3 : Mélange). On a donc un mélange complet des éléments des deux structures. Ce cas n'a pas été rencontré dans les langues étudiées.

4.8.2 Les structures de syntagmes

À partir de ces deux opérations de composition, nous allons construire les structures composées de deux syntagmes. Nous avons recensé pour l'instant deux types de syntagme : le syntagme relatif (SR) et le syntagme absolu (SA)[62]. Commençons par la composition interne. Ce sont donc des structures où un syntagme est inséré dans un autre syntagme. Cette structure est illustrée par la construction allemande *substantif-adjectif antéposé*. L'insertion est généralement effectuée au niveau de la frontière entre marqueurs de début/fin (plutôt libre) et le noyau, et très rarement entre les marqueurs de frontière. Ainsi l'insertion de l'adjectif allemand dans un syntagme nominal se fait entre le déterminant et le substantif. La nature de l'élément inséré est généralement inférieure ou égale à celle de l'élément "récepteur". Une proposition ne se trouvera donc pas insérée dans un syntagme. La recherche de ces éléments insérés est assez facile. Une fois la structure des syntagmes simples identifiés, il suffit de rechercher des syntagmes qui peuvent venir s'intercaler dans les marqueurs de frontière et le noyau de la première structure. Le tableau 3.3 de la section 3.3.2 donne quelques exemples de structures syntagmatiques formées par composition interne, ainsi que l'algorithme utilisé.

Passons maintenant à la composition externe. La recherche va donc se faire sur des syntagmes contigus. Le principal problème va être de pouvoir différencier les syntagmes contigus qui sont en relation avec des syntagmes contigus qui ne sont pas en relation. Un élément va nous faciliter la tâche. Il existe un type de syntagme particulier qui est toujours en relation avec un autre SR. Nous appellerons ce syntagme un Syntagme Subordonné (SSub). Ces SSub ne peuvent donc se rencontrer que dans les structures syntagmatiques (ils ne peuvent exister seuls). Ce type de syntagme se différencie formellement du SR par sa morphologie et par son critère positionnel. La relation entre un syntagme régissant et un syntagme subordonné peut être marquée formellement par deux critères : le critère morphologique et le critère positionnel.

Comment par la morphologique. La marque morphologique du subordonné peut être de deux types : soit elle dépend des caractéristiques (genre, nombre, cas par exemple) de son régissant, soit elle est indépendante des caractéristiques du régissant). Le premier cas correspond aux structures d'accord. Nous renvoyons là aussi le lecteur à la section 3.3.2 qui donne quelques exemples de telles structures (tableau 3.4). Nous trouvons ce cas, par exemple en français, dans la relation entre un substantif et un adjectif (*les -s -s* par exemple). L'adjectif prend généralement le genre et le nombre du substantif. Il peut aussi dépendre du cas du substantif (allemand). Dans le deuxième cas, la marque portée par l'élément

dans un texte. Il doit exister une relation de dépendance entre les deux éléments considérés.

[62]Nous pouvons considérer les SAD et SAF de la même manière dans cette section.

subordonné ne dépend pas du régissant. Cette marque peut dépendre du subordonné ou non. Ce cas est illustré par la structure génitive turque où le substantif porte le suffixe (*-i*) quels que soient le genre et le nombre du régissant. Le cas est similaire pour la structure génitive allemand mais la marque est dépendante des caractéristiques du subordonné, alors qu'elle est invariable en turc.

Le deuxième critère formel qui peut indiquer une relation *régissant-subordonné* peut être de nature positionnelle. Supposons que, dans la structure *régissant-subordonné*, le régissant soit toujours le premier élément et donc que le subordonné n'apparaisse qu'en deuxième position. Cet élément subordonné peut donc apparaître à la fin d'une entre-ponctuation. Mais comme cet élément nécessite un régissant, il ne pourra pas apparaître en début d'entre-ponctuations. Le subordonné possède les mêmes caractéristiques qu'un marqueur de frontière morphologique. Le tableau 4.15 nous montre quelques exemples d'adjectifs antéposés (turc) ou postposés (vietnamien, français). Le tableau a été construit en travaillant au niveau syntagmatique. Si le mot appartient à un syntagme qui commence une entre-ponctuation, il est comptabilisé comme début. Ainsi dans *le français moyen ...*, *français* est considéré comme débutant l'entre-ponctuation (+1 dans la colonne *début*). L'on voit que ces éléments se comportent exactement comme des marqueurs de frontière, à la différence qu'ils sont de nature lexicale. Ce type de tableau est très similaire au tableau 4.3 des marqueurs caractéristiques de frontière. La catégorie des mots ainsi définie peut être carac-

Langue	Mot	Effectif	Début	Fin
Turc	iki	198	18	1
	tüm	171	56	0
	yüksek	74	22	0
Vietnamien	dothaui	125	0	35
	gì	279	0	80
	khác	133	0	48
Français	français	211	21	65
	économique	127	1	40
	nationale	122	3	47
	N-ique	1895	39	576

TAB. 4.15 – Exemple de Syntagmes Subordonnés : les adjectifs en turc, vietnamien et français. Ces éléments sont caractérisés par leur position fixe par rapport à leur SR.

térisée par une morphologie spécifique (les terminaisons *-ique*, *-ale* en français) ou non (comme en turc ou en vietnamien). Ces éléments sont considérés comme lexicaux car la catégorie qu'ils définissent possède un nombre important d'éléments. Nous retrouvons là la distinction entre classe fermée (morphologique) et classe ouverte (lexicale). Mise à part cette distinction lexicale/morphologique, ces éléments sont considérés comme étant des marqueurs de frontière de structure syntagmatique. Il peut bien sûr y avoir combinaison entre ces deux critères (morphologique et positionnel) qui caractérisent les SSub. Ainsi, en allemand, le groupe génitif masculin caractérisé par la structure *des N-es* est marqué morpho-

logiquement, et possède une position fixe postposée par rapport à son régissant.

Le critère positionnel permet de catégoriser les SSub en deux catégories (similairement aux SA (SAD et SAF)) : les Syntagmes Subordonnés de Début (SSubD) pour les SSub antéposés, et les Syntagmes Subordonnés de Fin (SSubF) pour les SSub postposés. Les Syntagmes Subordonnés possèdent une morphologie assez différente des Syntagmes Relatifs et aussi souvent moins riche : les séquences de marqueurs de frontière sont moins développées. Elles peuvent être nulles pour certaines structures de langues (vietnamien, turc). Comme dans toutes les structures trouvées, il peut exister des marqueurs de frontière caractéristiques des structures subordonnés.

De même que nous nous sommes interrogé sur l'utilité d'introduire le niveau propositionnel, nous pouvons faire de même en ce qui concerne le Syntagme Subordonné . Son utilité est apparue en travaillant sur les langues turque et surtout vietnamienne. Dans cette langue, les adjectifs et adverbes ne possèdent pas de morphologie particulière (ni début ni fin particulière). Ce sont même des mots invariables. Pourtant, ces mots avaient cette caractéristique positionnelle qui les rendaient similaires à des marqueurs de fin. Nous avons donc introduit un Syntagme Subordonné en vietnamien, très utile pour comprendre la structure de cette langue. Ayant trouvé des traces de cette structure dans les autres langues, elle a ensuite été généralisée.

Nous avons écrit plus haut que les SSub dépendent d'un Syntagme Relatif . Ceci est partiellement vrai. Ils peuvent de la même manière être dépendant d'un Syntagme Absolu[63]. Dans certaines langues, les SSub dépendant de SR sont distincts (souvent morphologiquement) des SSub dépendant de SA (adjectifs et adverbes en français). Dans d'autres langues (vietnamien, turc), les SSub sont identiques (du moins les différentes catégories de SSub partagent un assez grand nombre d'éléments communs). Ainsi le mot vietnamien *xa* lorsqu'il dépend d'un substantif est un adjectif (*lointain*), lorsqu'il dépend d'un verbe est un adverbe (*loin*). Il en est de même pour la plupart des autres adjectifs/adverbes de cette langue. Il y a donc une ressemblance entre la notion d'adjectif et d'adverbe : ils sont tous deux de catégorie SSub, mais ils différent par la catégorie de leur régissant.

À ce stade du travail (la gnénération des structures SSub n'a pas été implémentée), nous ne savons pas exactement quelles sont toutes les structures que recouvre cette notion. Faut-il y inclure toutes les structures dépendant d'un syntagme (SR ou SA), où seulement celles qui se distinguent formellement des SA/SR. Doit-on par exemple considérer seulement les adjectifs/adverbes français comme SSub, ou bien y inclure aussi les groupes prépositionnels ? Nous penchons plutôt pour la première solution.

4.8.3 Les structures de propositions

Nous allons maintenant recenser les constructions composées de deux propositions. Pour illustrer ce propos, prenons l'entre-ponctuation suivante :

, *par exemple l'écrivain* ***souhaitait*** *que sa pièce* ***soit enregistrée***

[63] Typiquement la catégorie des adverbes.

par une seule caméra.

Dans cet exemple aucun syntagme de la proposition *par exemple l'écrivain* ***souhaitait*** n'est en relation avec un syntagme de la proposition *que sa pièce* ***soit enregistrée*** *par une seule caméra*, mais il existe une relation entre les deux propositions. De façon similaire aux structures formées de syntagmes, il existe une proposition régissante et une proposition subordonnée. La construction française la plus caractéristique et la plus fréquente est celle de la subordonnée conditionnelle *si P1, P2* :

– *Si on n'exploite pas les idées sur le moment, on doit y renoncer.*

De même que les constructions de syntagmes peuvent s'enchaîner, plusieurs propositions peuvent être en relation.

– *Hier kam man noch besser als unten zu der Überzeugung, daß die Türken verloren wären, wenn es ihnen nicht gelänge, mit ihren Belagerern einig zu werden.*

Comment de telles structures sont identifiées ? Une première méthode simple consiste à rechercher les entre-ponctuations où existent deux syntagmes absolus. L'on voit apparaître alors des régularités morphologiques dans ces couples, qui caractérisent les débuts de proposition subordonnées (figure 4.17). Cette

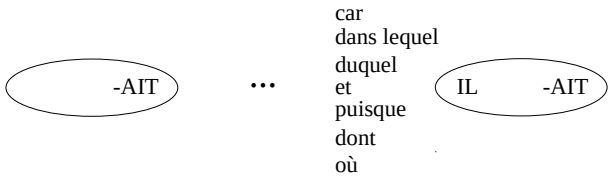


FIG. 4.17 – Exemple de recherche de structure composée de deux propositions en français. On recherche les éléments précédant le deuxième syntagme absolu. Dans l'exemple le deuxième syntagme absolu est formé par la simple structure *il N-ait*.

méthode ne donnerait pas de résultat sur une langue comme l'allemand, où les propositions sont très souvent précédées d'une marque de ponctuation. On a alors une proposition par entre-ponctuations. La détection des propositions subordonnées est alors équilavente à celle des SSub : certaines structures ne peuvent commencer (ou finir) une phrase. Nous parlons ici de phrases et non plus d'entre-ponctuations, ce dernier niveau n'étant plus adéquate pour l'observation des constructions de propositions. L'ordre entre le régissant et le subordonné est soit libre soit fixe selon la construction et la langue. Nous avons donc deux types de propositions : une proposition régissante (relative (*sic*) si l'on reprend la terminologie du niveau syntagmatique) et la proposition subordonnée. La construction de ces structures permet de catégoriser les marqueurs de frontière de proposition en plusieurs catégories. Certains marqueurs de frontière ne se rencontrent qu'en début/fin de proposition subordonnée (les conjonctions de coordination par exemple). D'autres n'apparaissent qu'en début/fin de proposition régissante (adverbes de phrases).

La distribution du couple de propositions est très similaire à la distribution d'une proposition. Dans les emplacements où une proposition peut apparaître, un couple [*proposition régissante, proposition subordonnée*] peut aussi

apparaître. Ainsi dans la structure *:il est certain que*, le mot *que* correspond à un marqueur de début de proposition. Mais la structure peut aussi bien être complétée par un couple de propositions comme dans la phrase :

> Il est certain ***que*** *si les coups avaient été portés par de simples particuliers, il eût immédiatement été requis une information pour coups et blessures ayant entrainé la mort sans intention de la donner.*

Plus une structure est grande, plus la combinaison de ces structures entre-elles semble difficile. Ainsi, s'il est possible d'avoir une composition interne entre syntagmes, ce type de composition pour la proposition n'a pas été rencontré dans nos différents corpus. La seule composition possible entre deux propositions est la composition externe (n'oublions pas que la relative concerne une relation entre un syntagme et une proposition (section 4.9.2)).

4.9 La prédiction des structures

La théorisation formelle de la structure des langues nous permet de mettre à jour toutes les possibilités de structures pouvant être rencontrées. Ce travail a pour but de recenser toutes les combinaisons de structures possibles des langues.Pour générer tous les types de relations possibles entre structures, il suffit de prendre chaque structure identifiée (morphème, syntagme, proposition) et de les combiner avec toutes les autres structures. Cette méthode est très similaire à celle utilisée par les physiciens dans la recherche des particules élémentaires. La théorie avait établi l'existence de 15 mésons (combinaison d'un quark et d'un antiquark). Seuls 14 avaient été observés. Des laboratoires se sont donc mis à la recherche du quinzième (combinaison d'un quark *charme* à un antiquark *beauté*) qui vient d'être découvert ou plutôt observé ([SciencesAvenir, 1998]). Comme on le voit, la découverte d'un objet est d'autant plus facile si l'on connaît (suppose) déjà son existence. Nous essayons donc de recenser toutes les structures (ou les objets plus généralement) que la théorie nous permet de construire, puis l'on confronte ces objets théoriques avec la "réalité" que sont les corpus. *À ce moment là du processus, nous voyons donc bien que c'est la théorie qui guide explicitement la recherche et non les données.* Dans le cas de la structure grammaticale des langues, les possibilités théoriques ne sont pas grandes puisque le nombre d'éléments servant à construire ces possibilités est peu nombreux (moins d'une dizaine d'éléments). La génération systématique de ces structures permet de rechercher toutes les structures théoriques. Cela permet, entre autre, de rechercher les structures très rares de la langue, et qui sont donc difficilement décelables si on ne les cherche pas spécifiquement. Cette théorisation des structures est très importante car elle permet de guider le processus de génération des structures. Elle limite le champ d'investigation : tous les faits observables (les régularités) ne sont pas pris en compte.

De façon similaire, pour établir les différentes manières dont deux éléments peuvent se combiner, nous nous sommes servi, dans la section 4.8.1, de notre conception de la langue comme objet linéaire. Il existe donc un aller retour entre les données et les structures théoriques, l'un servant à construire l'autre et réciproquement. On trouvera un autre exemple d'une génération des possibilités

théoriques dans [Mel'čuk, 1987, page 119] ou la liste des combinaisons possibles des dépendances syntagmatiques entre deux éléments est ainsi produite.

4.9.1 La génération des couples de syntagmes

Pour illustrer ces propos, nous allons prendre comme exemple la génération des couples de syntagmes. Nous avons vu qu'il existait en tout et pour tout trois types de syntagmes : le Syntagme Absolu, le Syntagme Relatif , et le Syntagme Subordonné . Dans les propos suivants, nous avons fusionné SAD et SAF en SA, et SSubD, SSubF en SSub, la position ne semblant pas jouer de critère discriminant. Nous allons donc générer tous les couples possibles composés de ces quatres éléments en effectuant leur produit cartésien (figure 4.18), *sans tenir compte de l'ordre linéaire des deux éléments.* Un couple est composé de deux syntagmes, dont l'un est le régissant de la structure, et le deuxième l'élément subordonné. Il ne nous reste plus qu'à rechercher dans une langue donnée l'exis-

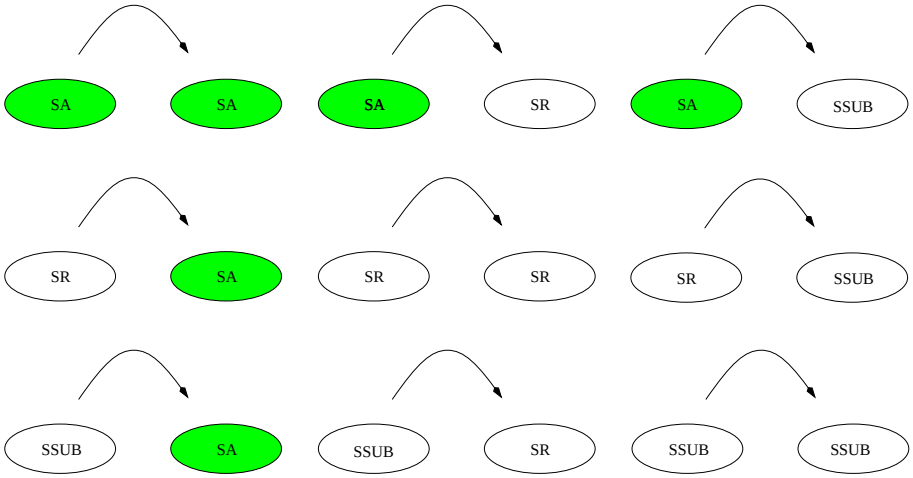


FIG. 4.18 – Liste de tous les couples de syntagmes simples possibles en théorie. Le sens de la flèche correspond au sens *Régissant-subordonné*, *sans renseignement sur l'ordre linéaire entre le régissant et son subordonné.*

tence de ces couples. Prenons le cas du français. Le tableau 4.16 nous montre les différentes structures trouvées dans le français. La forme $X \rightarrow Y$ indique que les couples éléments X et Y sont en relation, et que l'élément à gauche de la flèche (X) est le régissant de la structure. L'ordre linéaire est quelconque. Il se pose ici le problème de déterminer le régissant dans de telles structures. Les choix sont expliqués dans la section 4.10. Ils correspondent généralement aux conventions. Deux structures n'ont pas été observées dans notre corpus *français01* : SR→SA, et SSub→SA[64]. Les structures où le subordonné est de type SA posent un problème : faut-il considérer ce SA comme un Syntagme ou bien comme une trace de la proposition qui le contient ? Dans ce dernier cas, la

[64] On peut peut-être la voir dans une phrase comme : *Il a beau->courir*, ou *beau* serait le SSub et *courir* le SA.

Structure	Séquence	Exemple
SA→SA	verbes (?)	*[pouvait] [travailler]*
SA→SR	verbe + substantif	*[augmenterait] [les dangers]*
	substantif + verbe	*[le programme] [annonçait]*
SA→SSub	verbe + adverbe	*[il parlait] [évidemment]*
SR→SA	?	
SR→SR	substantif + substantif	*[dans l'usine] [de la vallée]*
SR→SSub	substantif + adjectif	*[le nationalisme] [azéri]*
SSub→SA	?	
SSub→SR	adverbe + substantif	*[conformément] [à la ligne]*
SSub→SSub	adverbe + adjectif	*[évidemment] [prioritaire]*

TAB. 4.16 – Quelques structures syntagmatiques en français. Le *?* marque les structures non rencontrées dans notre corpus. Les crochets délimitent les syntagmes.

relation deviendrait *X→Proposition*. Nous reviendrons sur ce problème dans la section suivante. Les couples concernés sont SA→SA, SR→SA, et SSub→SA. Le cas de la construction SR→SA pourrait faire penser à la structure de la subordonnée relative, mais ce n'est pas le cas : cette structure correspond à un couple *Syntagme→Proposition* (section suivante). Un couple pose problème : le couple SSub→SR. On peut considérer que le couple SSub→SR existe en français dans la construction : *Adverbe→Groupe Nominal* (*peu de X beaucoup de X, énormément de X, conformément à X*). Nous avons donc affaire la plupart du temps à une structure très limitée dans son utilisation, que l'on pourrait schématiser par *une Quantité de quelque chose*. On pourrait considérer ces constructions d'une autre manière en posant que les éléments comme *peu*, *beaucoup* font partie des marqueurs de français d'un SR, mais ce choix n'est pas retenu pour deux raisons : d'une part, la nature de ces éléments peut être lexical (comme *énormément*), et il peut venir s'ajouter une construction SSub→SSub au SSub de la structure (*trop peu de X*). D'autre part, on notera un fait important dans cette construction : lorsque cette structure est en position sujet, c'est l'élément subordonné (le SR) qui s'accorde avec le verbe :

> Un homme qui aime dire tout haut ce que *beaucoup de ses collègues pensent* tout bas.

Les deux constructions (SSub→SA et SSub→SR), si elles existent dans la langue, semblent avoir un effectif très faible, et ne correspondre qu'à des constructions bien particulières. Nous entrons dans des considérations qui ne peuvent être prises en compte qu'après une étude très fine de la langue. Se pose ici non pas le problème de l'identification des structures, mais celui de leur reconnaissance. Si les structures fréquentes d'une langue sont assez faciles à caractériser (considérer la séquence *le président de la république* comme étant une construction *SR→SR*), car l'on possède beaucoup de renseignements sur celles-ci, les structures plus rares sont plus délicates à étudier en se basant sur des critères formels.

4.9.2 La génération des couples transhiérarchiques

L'étude suivante concerne les couples où les deux éléments n'appartiennent pas à un même niveau hiérarchique. Nous avons trois types de structures élémentaires : le morphème, le syntagme, et la proposition. Nous allons donc regarder s'il existe des structures qui comprennent un régissant d'un certain type et un subordonné d'un autre type. Pour cela, nous générons les neuf possibilités théoriques (tableau 4.17).

régissant \ subordonné	Morphème	Syntagme	Proposition
Morphème	√		
Syntagme		√	√
Proposition		√	√

TAB. 4.17 – Les différentes structures composées de différents niveaux de la hiérarchie. La marque √ indique que la structure a été observée.

La première observation concerne le morphème : il ne se combine avec aucune autre structure élémentaire. Il se combine uniquement avec lui même pour former le syntagme. La combinaison *Proposition→Proposition* est expliquée à la section 4.8.3. L'observation la plus intéressante porte sur les combinaisons possibles entre le syntagme et la proposition. Nous n'avons pas trouvé de structure correspondant à la combinaison *Proposition→Syntagme*, le syntagme étant alors incorporé (conventionnellement[65]) dans la proposition. Par contre, la combinaison *Syntagme→Proposition* existe et est très fréquente. La structure typique de ce cas étant en français la proposition subordonnée relative. En toute généralité, le type du syntagme peut être absolu, relatif, ou subordonné. *Nous avons donc une unité dépendante d'une deuxième unité inférieure hiérarchiquement.*

Le principal problème rencontré porte sur les SA : doit-on considérer systématiquement les SA comme des marques de la présence d'une proposition, ou bien peuvent-il être vus comme des syntagmes. Autrement dit, existe-t-il des structures *SA→X*, et *X→SA*, ou bien faut-il y voir des structures *Proposition→X* et *X→Proposition* ? Le problème ne se pose que lorsqu'il n'existe pas de marques de frontière de propositon dans l'entre-ponctuations. Prenons l'exemple français suivant :

> *[Le thème de l'aménagement du territoire va prendre de plus en plus d'importance dans les années à venir] tant on sent les déséquilibres s'accentuer avec une grande rapidité.*

Nous avons délimité une première proposition entre crochets. La séquence restante est plus délicate. Faut-il la considérer comme une proposition ou deux ? Le problème provient du verbe à l'infinitif *s'accentuer* (catégorisé comme SA[66] par notre algorithme (section 6.4.4)). Faut-il voir une relation *SA→SA* entre *sent* et

[65] En français tout du moins. Mais le cas est à étudier.

[66] Considérer l'infinitif commeétant un verbe ne va pas de soi : *On ne répétera jamais suffisamment que l'infinitif n'est pas un verbe.* [Tesnière, 1959, page 419]. Mais il connait la proposition infinitive (chapitres 180 à 190)

s'accentuer ou bien une relation *Proposition→Proposition* ? Nous avons pris le parti de maximiser le nombre de propositions dans les entre-ponctuations, c'est-à-dire considérer les structures *SA→SA* comme étant des structures *Proposition → Proposition*. Ceci pour deux raisons. La première se place dans un point de d'analyse. Il nous semble qu'introduire des propositions peut faciliter l'analyse. En effet, inclure un élément propositionnel permet d'inclure les contraintes liées à ce niveau dans l'analyse. La deuxième raison provient de la comparaison entre séquences de SR et séquences de SA. Dans le premier, les relations entre les différents SR varient selon les séquences (figure 4.19).

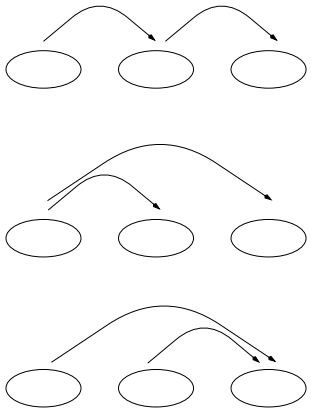


FIG. 4.19 – Les différentes relations possibles dans une séquence de trois SR en français. Nous trouvons toutes les possibilités (La flèche va du régissant au subordonné).

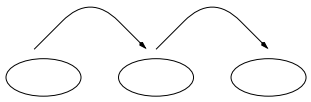


FIG. 4.20 – La seule mise en relation possible dans une séquence de trois SA. Un SA est considéré comme régissant du SA suivant.

Dans les séquences de SA, les relations semblent être fixes. Un SA est toujours considéré comme le régissant du SA suivant (figure 4.20). Il existe donc une différence importante dans le comportement des séquences de SA et des séquences de SR. Cette observation a été faite sur le français, et devrait être validée sur d'autres langues. La relation entre deux SA semble être plus contraint que celle entre deux SR. Nous voyons là un indice en faveur de l'hypothèse de la maximisation des propositions. Si cette hypothèse retenue, il est possible qu'elle soit dépendante de la langue étudiée. Prenons le cas des propositions allemandes telles que :

- ***Er hat dir*** *einen Mund* ***gegeben***.
- *, daß der Adjutant des Miralai von ihnen* ***gefangen genommen worden sei***.

Dans la première proposition, les SA sont discontinus, et sont tous deux caractérisés par une position absolue (début pour *Er hat dir* et fin pour *gegeben*). Dans la deuxième proposition, le morphème initial *daß* et la séquence finale *gefangen genommen worden sei* sont caractéristiques d'un début et d'une fin de proposition. Il semble que l'option de maximisation ne soit pas adéquate à de telles structures propositionnelles. Une étude plus complète reste donc à réaliser. Toutes ces questions concernent essentiellement la mise en relation de structures (syntagme et proposition). Or cette mise en relation est très difficile à réaliser en ne considérant uniquement que des critères formels, beaucoup plus difficile que la mise en relation des éléments qui forment un syntagme. Le travail sur des langues inconnues devient alors très difficile.

4.10 La notion de relation

Nous avons beaucoup parlé de relation. Introduire cette notion a été nécessaire dès que nous nous sommes intéressé aux structures composées. Les notions classiques de *régissant* et de *subordonné* sont apparues dans le processus de découverte de ces structures. Ce processus se déroule comme suit : nous partons d'une structure donnée (syntagme ou proposition) et nous cherchons d'autres structures apparaissant souvent avec elle. Cette méthode permet de mettre à jour les structures composées de la langue (sections 4.8 et 6.4.6). Nous avons utilisé le terme *relation* pour indiquer le lien entre ces éléments. Il y a *relation* entre les deux éléments qui composent une structure. Un ordre (conceptuel et non linéaire) de fait s'est imposé : l'élément qui permettait de trouver cette structure est considéré comme premier dans la structure. Nous avons alors repris la terminologie en vigueur (au moins chez Tesnière) en le désignant comme *régissant* de la structure composée, et avons considéré l'élément ajouté comme *subordonné*. Ainsi, dans la structure composée incluant un SR et un SSub, le SR est considéré comme le régissant et le SSub comme le subordonné. Ceci explique le fait que les actants soient considérés comme subordonnés au SA de la proposition. L'identification des actants se fait en partant de la structure des SA (section 6.4.6). L'élément permettant cette identification est donc le SA et les marqueurs de frontière de proposition. Pour cette raison, le SA est considéré comme l'élément central à la proposition. Les SR jouant le rôle d'actant sont donc considérés comme des éléments subordonnés au SA. On applique le même principe à toutes les structures composées.

Il est difficile de trouver une définition de la relation entre éléments. [Tesnière, 1959] utilise le terme de *connexion* entre mots en faisant le parallèle entre les liaisons chimiques entre atomes, mais ne donne pas de définition précise de la connexion. [Martinet, 1970] ne définie pas la relation mais le deuxième élément de cette relation appelé *expansion* et définie l'élément subordonné comme :

> On appelle **expansion** tout élément ajouté à un énoncé qui ne modifie pas les rapports mutuels et la fonction des éléments préexistants. [Martinet, 1970, page 128]

Le propos suivant de [Mel'čuk, 1987] résume la situation :

> I am unable to propose a rigorous definition of syntactic dependency. However, since this notion is extremely important and, at the same time, not quite clear, some preliminary considerations seem to be in order. [Mel'čuk, 1987, page 129]

Il propose la typologie des relations (il utilise le terme de *dépendance*) suivante :

- dépendance morphologique
- dépendance syntaxique
- dépendance sémantique

Nous n'allons pas détailler ici ces différentes dépendances. Les critères utilisés ne sont pas tous formels (en particulier pour la dernière dépendance). On notera que Mel'čuk définit aussi la dépendance comme une relation entre *deux* éléments. Notre typologie des différents types de relations repose sur la nature des éléments utilisés dans la relation :

- relation morphologique (entre deux morphèmes)
- relation syntagmatique (entre deux syntagmes)
- relation propositionnelle (entre deux propositions)
- relation syntagmo-propositionnelle (entre un syntagme et une proposition)

La question est de savoir si une telle typologie nous est utile, c'est-à-dire s'il existe une différence (formelle ?) entre ces trois types de relations. Les relations capturées par notre méthode de découverte sont majoritairement des relations de subordination (de dépendance). La deuxième relation traditionnelle, celle de coordination, est généralement moins marquée formellement (plus exactement les régularités formelles sont moins fréquentes), et est beaucoup plus difficile à trouver. Cette notion de relation nécessite encore un travail important.

4.11 La représentation de la structure

Pour l'instant nous n'avons décrit les structures composées qu'en termes de couples d'éléments : couple de syntagmes, couples de propositions. Est-ce que ce formalisme suffit à décrire toutes les structures de la langue ? Cela dépend de la relation que les deux éléments entretiennent. La réponse est affirmative si l'on considère la relation de dépendance que nous avons étudiée : *Régissant→Subordonné*. D'une manière générale, toute relation concernant *n* éléments peut être décomposée en *n-1* relations entre couples d'éléments. Dans un processus de découverte, il est très difficile de mettre à jour des régularités qui concernent plus de deux éléments. La découverte des relations concernant des séquences de plus de deux éléments se fait en passant par la connaissance des relations existant entre deux éléments. Le cas qui illustre parfaitement ces propos est celui de la relation entre un SA (verbe) et ces actants, prenons ces deux premiers actants (sujet et verbe). S'il est facile de trouver la relation entre le prime actant et le verbe, et entre le second actant et le verbe, la structure composée des trois éléments est très difficile à trouver : les couples *contigus* de structures sont beaucoup plus fréquents que les triplets. Plus la structure est grande (en terme de taille et non pas de hiérarchie), plus elle accepte d'éléments subordonnés qui viennent parasiter la structure étudiée. Les structures où interviennent plusieurs éléments (comme la structure actancielle) peuvent

être reconstituées en regroupant les différents couples qui partagent un même élément régissant.

Cette structure de couples est suffisamment puissante pour prendre en compte les séquences de syntagmes composées d'un nombre quelconque d'éléments, en particulier grâce aux couples qui possèdent deux éléments de même nature (SR-SR). Nous retrouvons une structure récursive, où la récursion est définie comme :

> Recursion is a particular kind of representation of a particular kind of repetition. [Franova and Kooli, 1998]

Nous avons bien une représentation particulière (les deux éléments sont de même nature) d'une répétition d'éléments de même nature. Notons que le schéma X-barre présenté dans [Chomsky, 1970] utilise aussi cette représentation en couples (section 7.2).

4.12 Un récapitulatif

Nous allons donner dans cette section un récapitulatif de toutes les structures que nous avons sélectionnées. Nous avons d'abord trois éléments de base :

- Le morphème
- Le syntagme
- La proposition

Le morphème est l'unité de base et n'est pas structuré. Il existe deux types de morphèmes : le morphème grammatical, qui est utilisé pour marquer les frontières de structures élémentaires, et le morphème lexical qui compose le noyau du syntagme. Pour les autres niveaux, chacun peut être caractérisé par des marqueurs de frontière et par des contraintes positionnelles. De plus, chaque construction de deux structures élémentaires peut aussi avoir des marqueurs de frontière caractéristiques. Le nombre de catégories est assez important mais l'on s'aperçoit que les "ressources" en marqueurs de frontière sont limitées et qu'une langue utilise des mêmes éléments pour marquer différentes structures. Ainsi les prépositions allemandes peuvent être utilisées comme marqueurs de début de SR (utilisation canonique), de SSub, de Proposition Subordonnée, et pour certaines de marqueurs de fin de Proposition Régissante. *De plus une langue donnée n'utilise pas toutes les catégories de marqueurs de frontière mises à sa disposition.* Une telle langue, si elle existait, serait très adaptée à une analyse syntaxique automatique, puisque toutes les structures seraient explicitement marquées.

Dans ces structures (syntagme, proposition), il existe des marqueurs caractéristiques qui aident à la découverte de ces structures. Dans ce recensement des structures, nous avons sans doute (certainement) oublié quelques cas, mais l'important est de mettre au point une théorie qui permet de les découvrir théoriquement. Des questions restent en suspens. Par exemple, faut-il introduire la catégorie des Syntagmes Subordonnés aux Syntagmes Subordonnés (des SSub-SSub) ? Nous n'en avons pas vu l'utilité pour les langues étudiées, les SSub semblant être leur propre subordonné. Mais il se peut que des langues utilisent un type de syntagme particulier pour cette structure. Dans ce cas, un nouveau

type de syntagme devra être ajouté. Le tableau 4.18 donne un récapitulatif des structures.

proposition	régissante		
	subordonnée	à une proposition à un syntagme	
syntagme	absolu	de début de fin	de proposition
	relatif		
	subordonné	de début de fin	de syntagme
morphème	lexical		
	grammatical	de début de fin	de syntagme ou de proposition

TAB. 4.18 – Les différentes structures.

Voyons maintenant quelle différence existe entre les différents syntagmes : syntagmes absolu, relatif et subordonné. Le Syntagme Absolu (SA) correspond à une structure syntagmatique caractérisée par sa position absolue dans une proposition, qui se traduit dans un texte écrit par un nombre d'occurrences très élévé apparaissant avant (SA de Fin) ou après (SA de Début) une ponctuation. Certaines structures de SA (par exemple la structure française [*ne ...pas*]) ne sont pas identifiables grâce à cette position absolue, mais grâce au processus de catégorisation (section 6.4.4). Dans toutes les langues étudiées, le SA correspond toujours à la structure verbale de la langue.

Le Syntagme Relatif a correspondu, dans un premier temps, aux syntagmes qui n'étaient pas des SA. Nous l'avons nommé relatif par opposition au terme *absolu*. Est alors apparu un troisième type de syntagme : le Syntagme Subordonné. Ce type de syntagme a été introduit pour prendre en considération le fait que certains SR n'étaient pas si "relatifs" que cela : ils possédaient une caractéristique positionnelle (ils n'apparaissaient pas soit avant une ponctuation soit après une ponctuation). Mais cette caractéristique était moins forte que dans le cas du SA. Nous avons appelé ce type de syntagme le Syntagme Subordonné, car la contrainte positionnelle est due au fait que ce syntagme nécessite un syntagme régissant (4.8.2). Les SSub peuvent aussi être identifiés grâce aux structures d'accord (critère morphologique) de la langue (section6.4.6).

Nous voyons donc qu'il existe trois types de syntagmes, deux étant caractérisés positionnellement, et un, le SR, correspondant aux syntagmes n'étant ni absolus ni subordonnés. La contrainte positionnelle s'appliquant aux SA et aux SSub, ces deux types peuvent se partitionner en deux : SA de Début (SAD) et SA de Fin (SAF), et SSub de Début (SSubD) et de Fin (SSubF).

La catégorisation des propositions est assez simple puisque le critère utilisé est la nature du régissant : aucun (proposition régissante), subordonné à un

syntagme ou bien à une proposition. il est clair que d'autres catégorisation peuvent être effectuées, en particulier lorsque les structures supérieures à la propositions auront été (découvertes) intégrées.

4.13 Une comparaison entre nos catégories et les autres catégories

Les classes de mots, unité traditionnelle de la langue, sont catégorisées en *partie du discours* (lat. *partes orationis*, gr. *meroi logou*). La notion est ancienne puisqu'on la trouve déjà dans les Poétiques d'Aristote. Depuis Denys de trace, elles sont au nombre de huit. Robert Estienne, en 1557, considérait neuf parties du discours en ajoutant l'article, qui n'existe pas en latin, catégorisation que la grammaire de Port-Royal [Arnauld and Lancelot, 1660] a reprise. Ces parties sont :

- nom
- verbe
- pronom
- article
- adjectif
- adverbe
- préposition
- conjonction
- interjection

Nous pouvons assez facilement recatégoriser ces classes dans notre catégorisation :

nom	noyau de SR
verbe	noyau de SA
pronom	marqueur de frontière de syntagme ou de proposition
article	marqueur de frontière de syntagme
adjectif	noyau de SSub (de SR)
adverbe	noyau de SSub (de SA) ou marqueur de frontière de proposition
préposition	marqueur de frontière
conjonction	marqueur de frontière
interjection	?

Quand nous mettons en parallèle la catégorie *nom* et *noyau de SR*, nous voulons dire que le nom correspond à un élément comprenant un noyau de SR avec ses marqueurs de frontière liés. Il faut rappeler qu'une catégorisation utilise des *mots* et que notre catégorisation utilise des morphèmes et des syntagmes. La catégorie de l'interjection n'est pas apparue dans notre travail. Se pose aussi le problème de catégories comme le pronom et l'adverbe : ces deux classes regroupent des éléments aux distributions très disparates. Si les divers pronoms d'une langue sont généralement des marqueurs de frontière (comme tous les éléments grammaticaux), ils peuvent marquer la frontière de différentes structures

(syntagme ou proposition). Quant à la classe des adverbes, elle semble regrouper tout ce que l'on ne peut pas classer ailleurs. Dans notre catégorisation sur le français, certains adverbes sont considérés comme SSub (généralement au verbe). D'autres sont vus comme des marqueurs de frontière de proposition[67] (*donc*, *puis*).

On trouve aussi chez [Tesnière, 1959, page 63] et [Hejmslev, 1966] une catégorisation intéressante concernant les mots lexicaux. Il existe pour Tesnière deux catégories *concrètes* : le substantif (notre SR) et le verbe (notre SA), et deux catégories abstraites : L'adjectif (SSub de SR) et l'adverbe (SSub de SA).

> L'adverbe est au verbe ce que l'adjectif est au substantif. [Tesnière, 1959, page 63]

Hejmslev adopte une vue différente : l'on trouve d'abord le verbe (SA), puis le substantif (SR) qui *modifie* le verbe, puis l'adjectif (SSub de SR) qui modifie le substantif, et enfin l'adverbe qui modifie l'adjectif (SSub de SSUb). Nous retrouvons bien chez ces deux auteurs notre notion de syntagme subordonné.

Il faut bien être conscient qu'il existe plusieurs catégorisations possibles des éléments linguistiques. Ces catégorisations dépendent des critères utilisés (comme la classification retenue par [Halliday, 1985, page 214]).

nonimals	noun	common proper pronoun
	adjective	
	numeral	
	determiner	
verbals	verb	lexical auxillary finite
adverbials	preposition	
	adverb	
	conjunction	linker binder continuative

TAB. 4.19 – La classification fonctionnelle des parties du discours de [Halliday, 1985, page 214]

[67] Ils sont classés comme marqueur de frontière de proposition, mais il est vraisemblable qu'ils appartiennent à une structure supérieure à la proposition.

Chapitre 5

La structure lexicale

Sommaire

5.1 Les régularités lexicales

Jusqu'à présent, l'étude des structures s'est faite en utilisant des éléments grammaticaux (mots et morphèmes marqueurs de frontière). Il existe un deuxième type d'éléments : *l'élément lexical*. Un élément lexical est composé d'une séquence de morphèmes comprenant un noyau syntagmatique (mot lexical (plein) ou syntagme). Pourquoi les éléments grammaticaux ont-ils été privilégiés jusqu'à présent ? Simplement parce que leur effectif permet d'avoir énormément d'informations sur eux. De plus, ces éléments sont assez invariants d'un corpus à un autre (de la même langue), ce qui n'est pas le cas des éléments lexicaux. Ils ont donc été longtemps ignorés. L'intérêt de leur utilisation est apparu lors de l'opération de segmentation. Mais les résultats étant suffisamment bons sans leur prise en compte explicite, ils n'ont pas été intégrés au traitement et ont sombré dans l'indifférence. Ils ont fait leur réapparition lorsqu'il a fallu trouver les relations entre syntagmes. Les marques morphologiques et positionnelles n'étant pas assez présentes dans certaines séquences, il a fallu rechercher d'autres informations. Cette recherche a commencé lorsque nous avons travaillé sur le turc. Nous prenions des entre-ponctuations au hasard et essayions de trouver leurs structures, c'est-à-dire mettre en relation tous les éléments de l'entre-ponctuation. Prenons l'entre-ponctuation suivante :

> *mesih'in acı çekip ölümden dirilmesi gerektiine dair açıklamalarda bulunuyor*

La segmentation produit la séquence suivante :

> *mesih'-in acı çek-ip ölümden diril-mesi gerekti-ine dair açıklamalarda bulunuyor*

Aucun mot n'est caractérisé comme début, le seul couple morphologique est *-ine dair* (*dair* est un marqueur de fin du syntagme *gerektiine dair*). La seule autre information disponible est celle des effectifs des éléments. Nous considérons alors les mots deux à deux et essayons de déterminer s'ils sont en relation. Prenons *acı* et *çekip*. Pour cela nous regardons l'effectif du couple. Il est de un. Regardons maintenant les séquences qui correspondent au patron suivant : *acı- çek-*, c'est-à-dire une séquence de deux mots contigus commençant par *acı-* et *çek-* Nous trouvons 25 occurrences (tableau 5.1).

Couple	Effectif
acı çekecek	1
acı çekecektir	1
acı çekecei	1
acı çekece¡ini	2
acı çeken	1
acı çekenleri	1
acı çeker	1
acı çekerse	1
acı çekip	1
acı çekiyor	1
acı çekmeden	1
acı çekmek	2
acı çekmesi	4
acı çektiniz	1
acı çektirdiler	1
acı çektiğine	1
acı çektiğiniz	1
acıları çekmemin	1
acıları çekmesi	1
acısını çektikten	1

TAB. 5.1 – Les régularités ne sont pas seulement morphologiques. Nous avons ici un couple lexical *acı- çek-*.

Nous voyons que la régularité des couples formés n'est pas grammaticale mais lexicale : ils sont formés par les noyaux syntagmatiques. Nous avons vu (section 1.7) qu'il fallait manipuler l'effectif avec prudence. Mais comme nous le verrons dans la section 5.3, cette prudence n'est plus de mise lorsque les éléments concernés sont de nature lexicale. La portée de ces informations est bien sûr bien moindre que les informations morphologiques (ces couples sont beaucoup moins fréquents), mais elles n'en demeurent pas moins essentielles pour améliorer la

découverte des relations. Ce fait nous a conduit à nous intéresser davantage aux ressources lexicales que contiennent les corpus. Nous nous sommes alors livré à quelques expériences, en particulier sur les couples de mots lexicaux (section 5.3).

Le lexique est décrit chez certains auteurs, [Pike, 1967], [Hockett, 1961], [Longacre, 1964], comme une des trois composantes de la linguistique, les deux autres étant la phonologie et la grammaire (étude des structures).

> It is here assumed that language is structured in three semiautonomous but interlocking modes, phonology, grammar, and lexicon (Pike's trimodalism). [Longacre, 1964, page 7]

Mais l'intégration des ces trois parties est délicate :

> To describe a language exhaustively (a task as yet seriously attempted by no one), three volumes are nedeed : a phonological statement, a grammatical statement, and a highly spophisticated dictionary. Attempts to incorporate the lexicon directly into the grammar will lead only to the oversimplification of the former or to the endless atomization of the latter. [Longacre, 1964, page 8]

Si nous sommes d'accord sur le fait d'intégrer l'information lexicale dans notre travail, le problème est de savoir comment organiser le lexique (les informations lexicales). Les sections suivantes donnent quelques pistes quant à l'intérêt de l'apport de l'information lexicale et de son intégration dans le processus de découverte.

5.2 L'aide à la segmentation

Nous n'avons pas tiré partie explicitement de l'information lexicale dans notre opération de segmentation des mots. La prise en compte des noyaux peut améliorer les résultats de la segmentation.

Voyons un simple exemple. La troisième étape de la segmention consiste à segmenter tous les mots du corpus, et peut générer un certain nombre d'erreurs (tableau 2.19 de la section 2.3). La prise en compte des éléments lexicaux peut alors réduire le nombre d'erreurs générées par notre segmentation. Par exemple, l'identification de la séquence *indiqu* comme noyau aurait évitée la segmentation du mot *indique* en *ind-ique*. La découverte des noyaux doit sans doute se réaliser en même temps que la découverte des affixes. L'amorçage (l'éternel problème dans le cadre de ce travail) d'un tel traitement peut peut-être se faire grâce aux hapax qui permettent une identification certaine (à plus de 99%) d'éléments comprenant un noyau lexical (dans un corpus de plus d'une dizaine de milliers de mots).

Si les retombées de ce traitement sont assez faibles en français, elles peuvent être d'une grande aide pour des langues possédant un système casuel (comme le latin ou le turc). Seule la prise en compte de données lexicales peut (parfois) nous permettre de déterminer si tel ou tel mot possède un morphème zéro ou non, information importante dans les langues casuelles. Prenons le cas du turc. Le morphème *-u* marque un cas (l'accusatif). Mais ce morphème segmente tous

les mots finissant par *-u*, comme *kuyu*, *huyu*, *tozu*. Or ces séquences sont toutes des noyaux lexicaux : le *-u* final ne correspond pas à la marque casuelle (qui est réalisée par *kuyunu* pour *kuyu* par exemple). Une telle segmentation peut parasiter la découverte des structures actancielles de la proposition, et surtout rend plus difficile la découverte du fameux *morphème zéro*[68] du syntagme ayant le rôle du prime actant.

5.3 L'aide à la mise en relation

Nous allons voir comment l'information lexicale peut être utilisée dans le cadre de la découverte des relations entre éléments. Pour cela nous allons étudier les couples composés de deux éléments lexicaux.

5.3.1 Les couples de lexicaux

Deux moyens ont été utilisés pour générer la liste des couples lexicaux. Premièrement, en utilisant un corpus étiqueté. Les éléments dont l'étiquette indique une nature non lexicale (les étiquettes des mots grammaticaux) sont éliminés du corpus (comme tout bon traiteur de langue qui s'intéresse aux éléments lexicaux). Deuxièmement, à partir d'un corpus non étiqueté. Les mots de moins de cinq lettres sont considérés comme élément non lexical (les éléments grammaticaux de plus de cinq lettres sont donc conservés). Nous avons comparé les deux différents résultats du français. Les différences sont très minimes. Nous avons alors travaillé en utilisant la deuxième méthode (elle évitait une recherche de corpus étiquetés, et cela nous permettait de nous *re*mettre dans les conditions des autres traitements : en travaillant sur un corpus non étiqueté). Le résultat du traitement donne des séquences telles que :

> *engagés responsabilité politique syndicat d'études programmation l'agglomération lyonnaise*

pour une entre-ponctuation initiale :

> *engagés sous la responsabilité politique du syndicat d'études et de programmation de l'agglomération lyonnaise (sepal)*

Une fois le corpus lexical construit, nous calculons les effectifs des couples de lexicaux contigus (tableau 5.2). Les mots n'ont pas été lemmatisés. Dans une langue comme le turc, cette lemmatisation (en fait une identification du noyau syntagmatique suffit) serait très utile, la variété morphologique étant très grande.L'effectif maximal des couples lexicaux français est de 70. Par comparaison, l'effectif du couple le plus fréquent du corpus est de 2423. Le premier couple lexical, *premier ministre*, occupe le rang 124 dans cette liste. L'effectif de ces couples décroît très vite. Les couples de lexicaux qui ont une seule occurrence représente 95% des couples.

Que faire de ces couples ? L'idée générale est que si deux éléments lexicaux sont souvent contigus, alors il existe une relation entre ces deux éléments. Nous verrons qu'un effectif de deux est suffisant pour induire l'existence d'une relation

[68] absence de morphème

Couple français	Effectif
premier ministre	70
milliards francs	67
secrétaire général	65
millions francs	51
affaires étrangères	46
françois mitterrand	33
conseil d'administration	32
chiffre d'affaires	32
banques centrales	32
milliards dollars	28

Couple allemand	Effectif
master lindsay	21
kennen lernen	17
gefangen nehmen	17
gefangen genommen	16
mutessarif mossul	11
makredsch mossul	11
fünfhundert piaster	10
beiden männer	10
mutter gottes	8
lautete antwort	7

TAB. 5.2 – La liste des dix plus fréquents couples lexicaux du corpus *français01* et *allemand01*. Certains mots grammaticaux allemands étant assez longs, peuvent apparaître dans les couples (*zurück*, *beiden*).

entre les deux éléments. Si cette méthode permet de dire que deux éléments sont en relation, elle ne permet pas de préciser la nature de cette relation. En pratique, il s'agit le plus souvent d'un relation de dépendance (quantitativement la plus fréquente), mais il peut aussi s'agir d'une relation de coordination.

5.3.2 Effectif contre information mutuelle

Des travaux ont porté sur le calcul de la liaison qui peut exister entre deux éléments lexicaux. Cette opération est nécessaire dans le domaine de l'extraction terminologique. Pour cela, il existe plusieurs méthodes afin de déterminer cette "force" entre éléments. Nous en avons déjà vu une : l'effectif, mais il existe d'autres mesures plus sophistiquées. Selon [Church and Hanks, 1990] l'information mutuelle est le meilleur critère pour mesurer la force entre deux éléments. Nous renvoyons à [Daille, 1994, pages 115-144] pour une étude détaillée de ces différentes mesures. La définition de l'information mutuelle de deux éléments a et b est :

$$im(a,b) = \ln \frac{P(a,b)}{P(a) \times P(b)} \tag{5.1}$$

où $P(x)$ est la probabilité d'apparition de l'élément x dans le corpus (en pratique obtenu par le quotient de son nombre d'occurrences par le nombre d'occurrences totales du corpus). Les travaux comparatifs effectués par [Daille, 1994] infirme le propos de Church, et désigne la fréquence comme meilleur critère. Il serait intéressant d'effectuer un travail similaire à celui décrit dans [Smadja, 1993], qui utilise l'information mutuelle pour mener à bien différents traitements (extraction de termes, recherches des variations), mais en utilisant cette fois le critère de la fréquence. Notre expérience dans le domaine nous pousse à croire que les résultats seraient aussi bons.

Les tableaux 5.3 et 5.4 présentent les couples français ayant le plus fort effectif et la plus forte information mutuelle. Dans le premier tableau, le classement donné par l'information mutuelle est inverse de celui de l'effectif. Le tableau 5.4

montre que les couples ayant la plus forte information mutuelle ont un effectif très réduit. Ils correspondent aux couples formés de mots n'apparaissant que dans le couple. On retrouve essentiellement les noms propres du corpus dans les premières places.

Couple	Effectif	Information Mutuelle (IM)
premier ministre	70	8.03
milliards francs	67	9.86
secrétaire général	65	9.76
millions francs	52	9.27
affaires étrangères	46	11.47
françois mitterrand	33	10.95
conseil administration	32	10.25
chiffre affaires	32	12.11
banques centrales	32	12.28
milliards dollars	28	14.04

TAB. 5.3 – Les dix couples lexicaux les plus fréquents du corpus *français01*.

Couple	Information Mutuelle	Effectif
pedro toledo	15.59	6
moshé many	15.37	7
barbara stanwyck	15.17	7
baby blood	15.17	7
wall street	15.00	9
karl otto	15.00	8
camil petrescu	15.00	7
serge leclaire	14.59	6
ordures ménagères	14.59	6
malik oussekine	14.52	7

TAB. 5.4 – Les dix couples lexicaux du corpus *français01* ayant la plus forte information mutuelle.

Les différentes mesures applicables offrent des classements différents. Il semble difficile d'ordonner les couples à travers ces mesures. La force d'un lien entre éléments ne peut se calculer à travers une simple mesure numérique. Pourquoi le lien entre les éléments de (*banques centrales*) (IM : 12.28) serait-il plus fort que le lien entre les éléments de (*premier ministre*) (IM : 8.03) ? On peut seulement conclure qu'il existe un lien entre les deux éléments de ces couples. Dans la suite de cette étude, nous avons retenu l'effectif pour caractériser un couple de lexicaux, cette mesure étant la plus simple, et surtout la plus efficace. Le travail à commencer avec les couples ayant un grand effectif : ils correspondent toujours à des éléments en relation. Puis nous avons essayé de descendre le seuil qui permettait de mettre en relation les lexicaux.

5.3.3 La mise en relation grâce aux éléments lexicaux

L'étude des couples de lexicaux a révélé un fait assez surprenant : les couples de lexicaux ayant un effectif de deux sont massivement composés d'élément en relation. Il faut se souvenir qu'environ 95% des couples de lexicaux (dans la liste des couples) sont des hapax . Les couples retenus ne sont donc pas très nombreux. Mais ils représentent environ 40% des couples de lexicaux du corpus (estimation faite sur le corpus *français03* étiqueté). Le tableau 5.5 donne quelques couples d'effectif deux.

Couple	Effectif
accords *de* commerce	2
accusés d'avoir	2
accélérer *le/son* processus	2
acheter *la* clinique	2
acteurs économiques	2
action d'occupation	2
action militaire	2
actions britanniques	2
actuellement détenus	2
activités *de* courtage	2

TAB. 5.5 – Couples de lexicaux ayant un effectif de 2. La quasi totalité des éléments formant ces couples sont en relation. Les éléments morphologiques du deuxième syntagme sont en italique (nous rappelons que *d'avoir* ne forme qu'un mot selon notre définition).

Il arrive que les éléments de ces couples ne soient pas en relation. Nous avons diagnostiqué deux types d'"erreur" :

- Les éléments ne sont pas en relation
- les deux éléments appartiennent à une structure plus grande

Certains couples (quelques pour mille) ne sont réellement pas en relation comme le couple *fidèle garde* qui provient des entre-ponctuations :

- , mais celui qui a l'esprit *fidèle les garde*
- . ce dieu *fidèle garde*

L'erreur la plus fréquente provient d'une structure particulière : [substantif + complément + Verbe], où le couple généré correspond aux lexicaux du complément et du verbe :

- the males of some few *quadrupeds possess* [...]
- very few male *quadrupeds possess* [...]

Cette structure génère plus de la moitié des erreurs.

Le deuxième type provient de structures lexicales incluant les couples considérés. Par exemple le couple *national développement* a un effectif de cinq, mais les deux éléments ne sont pas en relation. Ce couple appartient à une structure composée de trois éléments : *du/le fonds national de développement* ayant un effectif de cinq. Dans le cas de $n$uplets, la séquence est jugée correcte si chaque élément est en relation avec un autre élément du $n$uplet. Une séquence

comme *vendéens essaient grappiller*, provenant de *les producteurs vendéens essaient de grappiller quelques subventions auprès du conseil général.* n'est pas valide puisque *vendéens* n'est en relation ni avec *essaient*, ni avec *grappiller*. La mise est relation est bonne à 100% (estimation faite sur 100 triplets pris

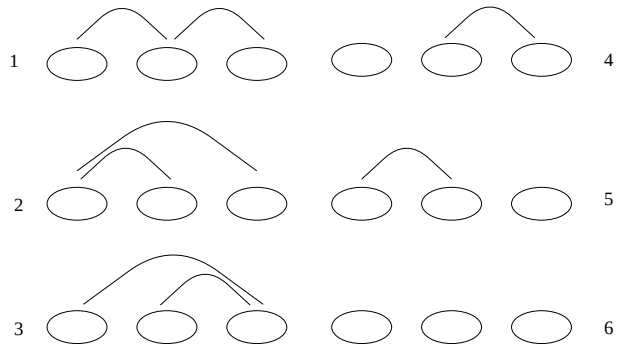


FIG. 5.1 – Les relations possibles entre trois éléments (en supposant qu'un élément n'entretient qu'une seule relation avec un autre élément). Si un triplet lexical à un effectif supérieur à un, il ne peut correspondre aux cas 4, 5, et 6.

au hasard en français). L'effectif des $n$uplets lexicaux décroît extrêmement vite. L'effectif du triplet le plus fréquent dans le corpus *français01* est de 17. Il existe seulement trois séquences de longueur 7 ayant un effectif de deux[69]. Si nous savons que les $n$uplets forment une séquence en relation, cela ne nous indique pas quelles sont les relations entre éléments du $n$uplet. Nous pouvons avoir toutes les possibilités (figure 5.1). Nous avons cru pouvoir déterminer les relations entre $n$uplets en utilisant les effectifs des couples composant le $n$uplet (en regroupant deux à deux les éléments les plus fréquents), mais le résultat est aléatoire et dépend de la structure des $n$uplets.

Sur cent couples ayant un effectif de deux pris au hasard et provenant du corpus *français01*, seuls deux n'étaient pas en relation (erreur type 1), et seize étaient inclus dans une structure supérieure (erreur type 2). Les deux couples n'étant pas en relation sont :

- (*politique, provisoire*)
- (*soviétique, proposé*)

Ils apparaissent dans les entre-ponctuations suivantes :

- *d'un dégel* ***politique provisoire***
- *son bureau* ***politique provisoire*** *est dirigé par mr khalifa abid*
- *l'union* ***soviétique*** *avait* ***proposé*** *35 %*
- *la délégation* ***soviétique*** *avait* ***proposé*** *que [...]*

Le taux de mise en relation est donc de 98% sur les couples d'effectif deux. Les autres langues n'ont pas été étudiées, car l'étude des relations nécessite une

69
- d'intérêt américains risquent s'orienter baisse avenir proche
- présidente croix-rouge française déléguée générale contre drogue
- successeur jean-marie tjibaou présidence mouvement indépendantiste nouvelle-calédonie

Triplet	Effectif
accords commerce coopération	2
activités courtage actions	2
américains risquent s'orienter	2
animer instances européennes	2
annoncé mercredi janvier	2
années titres possession	2
appartenaient peloton voltigeur	2
apporté soutien financier	2
arrêté ministres l'économie	2

TAB. 5.6 – Triplets de lexicaux. Ils correspondent systématiquement à des éléments en relation.

Quadruplet	Effectif
activités courtage actions britanniques	2
américains risquent s'orienter baisse	2
appartenaient peloton voltigeur motocycliste	2
atteinte exigences éthiques d'accueil	2
banque bilbao banque biscaye	2
blessures entrainé intention donner	2
budget conseil régional d'ile-de-france	2
candidat unique l'opposition l'élection	2

TAB. 5.7 – Quadruplets de lexicaux. Ils correspondent systématiquement à des éléments en relation.

compréhension totale de l'énoncé.

Cette étude ayant été faite principalement sur le français (très légèrement validée sur l'anglais, le turc et l'allemand), elle reste à valider sur les autres langues. Nous avons voulu généraliser ces couples en n'imposant pas d'ordre. Peu d'erreurs ont été générées, mais la généralisation a été très faible, les couples se rencontrant dans le même ordre la plupart du temps.

La distribution des éléments lexicaux est donc très contrainte, beaucoup plus que celle des éléments morphologiques au niveau des séquences de syntagmes. Elle semble utiliser des contraintes terminologiques. Est-ce que ces structures sont invariantes d'un corpus à un autre ? Si un corpus génère un couple *X, Y* dont l'effectif est de dix (disons *ministre de l'économie*), peut-on en conclure que ces éléments lexicaux sont toujours en relation quel que soit le corpus analysé ? Une autre question se pose : est-ce que les séquences de *n*uplets lexicaux sont toujours étiquetés de la même façon ? Si dans l'absolu la réponse est non, des études sont à effectuer pour quantifier ces dires. Il serait intéressant de voir si quelques pre-traitements effectués sur un texte à analyser permettraient une amélioration de l'analyse de ce texte. Les informations lexicales permettent aussi de valider certaines hypothèses faites sur le matériau grammatical. Une séquence morphologique très fréquente en français est la suivante : [SR] [*de X*]. Aucune information possitionnelle ne caractérise une structure syntagmatique commençant par le morphème libre *de*. Seule l'effectif de cette structure (c'est la structure la plus fréquente de nos corpus) nous incite à mettre en relation les deux syntagmes. Cette hypothèse est validée par les couples lexicaux : une grande majorité de ces couples admet le morphème *de* comme début du deuxième syntagme. De manière plus générale, l'étude des séquences morphologiques des couples de lexicaux peut nous aider à mieux comprendre le rôle de ces éléments morphologiques.

Si notre travail est principalement axé sur des données morphologiques, l'utilisation de données lexicales semble complémentaire. [Zuret, 1998] développe un algorithme d'apprentissage des relations qui se base sur ses propriétés lexicales. Nous donnons un avantage aux informations morphologiques, car elles semblent permettre plus facilement d'amorcer un système d'apprentissage.

5.3.4 Les variations morphologiques

Si cette méthode permet d'établir l'existence de certaines relations, elle ne donne aucun renseignement sur la relation entre deux éléments. De plus, cette relation peut être différente d'une occurrence à une autre. Ce changement s'accompagne généralement d'un changement morphologique d'un des syntagmes. Ainsi, le couple [*l'histoire, sciences*] apparaît dans les entre-ponctuations suivantes :

- fait revivre *l'histoire des sciences* et techniques
- sur le thème *histoire et sciences* sociales

Les deux éléments sont bien en relation, mais, dans un cas, c'est une relation de dépendance et dans l'autre cas une relation de coordination. Ce phénomène concerne surtout les couples ayant un effectif de deux, et devient très rare pour les autre *n*uplets (pas d'exemple rencontré). Il s'accompagne d'une variation

morphologique très caractéristique.

Ces variations morphologiques peuvent aussi intervenir sans modifier le type de la relation. Le couple allemand (*lautete, antwort*) apparaît avec les variations suivantes :

- [lautete] [*die* Antwort]
- [lautete] [*meine* Antwort]
- [lautete] [*seine* Antwort]

La variation peut aussi venir de l'ajout d'un élément (grammatical ou lexical) entre les deux éléments du couple :

- (turc) yüksek *bir* [sesle]
- (français) le ministe *néerlandais* des affaires étrangères
- (allemand) kennen *zu* lernen

La prise en compte de la variation morphologique est intéressante car assez fréquente ([Daille et al., 1996]). Elle permet une identification assez facile des SSub des langues.

5.3.5 Les couples lexico-morphologiques

Deux types de couples ont été étudiés : les couples composés de deux éléments morphologiques (les couples morphologiques) et les couples composés de deux éléments lexicaux (les couples lexicaux). Étudions maintenant le troisième type de couples possible : les couples composés d'un élément lexical et d'un élément morphologique. L'élément morphologique peut correspondre à un morphème libre ou lié. Il appartient à un syntagme différent de celui contenant l'élément lexical. Le tableau 5.8 en montre quelques exemples pour le français. Ce tableau a été construit manuellement à partir des couples fréquents comprenant un mot lexical comme premier élément. Pour réaliser une génération automatique de ces couples, l'identification des noyaux syntagmatiques est nécessaire. Ces structures

Couples	Effectif (éléments contigus)	Relation
ministre- de	127	127
mis- en	116	116
direct- de	73	61
conseil- -al	58	58
gouvern- de	41	39
comité -al	38	38
donn- à	23	23
renonc- à	18	18

TAB. 5.8 – Couples d'éléments noyau-morphème grammatical du corpus *français01*.

sont intéressantes parce que les deux syntagmes contigus à partir desquels les couples sont construits sont *très souvent* en relation. La validité de la relation dépend des deux éléments utilisés. Les éléments lexicaux ne sont pas toujours pertinents (*direct-* est reconnue dans des mots comme *directement*, *direction*,

directs) : chaque réalisation de *direct-* comme adjectif ou adverbe se traduit par une mauvaise mise en relation. La génération automatique de ces couples doit donc utiliser d'autres contraintes (ne retenir que les SR et SA par exemples). Elle demande donc une analyse du corpus en syntagmes et une catégorisation des syntagmes en SA, SR, SSub. Par contre, certains couples sont très fiables. Examinons la structure *donn- à*. L'effectif de cette structure est de 23. On peut ajouter l'effectif des structures *donn- aux* (16) et *donn- au* (5). Ce qui fait un total de 44 pour un corpus d'environ 300000 mots, ce qui est assez faible. Dès qu'une étude sur corpus porte sur des éléments lexicaux, la taille du corpus doit alors être très conséquente. Cherchons à intercaler des mots entre ces deux éléments. Nous trouvons alors 43 séquences inclues dans une entre-ponctuation. Dans notre corpus *français01*, la relation n'est mise en défaut que trois fois, quelque soit le nombre d'élément intercalés. Le syntagme commençant par *à* correspond à l'élément que l'on donne sauf dans le cas suivant :

- ***donnait à nouveau*** *le feu vert à edf*

Les éléments intercalés correspondent soit à un adverbe, soit au deuxième actant (la chose que l'on donne).

- ***donner*** *une seconde existence et un rayonnement international* ***à des manifestations***

Les trois exceptions sont :

- ***donne*** *une idée des dégâts que des virus pourraient causer s'ils parvenaient* ***à déjouer*** *tous les verrous de sûreté mis par les techniciens pour protéger les ordinateurs ou*
- ***données*** *que la décision de renoncer* ***à une opa*** *dans l'immédiat a été prise*
- ***donner*** *du liant et de mettre* ***à l'aise*** *ses clients*

L'ajout de contraintes structurelles est donc nécessaire afin d'améliorer cette mise en relation. Néanmoins, il semble que certains noyaux aillent un très fort pouvoir attracteur sur certaines séquences morphologiques : le noyau *donn-* attire à lui les syntagmes commençant par *à*. Ce couple a été étudié sur le corpus *français02* qui contient vingt millions de mots (tableau 5.9).

On retrouve ces données dans les travaux sur le rattachement au verbe de groupes prépositionnels en anglais. La structure étudiée est *SV SN SP* : (SV : syntagme verbal, SN : syntagme nominal, SP : syntagme prépositionnel). Le syntagme prépositionnel peut se rattacher au syntagme verbal ou au syntagme nominal. La technique habituelle (avec quelques variantes) illustrée dans [Hindle and Rooth, 1993], [Collins and Brooks, 1995] est de calculer l'effectif du couple (*verbe, préposition*), et (*nom, préposition*). Cet effectif permet alors de choisir la relation la plus probable. Les taux sont de l'ordre de 80% de réussite.

On voit donc que, si ces structures ne permettent pas une mise ne relation sûre, elles contiennent des informations intéressantes. Reste à savoir comment les utiliser au mieux.

nb mots intercalés	nb séquences	nb de séquences non en relation
0	689	0
1	475	8
2	437	14
3	347	21
4	178	10
5	109	40
10	39	20
Total	2274	113

TAB. 5.9 – Évaluation du taux de mise en relation de la structure *donn- à*. Les éléments intercalés ne comprennent pas de ponctuation. Les cas d'erreur proviennent soit des mots *donne* et *données* en tant que substantif, soit d'un verbe de la séquence intercalée qui attire lui même le *à* (*commenc-*). La relation se dégrade fortement après une séquence intercalée de cinq mots.

5.4 La classification des éléments lexicaux

Les catégories générées dans ce travail l'ont étées en utilisant des critères purement formels. Certains travaux essayent, non pas de générer des catégories formelles, mais lexicales. Ces classes contiennent des mots ayant des affinités sémantiques comme les classes suivantes trouvées dans [Huckle, 1995] :

- *boy, girl, man , woman*
- *months, years, days, hours, o'clock, times*
- *six, twelve, twenty, two, three, four, ten, five, seven*

On trouve parfois le terme de classification sémantique (*semantic clustering*). Nous préférons la dénommer *classification lexicale*, puisqu'elle consiste à classer les éléments lexicaux de textes. Certains travaux essayent de regénérer (ou d'aider à une génération) une ontologie d'un domaine [Bouaud et al., 1997]. La difficulté est d'évaluer la pertinence des classes de mots obtenues, tâche d'autant plus difficile que le nombre de classes obtenues peut atteindre plusieurs centaines voir plusieurs milliers. Seuls les travaux se rapportant à un domaine bien précis (comme ceux de [Bouaud et al., 1997] qui compare leurs résultats à une ontololgie déjà existante) peuvent être évalués. Il faut plutôt prendre ces travaux comme des expérimentations sur la langue (au stade actuel). Les corpus utilisés peuvent être annotés et/ou étiquetés [Bouaud et al., 1997] ou non [Schütze, 1993], [Pereira et al., 1993], [Honkela, 1997]. La taille des corpus utilisés dans ces études peut atteindre plusieurs centaines de millions de mots. Le principe est similaire aux algorithmes décrits en 3.3.3, la fenêtre définissant le contexte pouvant atteindre une centaine de mots.

Quatrième partie

Les algorithmes

Introduction

Où comment se servir de tout ce que l'on vient de dire. Ceci n'est qu'une utilisation possible des concepts développés dans la partie précédente. Elle est minimaliste et n'a qu'un objectif de *validation* des concepts développés précédemment.

Les algorithmes présentés ici sont axés sur la structure syntagmatique. L'importance de la structure propositionnelle n'est apparue qu'assez tard dans le travail. Il semble indispensable d'intégrer mieux cette structure dans le processus de découverte. Les résultats de ces algorithmes sont donnés en annexe pour différentes langues. Ils ont été obtenus de manière totalement automatique. Aucune supervision n'a été effectuée.

Chapitre 6

La catégorisation des éléments

Sommaire

Dans ce chapitre, le détail du processus de découverte des structures est donné. Nous allons illustrer ce processus principalement à travers le français. Les résultats sur diverses autres langues sont donnés en annexe. Le chapitre 4 nous offre les catégories à construire. Ces catégories sont nées de l'observation des corpus. Elles ont été sélectionnées parmi d'autres parce qu'elles possèdent des caractéristiques formelles très fortes qui facilitent leur traitement. Ces caractéristiques rendent les éléments (mots et morphèmes) des classes assez facilement identifiables. L'originalité de ce travail de catégorisation repose sur la prise en compte de la polycatégorisation des éléments. Nos algorithmes peuvent affecter à un même élément plusieurs catégories, ce qui n'est pas le cas dans les autres travaux (en TAL et en catégorisation). Les éléments que nous cherchons à catégoriser sont les *mots* et les *affixes* du corpus. La prise en compte de la polycatégorisation nous a obligé à délaisser les algorithmes de catégorisation généralement utilisés (algorithme de *clustering*), et à développer notre propre

méthode qui repose sur la construction de contextes prototypiques pour chaque catégorie identifiée.

Les algorithmes développés sont simples, mais ils permettent de valider les considérations théoriques décrites dans le chapitre 4. Afin d'améliorer ces résultats, une implémentation plus poussée serait nécessaire. Elle n'a pas été réalisée, l'objectif de ce travail n'étant pas la réalisation d'un système opérationnel. Nous ne pensons pas que, dans le domaine de l'analyse syntaxique, un système généré automatiquement puisse rivaliser avec un système conçu par un humain. Le problème du goulot d'étranglement (le fameux "bottleneck"), que certains [van den Bosch et al., 1996] pensent résoudre par une automatisation de l'acquisition des connaissances, ne semble pas se poser en analyse syntaxique puisque, comme l'a montré [Vergne and Giguet, 1998] peu de règles permettent de gérer une grande partie des mises en relation entre mots, et que les relations restantes, qui nécessitent, il est vrai, une assez grande quantité de règles, ne peuvent être traitées qu'avec des règles très fines qui semblent difficiles à générer automatiquement (la construction *ne ...que* française, par exemple). Nous décrirons les différents algorithmes utilisés puis nous donnerons les évaluations à la section 6.5. Comme pour l'algorithme de segmentation, la mise au point de ces algorithmes s'est faite sur plusieurs langues simultanément.

6.1 La tokenisation

Une fois le corpus obtenu, le premier traitement consiste à le formater afin de le préparer aux traitements suivants. Cette préparation consiste premièrement à segmenter le corpus en mots[70] en insérant un et un seul blanc comme séparateur de mots. Deuxièmement à segmenter le corpus en entre-ponctuations et à mettre une et une seule entre-ponctuations par ligne, le signe de ponctuation se trouvant en tête de la ligne. Le choix de ce format est historique, et n'a pas été modifié par la suite (Ce segment (l'entre-ponctuations) est très adapté à l'étude du syntagme). Voici la première phrase du corpus français (*français01*) et sa version formatée :

Le programme de tokenisation (écrit en flex) est donné en annexe B. Cette opération de formatage est appliquée sur les systèmes alphabétiques, mais aussi sur les autres systèmes. Pour les systèmes non alphabétiques (chinois, japonais), le mot est défini étant comme le symbole graphique (section 1.8). On trouve des travaux qui segmentent des textes chinois en "mots" ([Sproat et al., 1994]), mais pourquoi vouloir segmenter un texte chinois en unités qui appartiennent à un autre système d'écriture et une autre langue. En effet, le but de la plupart des travaux est d'obtenir une segmentation des signes chinois qui correspond à une segmentation en mots anglais. Le principal problème durant cette opération de formatage est dû aux systèmes mixtes comme le japonais (idéographique et syllabique). La découverte du système d'écriture est un préalable à toute autre manipulation informatique. Ayant travaillé surtout sur des langues utilisant un système alphabétique, nous n'avons pas développé de méthode permettant une découverte automatique d'un système d'écriture (recensement des

[70] La définition du mot est donnée à la section 1.9.

```
Du reste, ne l'avoue-t-il pas en partie
lorsqu'il déclare : " A ce poste, les aller-retour sont gênants " ?
Une incompétence avouée en matière de choix des gardiens de but,
un grand ancien qui se laisse désirer, un remplaçant en quête de
promotion...
```

```
. du reste
, ne l'avoue-t-il pas en partie lorsqu'il déclare
: a ce poste
, les aller-retour sont gênants
? une incompétence avouée en matière de choix des gardiens de but
, un grand ancien qui se laisse désirer
, un remplaçant en quête de promotion
.
.
.
```

signes de ponctuations, des signes composant les mots).

6.2 Les opérations morphologiques

Les opérations morphologiques (segmentation, réécriture des corpus, génération des couples morphologiques) sont décrites en détail dans le chapitre 3. Nous ne reviendrons donc pas dessus. Nous rappellerons seulement les résultats obtenus par celles-ci :

- une liste d'affixes (section 2.2.2)
- un corpus segmenté (section 3.1)
- une liste de couples morphologiques (section 3.2)

Le corpus segmenté correspond au corpus tokenisé dont les mots ont été segmentés. En voici un exemple en français :

```
. Les err-eurs des spéci-alistes de la planifi-cation urbaine
au cours des dernières dé-cenn-ies ont été nombr-euses
```

Les dix couples morphologiques les plus fréquents du corpus *francais01* et *vietnamien01* sont donnés au tableau 6.1 Dans les langues où aucune segmentation n'est réalisée, la liste des couples correspond à celle des mots contigus du corpus (exemple vietnamien). Les couples comprenant une ponctuation sont éliminés de la liste pour la suite du traitement (ils ne sont pas utilisés).

6.3 La recherche des éléments prototypiques

Nous allons maintenant étudier le critère positionnel des éléments. Pour ce faire, il suffit de recenser chaque élément du corpus (mot, morphème, couple morphologique), et de calculer leurs positions par rapport aux ponctuations (algorithme 6). Pour chacun de ces éléments, une liste comportant leur effectif est calculée. Ce recensement sert à ne prendre en compte que les éléments fréquents dans un premier temps. En effet, les hapax étant assez nombreux (dans les

de la	2423	ñöùc gieâsu	750
à la	980	caùc ngöôi	653
de l'-N	901	noùi vôùi	351
l'-N N-e	571	thieân chuùa	349
des N-es	561	anh em	344
les N-es	555	caùc oâng	221
la N-e	522	ngöôøi ta	169
à l'-N	515	chuùng toâi	151
et de	463	baáy giôø	135
dans le	390	moân ñoà	126

TAB. 6.1 – Les dix couples morphologiques les plus fréquents du corpus *français01* et *vietnamien01*.

listes générées), leur élimination permet un gain de temps appréciable dans les traitements. Le comptage en fin de ligne correspond aux occurrences avant une ponctuation (grâce au formatage du corpus). Le même algorithme est appliqué au niveau des morphèmes et des couples morphologiques.

Algorithme 6 Génération des positions des éléments

pré-requis un corpus
  pour tout mot du corpus **faire**
    compter son nombre d'occurrences
    compter son nombre d'occurrences en fin de ligne
    compter son nombre d'occurrences après une ponctuation
    compter son nombre d'occurrences situées après une ponctuation et en fin de ligne (singleton)
  fin pour

Le résultat de ces opérations fournit trois listes :
- la liste des mots et leur position
- la liste des morphèmes et leur position
- la liste des couples morphologiques et leur position

Nous appellerons par la suite ces données les *listes positionnelles*. Les tableaux 6.2 donnent quelques exemples des fichiers générés du corpus *français01*. Une marque (*D2*[71] ou *F2*) est ajoutée en fin de ligne si l'élément est considéré comme prototypique, c'est-à-dire que son nombre d'occurrences de début ou de fin est supérieur à la moitié[72] de son effectif total (par exemple l'élément *nous N-ons* est un élément prototypique de début absolu). La colonne *D&F* recense le nombre d'occurrences d'un élément compris entre deux ponctuations. Les éléments qui apparaissent souvent dans cette position correspondent généralement

[71] "Historiquement" les D1 et F1 sont les marqueurs de frontière de syntagme (niveau 1) et les D2 et F2 les marqueurs de frontière de proposition (niveau 2).

[72] Si aucun élément n'est sélectionné avec cette valeur, nous la diminuons de 10 en 10 jusqu'à sélectionner des éléments morphologiques (le cas se produit dans le corpus *latin01*, où la ponctuation est inexistante).

Mot	Eff.	Début	Fin	D&F	
de	14943	648	3	0	
la	8427	1300	0	0	
le	6504	1893	0	0	
...					
et	5311	760	115	34	
des	4750	304	0	0	
...					
il	1605	1195	0	0	D2
pas	1523	54	88	0	
est	1491	128	34	1	

Morphème	Eff.	Début	Fin	D&F
N-e	4235	385	1225	61
N-es	2866	90	689	23
N-er	1844	137	449	14
N-é	1474	200	385	12
N-ent	1324	153	242	20
N-ement	1115	135	318	37
N-ant	935	279	188	32
N-ée	860	128	266	14
N-ie	836	94	299	22
N-ique	802	35	394	10

Couple	Eff.	Début	Fin	D&F	
de la	2423	90	1	0	
à la	980	108	0	0	
de l'-N	898	46	272	12	
N-e de	664	54	2	0	
...					
il est	176	133	2	1	D2
ont N-é	175	24	15	2	
de ses	174	14	0	0	
...					
N-er un	100	5	0	0	
nous N-ons	99	58	2	1	D2
les con-N	99	32	11	0	

TAB. 6.2 – Calcul des positions des différents éléments (morphèmes, mots, couples morphologiques).

à des interjections ou à des adverbes et groupes adverbiaux (corpus anglais : *why*, *oh*, *yes*, *however*, *therefore*). Seuls les couples assez fortement liés seront pris en compte. Si le nombre d'éléments intercalés est supérieur à l'effectif du couple, le couple est éliminé (algorithme 7). Ces couples sont considérés comme peu fiables. Ils sont constitués d'éléments (mot ou affixes) très fréquents de la langue. Cette heuristique enlève en fait assez peu d'éléments de la liste (généralement une conjonction suivie d'un suffixe non discriminant comme le couple anglais *and N-e*), mais le bruit généré par ces éléments pouvait parfois dégrader énormément les résultats. Ces données vont nous servir de point de départ dans la catégorisation des éléments.

Algorithme 7 Élimination des couples mineurs

pré-requis CM : la liste des couples morphologiques
  pour tout élément c de CM **faire**
    calculer le nombre I de mots intercalés entre les deux éléments de c
    si $I \geq$ effectif(c) **alors**
      éliminer le couple c de CM
    fin si
  fin pour

6.4 La catégorisation des marqueurs de frontière

Nous allons maintenant détailler le processus de catégorisation des éléments. Dans cette section, le terme *génération d'une structure* signifie instancier cette structure (dans notre cas les différents types de syntagmes) pour une langue donnée. C'est-à-dire trouver les éléments (mots, morphèmes) qui interviennent dans sa composition.

6.4.1 L'ordre de catégorisation

La catégorisation des éléments ne se fait pas dans un ordre quelconque (figure 6.1). Plus une structure est formellement marquée, plus il est facile de mettre au point un algorithme qui la génère. Nous commençons par la génération des Syntagmes Absolus (SA de début et SA de fin). Puis, nous travaillons sur les Syntagmes Relatifs. Enfin, sur les Syntagmes Subordonnés (des SR et des SA). Pourquoi commencer par les SA ? Il y a deux raisons à cela. Premièrement, ils sont formellement mieux marqués que les SR : ils possèdent une contrainte positionnelle supplémentaire qui est très forte. Ils sont donc plus facile à générer que les SR. Deuxièmement, les SA aident à la construction des SR. Il est parfois même nécessaire de connaître les SA pour parvenir à construire les SR (section 6.4.5) L'ordre entre SAD et SAF provient simplement du fait que les marqueurs de début de proposition sont généralement plus fréquents. Sinon, l'ordre n'est pas important. Une fois les SAD et SAF traités, nous procédons à une analyse du corpus pour marquer ces éléments (section 6.4.8). Pour ce faire, nous insérons une marque de début et de fin de syntagme dans le corpus. Cette marque dépend de la nature du syntagme : *SAD- -SAD* pour les SAD, *SAF-*

-SAF pour les SAF, et *SR- -SR* pour les SR. Après le traitement des SAD, l'entre-ponctuations suivante :

> *, il en-visag-eait de négocier un ac-cord-cad-re avec la fédér-ation qui les re-group-e*

est réécrit en :

> *, SAD-il en-visag-eait-SAD de négocier un ac-cord-cad-re avec la fédér-ation SAD-qui les re-group-e-SAD*

Par l'insertion de ces marques, nous mettons à disposition la connaissance déjà acquise pour l'étape suivante. Ainsi, le travail sur les SA bénéficie à la génération des SR qui peut utiliser cette segmentation.

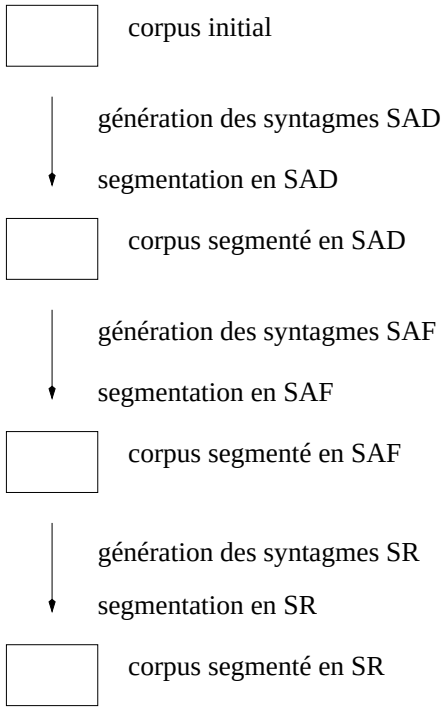


FIG. 6.1 – Ordre de traitement des syntagmes SA et SR.

Nous commençons donc par l'unité la plus haute de la hiérarchie . Une fois cette unité traitée, nous nous occupons des SR. Les SSub ne peuvent être traités que si leurs régissants sont déjà traités. Ils sont donc traités en dernier et de manière différente (section 6.4.6).

Pour chacune des structures SA(D|F) et SR, le principe de construction est similaire. L'algorithme comporte trois étapes principales (algorithme 8). La première est celle de la construction des *contextes prototypiques*. Ces contextes sont construits à partir des listes positionnelles d'éléments. Ils servent d'amorce au processus de catégorisation. Leur construction est expliquée à la section suivante. La deuxième étape consiste à rechercher les autres éléments qui peuvent former la structure en utilisant ce contexte prototypique. Nous partons d'un contexte qui contient un noyau lexical, et nous recherchons tous les marqueurs

de début et de fin possibles pour cette structure en utilisant les contextes appropriés. Cette opération est appelée *généralisation du contexte* (section 6.4.3). La troisième phase consiste à segmenter le corpus en utilisant la structure générée (section 6.4.8).

Algorithme 8 Algorithme de génération d'une structure

pré-requis C : un corpus segmenté en morphèmes
  Construire le contexte prototypique pour la structure
  tant que de nouveaux éléments ont été catégorisés **faire**
    Généraliser le contexte
    Analyser le corpus avec les structures trouvées.
  fin tant que

6.4.2 La génération des contextes prototypiques

Voyons comment les contextes prototypiques de SA sont construits. Les contextes prototypiques de SR utilisant les SA, ils seront décrits à la section 6.4.5. Pour cela, nous utilisons les listes positionnelles générées à la section 6.3. Nous commençons par recenser les éléments prototypiques de SA (SAD ou SAF). Ce sont les éléments marqués D2 ou F2 dans les listes positionnelles. L'algorithme consiste à rechercher tous ces éléments et à les regrouper par classes distributionnelles. Il se peut en effet que la liste des éléments prototypiques contiennent des éléments hétérogènes. Ainsi la liste française des couples morphologiques marquées D2 (SAD) (tableau 6.3) contient des structures verbales, mais aussi des structures nominales du type *mr X*, cette structure étant souvent précédée d'une virgule dans le corpus.

Couple	Eff.	Début	Fin	D&F	
il est	176	133	2	1	D2
il y	168	111	0	0	D2
le monde	160	87	20	0	D2
il a	148	131	7	7	D2
le président	126	64	5	2	D2
nous N-ons	99	58	2	1	D2
ils N-ent	95	68	19	7	D2
le ministre	90	52	5	1	D2
il ne	84	64	0	0	D2
il N-e	84	57	14	8	D2

TAB. 6.3 – Liste de certains couples morphologiques prototypiques de SA.

Pour éviter que le contexte prototypique ne soit construit avec des éléments hétérogènes, une classification est donc opérée (algorithme 9). Celle-ci utilise un simple algorithme de clustering qui regroupe les couples partageant un même environnement. Cet environnement est constitué soit des mots intercalés entre les deux éléments du couple, soit des mots apparaissant à gauche ou à droite.

La sélection du contexte s'effectue en prenant le contexte comprenant le plus de mots.

Algorithme 9 Algorithme de classification des couples morphologiques prototypiques.

pré-requis le corpus C
pré-requis la liste des couples morphologiques caractéristiques du syntagme choisi
pour tout couple de C **faire**
Générer les n mots les plus fréquents :
-intercalés entre les deux éléments de la structure
-à droite de la structure
-à gauche de la structure
fin pour
Prendre la liste l la plus longue.
Créer un cluster par élément
tant que Il y a plus d'un cluster **faire**
trouver les deux clusters qui ont les deux listes les plus proches
créer un nouveau cluster contenant les deux clusters
associer au nouveau cluster une liste composée des deux listes des deux clusters
éliminer les deux clusters de la liste de clusters.
fin tant que
Éliminer les clusters singletons
sortie : une liste de liste d'éléments

Le tableau 6.4 montre le résultat obtenu sur le français en considérant les éléments intercalés. Nous voyons bien que tous les couples ne partagent pas à une même distribution. Chaque ligne est constitué d'un couple et d'une liste. Le premier couple correspond au couple considéré. La liste suivante correspond aux éléments intercalés suivis de leur effectif dans ce contexte (*ne* apparaît 18 fois entre *il* et un mot finissant par *ait*). Les éléments *il N-a* et *qui N-it* ont ainsi trois éléments en commun dans leur liste. Un cluster comprenant ces deux éléments est ainsi créé. Ce cluster possède alors la liste d'éléments suivante : *se, ne, le, a, leur*. Le résultat final de cette clusterisation est donné par le tableau 6.5. Les listes générées vont ensuite servir de point de départ à la généralisation de ces structures. En pratique, seule la liste la plus longue est utilisée.

Il existe aussi une autre contrainte sur les éléments retenus. Seuls ceux qui possèdent un noyau syntagmatique sont retenus, le principe général de la catégorisation étant de partir d'une structure comprenant un noyau syntagmatique et de rechercher les marqueurs de frontière de ce noyau. Comment savoir si un élément obéit à cette contrainte ? La réponse est facile pour les langues qui ont générés une liste de morphèmes : tout mot segmenté est considéré comme possédant un noyau syntagmatique (noyau lexical) qui est la séquence ne correspondant pas à un affixe de la langue. Par exemple, la partie *transform* comporte au moins un noyau syntagmatique puisque la segmentation du mot *transformation* est *transform-ation*. Il suffit donc de travailler avec la liste des morphèmes

il N-ait :	ne 18 y 4 n'y 4 le 2
nous N-ons :	ne 17 le 7 nous 6 les 3
ils N-aient :	ne 1 se 1 y 1 la 1
il N-e :	est 25 faut 16 ne 14 se 14
mr N-is :	georges 2 andré 1 gorbatchev 1 maurice 1
on N-ait :	ne 10 se 3 les 2 lui 2
elles N-ent :	ne 3 trans- 1 an-nihil-ent 1 plong-ent 1
on N-e :	ne 15 lui 6 peut 4 se 4
mme N-e :	dupu-y 1 hélèn-e 1 nicol-e 1
il N-rait :	ne 5 lui 1
elle N-e :	est 20 se 8 devrait 4 le 3
mr N-i :	rajiv 2 tadeusz 2 pierre 2 jean 2
il N-ra :	ne 3 le 3 leur 1 se 1
je N-ais :	ne 6 n'ai 3 leur 2 le 2
sans N-er :	faire 2 os-er 2 doute 1 bourse 1
qui N-it :	se 15 ne 7 le 7 a 4
mr N-o :	marian-o 3 karl 2 pedr-o 2 jean-pier-re 2
ils N-ent :	se 9 ne 6 sont 4 ont 2

TAB. 6.4 – Calcul du contexte des couples morphologiques. Le contexte est ici composé des éléments intercalés.

et des couples morphologiques qui admettent un affixe. Quatre séquences morphologiques correspondent à cette contrainte :

- D-N
- N-F
- D N-F
- D-N F

Ces structures matchent des éléments comprenant assurément un noyau lexical. La structure [*N-F*] matche tous les *mots* finissant par la séquence *F* (qui est un affixe de la langue). La séquence [*D N-F*] matche tous les couples de mots dont le premier est *D* et le second un mot finissant par *F*. Le terme *N* est donc un élément qui comprend une séquence correspondant à un noyau. Dans la liste du tableau 6.3 qui nous montre la liste des éléments prototypiques de SAD, les éléments *il est*, *il y*, *le monde*, *il a*, *le président*, *le ministre*, *il ne* ne sont donc pas pris en compte dans la construction des contextes. Seuls les éléments comprenant un affixe le seront :*nous N-ons*, *ils N-ent* et *il N-e*.

Pour les langues qui n'admettent pas de segmentation morphologique (comme le vietnamien) la construction des contextes prototypiques est différente. Elle n'a pas été implémentée et le principe a seulement été testé manuellement (avec succès). Pour remplacer les séquences morphologiques, nous construisons des *classes lexicales*. Nous partons d'un couple de marqueurs de frontière prototypique (un seul élément n'est pas assez discriminant en général) et cherchons la liste des mots suivant (pour les débuts) ou précédant (pour les fins) le couple de marqueurs de frontière. Ces mots ne doivent pas être eux-mêmes des marqueurs de frontière (on obtient bien en pratique des éléments lexicaux). La liste

il N-ait	il N-ra ils N-aient il N-ait qui N-it je N-ais nous N-ons
nous N-ons	il N-ra on N-ait il N-ait qui N-it je N-ais nous N-ons
ils N-aient	on N-e il N-ait il N-ra ils N-aient il N-e on N-ait qui N-it ils N-ent
il N-e	il N-ra ils N-aient on N-e il N-e on N-ait qui N-it ils N-ent elle N-e
mr N-is	mr N-is
on N-ait	on N-e il N-rait nous N-ons il N-ra ils N-aient il N-e on N-ait qui N-it ils N-ent
elles N-ent	elles N-ent
on N-e	on N-e il N-rait il N-ra ils N-aient il N-e on N-ait qui N-it ils N-ent
mme N-e	mme N-e
il N-rait	on N-e on N-ait il N-rait
elle N-e	il N-ra il N-e qui N-it elle N-e
mr N-i	mr N-i
il N-ra	on N-e il N-ait nous N-ons elle N-e il N-ra ils N-aient il N-e on N-ait qui N-it je N-ais ils N-ent
je N-ais	il N-ra il N-ait qui N-it je N-ais nous N-ons
sans N-er	sans N-er
qui N-it	on N-e il N-ait nous N-ons elle N-e il N-ra ils N-aient il N-e on N-ait je N-ais qui N-it ils N-ent
mr N-o	mr N-o
ils N-ent	il N-ra ils N-aient on N-e il N-e on N-ait qui N-it ils N-ent

TAB. 6.5 – Résultat de la clusterisation des éléments

remplace l'élément N des séquences morphologiques.

Nous voyons qu'il existe trois objets qui permettent la construction des contextes :

- les morphèmes seuls (appelée structure morphémique par la suite)
- les couples morphologiques
- les classes de lexicaux

La structure morphémique correspond à un modèle [*N-m*] ou [*m-N*] où N représente un noyau quelconque et m un morphème (par exemple la structure morphémique turque [*N-dır*] qui caractérise un verbe turc). Pour une langue donnée, la recherche de la structure à considérer se fait dans l'ordre suivant : morphème, couple, et classe. Si la langue admet des morphèmes seuls comme éléments prototypiques, nous traitons d'abord ces éléments. Ce cas ne se produit pas en français, mais il se produit en turc (annexe C.5). Puis nous passons aux couples morphologiques. Enfin si la langue ne produit aucun couple morphologique (comme le vietnamien), nous construisons les classes de lexicaux. Une fois ces éléments prototypiques identifiés, les traitements suivants sont similaires dans le principe.

6.4.3 Le mécanisme de catégorisation

Prenons la liste des éléments prototypiques obtenue grâce à la section précédente. Elle comprend les éléments suivants :

> *on N-e, il N-ait, nous N-ons, elle N-e, il N-ra, ils N-aient, il N-e, on N-ait, qui N-it, je N-ais, ils N-ent*

Cela nous donne les renseignements suivants : Les éléments *on, il, nous, elle, ils, qui, je* sont des marqueurs de début libres de proposition[73] ou de SAD. Les morphèmes *-e, -ait, -ons, -ra, -aient, -it, -ais, -ent* sont des marqueurs de fin liés de SAD (figure 6.2).

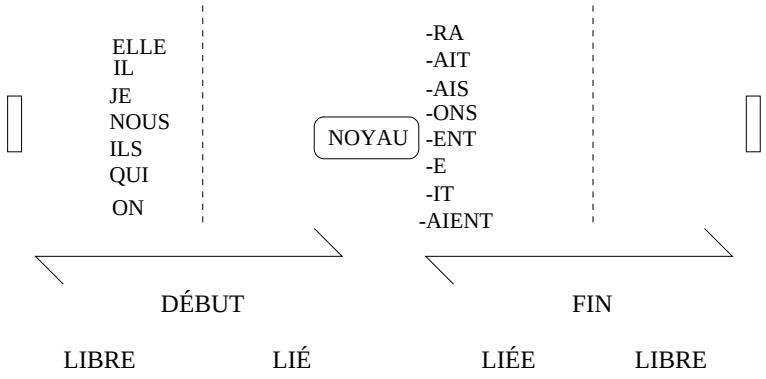


FIG. 6.2 – Une structure *D N-F* correspond à un marqueur de début libre (D) suivi d'un noyau syntagmatique (N) suivi d'un marqueur de fin lié *F*.

À partir de ces éléments qui contiennent un noyau syntagmatique, nous allons maintenant essayer de trouver de nouveaux marqueurs de frontière à ce

[73] À ce stade, il est impossible de distinguer marqueur de début de SAD et de proposition

noyau. Nous utilisons la ponctuation pour délimiter le contexte syntagmatique (les barres à gauche et à droite des figures). Les marqueurs de début peuvent apparaître à trois endroits : entre le noyau et une ponctuation (situation des marqueurs déjà trouvés), mais aussi avant les marqueurs de début déjà sélectionnés (1), et entre ces marqueurs et le noyau (2) (figure 6.3).

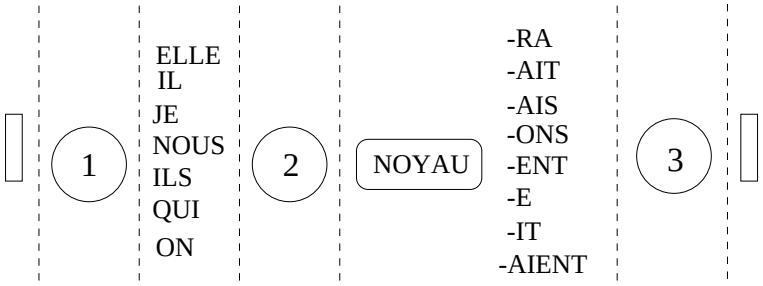


FIG. 6.3 – Les différentes positions pour le contexte SAD français. Les éléments apparaissant aux positions (1) et (2) correspondent à des marqueurs de début (ils sont à gauche du noyau), et les éléments apparaissant à la position (3) sont des marqueurs de fin.

Nous appelons *position* les endroits où un élément peut s'intercaler dans le contexte entre les différents éléments le constituant (ponctuation, mot, morphème). La liste de toutes les positions est donnée à la figure 6.4. *Cette figure décrit les différents contextes utilisés pour générer les SA*. Toutes ces positions ne sont pas fructueuses pour toutes les langues, mais il est nécessaire de les prendre en compte systématiquement puisque nous ne savons pas *a priori* lesquelles sont pertinentes pour une langue donnée. Il n'est pas nécessaire de chercher les éléments s'intercalant entre une ponctuation et la position (1). Ils apparaissent eux-mêmes en position (1). De même, nous ne cherchons pas à "étaler" les différents marqueurs de fin. La position (3) suffit à tous les recenser. La section 6.7 revient sur ce propos (tous les marqueurs de frontière ne sont pas équivalents dans un syntagme).

Prenons le cas de la recherche de marqueurs de début libres (donc des mots) en position (1). Le contexte utilisé pour catégoriser ces éléments est le suivant : nous allons rechercher tous les mots qui apparaissent dans le corpus entre une ponctuation et l'une des séquences matchant les modèles suivants :

> *on N-e, il N-ait, nous N-ons, elle N-e, il N-ra, ils N-aient, il N-e, on N-ait, qui N-it, je N-ais, ils N-ent.*

Tous ces éléments sont-ils réellement des marqueurs de début ? La première idée est de ne sélectionner que les éléments qui apparaissent fréquemment dans ce contexte. L'inconvénient de cette méthode est qu'elle ne permettra de catégoriser que les marqueurs (très) fréquents. De plus elle n'est pas absolument fiable même avec un seuil très élevé. Dans notre corpus espagnol, le modèle *que N-e* qui n'est pas assez caractéristique des SAD permet la catégorisation de *tierra* comme marqueur de début. La séquence

> *[ponctuation] tierra que N-e*

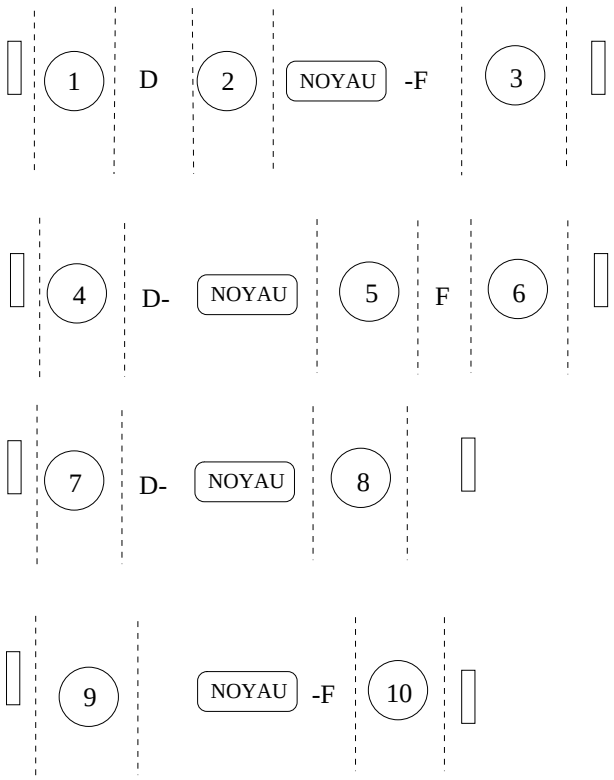


FIG. 6.4 – Liste de toutes les positions possibles (1 à 10) pour les différentes structures (morphème seul et couples). Les positions 1, 2, 4, 7, et 9 correspondent à des marqueurs de début, les positions 3, 5, 6, 8 et 10 à des marqueurs de fin. Les contextes sont limités par des ponctuations. Les traits pointillés verticaux indiquent les séparateurs de mots.

se rencontre 7 fois, ce qui correspond à un seuil élevé en pratique. Si un tel seuil était utilisé, le nombre de mots pris en compte serait très faible (les deux ou trois mots les plus fréquents de la catégorie). Or *tierra* n'est pas un marqueur de début de SAD ni de proposition (même avec la meilleure volonté du monde). Son effectif est uniquement due à l'entre-ponctuations :

, tierra que fluye leche y miel

Nous n'utilisons donc pas l'effectif d'un mot, mais un critère que nous appelons la *diversité morphologique*. Prenons l'exemple de la position (1) de la figure 6.3. Les éléments retenus sont :

comme, et, mais, si

Le mot *comme* est sélectionné, non pas grâce à son effectif, mais parce qu'il apparaît avec quatre séquences différentes : *il N-e*, *il N-ait*, *on N-e*, *nous N-ons* (tableau 6.6). C'est ce nombre de couples morphologiques que nous appelons la

Séquence	Effectif
, comme il aime	1
, comme il l'avait	1
, comme il étrangle	1
, comme nous l'avons	2
, comme on agace	1
. comme il n'avait	1
. comme il n'existe	1

TAB. 6.6 – Le mot *comme* n'est pas sélectionné grâce à son effectif d'apparition dans le contexte (8), mais grâce à la variété morphologique de son contexte qui comporte quatre structures différentes : *il N-e*, *il N-ait*, *on N-e*, *nous N-ons*.

diversité morphologique d'un élément. Un mot doit apparaître dans une position donnée grâce à *quatre* couples différentes. Le chiffre de *quatre*[74] a été retenu car il assure une assez grande diversité et est assez faible pour permettre la catégorisation de nombreux éléments. Ceci palie le fait que certains couples ne caractérisent pas suffisamment une structure (comme *que N-e* en espagnol ou en français). Nous utilisons donc plusieurs couples pour augmenter le degré de confiance de la catégorisation.

6.4.4 La génération des structures SA

Nous allons détailler la génération de la structure SAD en français. Pour les SAF, le principe est le même, les différences proviennent des contextes utilisés (le travail se fait en considérant les fins d'entre-ponctuations au lieu des débuts). L'algorithme consiste à rechercher les éléments apparaissant dans les positions (1), (2) et (3) de la figure 6.4. Après chaque recherche d'éléments, nous intégrons ceux-ci au contexte afin d'augmenter la *diversité morphologique* possible

[74] Dans la dernière version, ce seuil n'est plus fixe mais dépend du nombre de couples morphologiques utilisés. On a s =f(nbcouples), avec $2<s<7$.

et catégoriser un plus grand nombre d'éléments. Une fois ce travail effectué sur toutes les positions possibles, l'opération est réitérée jusqu'à ce qu'aucun autre élément ne soit catégorisé. Pour la deuxième itération, les contextes sont donc augmentés de tous les nouveaux éléments obtenus dans la première itération. Le nombre d'itérations se situe généralement entre trois et cinq selon les langues (en particulier selon la diversité morphologique du syntagme traité). Nous commençons par la position (2), puis (1), et enfin (3), les marqueurs de début étant plus fréquents dans les langues étudiées que les marqueurs de fin.

Les éléments intercalés Nous travaillons d'abord sur la position (2) car ce contexte est très fiable en pratique (on applique le principe général : commencer par ce qui est facile et sûr). Nous recherchons donc les mots qui peuvent s'intercaler entre les deux éléments des couples prototypiques. Nous appellerons *LI* cette liste (Liste d'Intercalés). Le résultat est le suivant :

Première itération : `leur se n'en y le lui en ne les nous n'y`
Deuxième itération : `leur se n'en y le lui en ne est les nous n'y`
Troisième itération : `leur se n'en y le lui en ne est les nous n'y`

Les nouveaux débuts Puis nous traitons la position (1) en intégrant la liste *LI* au contexte. Le contexte utilisé est donc composé de deux éléments (les deux éléments des couples), ou de trois : le premier élément du couple, un élément de la liste *LI* et le morphème final du couple (figure 6.5).

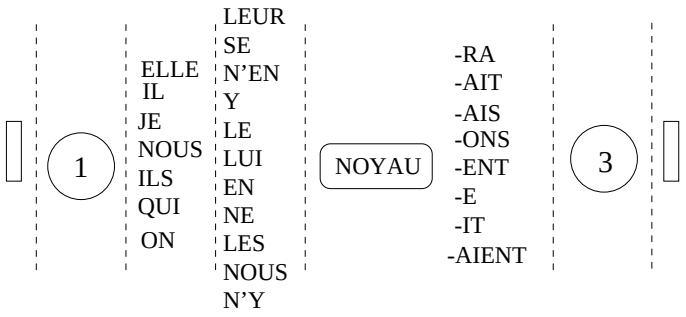


FIG. 6.5 – Contexte utilisé pour rechercher les marqueurs de début apparaissant en position (1).

La liste *LI* augmente donc le nombre de contextes dans lequel un élément peut apparaître. Les éléments trouvés sont :

Première itération : `où ce mais comme car et quand si`
Deuxième itération : `où comme car ce mais tout et quand si`
Troisième itération : `où comme car ce mais tout quand et si`

Les marqueurs de fin Puis nous nous occupons de la position (3) : les marqueurs de fin libres. On intègre bien sûr dans le contexte les nouveaux débuts trouvés (figure 6.6).

L'on peut donc avoir des contextes composés de quatre éléments. Le résultat est le suivant :

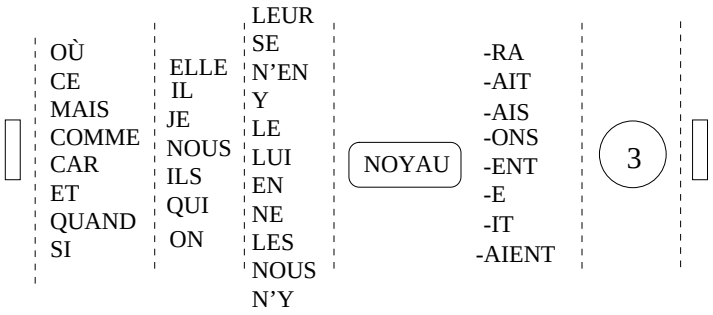


FIG. 6.6 – Contexte utilisé pour rechercher les marqueurs de début apparaissant en position (3) à la premier itération.

Première itération : `pas`
Deuxième itération : `pas`
Troisième itération : `pas`

On pouvait s'attendre à trouver plus d'éléments (*donc*, *plus*, *encore*), mais ces éléments apparaissent rarement en fin d'entre-ponctuations. Pour les catégoriser, il faut attendre le traitement des SR et SSub. Le résultat est très différent pour une langue comme l'allemand, ou les marqueurs de fin de SA sont très nombreux.

Les nouveaux couples On ajoute la liste des fins au contexte (figure 6.7). En utilisant les éléments des positions (1), (2), et (3) pour construire des contextes, nous cherchons à inclure dans la liste des couples, de nouveaux couples apparaissant dans ces contextes. Cela permet d'inclure de nouveaux marqueurs de frontières liés (de nouveaux morphèmes) dans la structure (tableau 6.7). Le couple *qui N-ent* qui n'est pas un couple prototypique du SAD français (Effectif : 253 , Début : 57, Fin : 27) est ainsi reconnu comme structure de SAD. La figure 6.7 illustre la recherche de nouveaux morphèmes liés avec le modèle de couples D N-F. Nous pouvons aussi réaliser une recherche avec les autres modèles de couples : *D-N F*, *N-F F*, et *D D-N* (implémentation non réalisée).

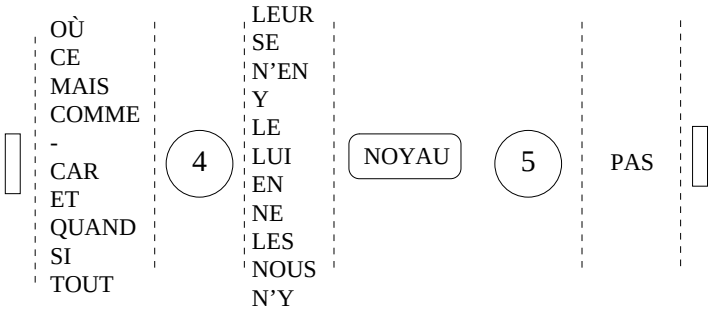


FIG. 6.7 – Le contexte utilisé pour intégrer de nouveaux couples morphologiques dans la structure. La position (4) est occupée par un mot, et la (5) par un morphème lié au noyau.

Première itération :	Deuxième itération :	Troisième itération :
il N-ait	il N-ait	il N-ait
qui N-e	il N-e	il N-e
il N-e	pour N-er	pour N-er
on N-ait	nous N-ions	nous N-ions
pour N-er	elle N-e	elle N-e
nous N-ions	je N-ais	je N-ais
qui N-aient	nous N-ons	nous N-ons
on N-e	en N-ant	en N-ant
qui N-era	ils N-aient	ils N-aient
je N-ais	qui N-it	qui N-it
nous N-ons	qui N-ait	qui N-ait
en N-ant	qui N-e	qui N-e
qui N-ent	on N-ait	on N-ait
ils N-aient	qui N-aient	qui N-aient
elles N-ent	on N-e	on N-e
qui N-it	qui N-era	qui N-era
qui N-ait	qui N-ent	qui N-ent
ils N-ent	elles N-ent	elles N-ent
	ils N-ent	ils N-ent

TAB. 6.7 – Les couples morphologiques de structure [*D N-F*] intégrés à la structure

Le résultat final La figure 6.8 donne le résultat de ces traitements. Nous appelons ce résultat le *schéma contextuel* de la structure SA. À ce stade du traitement, toutes les structures contenues dans le corpus ne sont pas trouvées, mais le résultat couvre une assez grande partie des SA du corpus (section 6.5). Il manque par exemple des éléments comme *lorsque*, *parce que* qui n'apparaissent que très peu avant une structure verbale. L'algorithme ne peut donc les inclure comme début de SAD. Ce résultat suffit pour commencer le traitement des SR.

OÙ	JE	LE		-ONS	
ET	IL	EN		-AIENT	
MAIS	ELLE	S'Y		-ANT	
CAR	ON	NE		-RA	
COMME	QU'ON	LES		-E	
DONT	QUI	NOUS	NOYAU	-ER	PAS
SI	NOUS	N'EN		-ENT	
QUAND	ILS	LEUR		-ERA	
S'IL	ELLES	SE		-IONS	
CE	EN	Y		-AIS	
CEUX	POUR	LUI		-IT	
CELA	EST	S'EN		-AIT	
		N'Y			

FIG. 6.8 – Le schéma contextuel des SA français.

Un élément peut apparaître dans plusieurs positions d'un même schéma (cas fréquent dans le schéma contextuel SAD allemand). Comme le montre le tableau 6.8, on ne trouve pas seulement des structures verbales conjuguées, mais aussi des infinitifs et des participes présents. Ces structures sont catégorisées SAD car elles partagent un même environnement morphologique. Le participe passé n'est pas inclu dans les SA à ce stade du traitement. Les pronoms sont intégrés dans les structures SAD telles que *ce qui ne correspondent pas*, dans lesquelles ils sont catégorisés comme marqueur de début. Le rôle particulier du pronom ne peut se détecter au niveau du syntagme : il faut attendre la génération des couples de syntagmes.

Analysons les éléments catégorisés dans les différentes positions. La position (1) comprend des marqueurs de début correspondant plutôt aux conjonctions, mais nous y trouvons aussi ces pronoms. La position (4) (les mots des couples morphologiques) comprend surtout des pronoms sujets, mais aussi des prépositions (*pour*, *en*). La position (2) correspond essentiellement aux pronoms clitiques. L'appartenance d'un élément à l'une ou l'autre des colonnes n'a pas d'importance. Nous verrons comment ce schéma permet de reconnaître (d'analyser) les syntagmes du corpus dans la section 6.7.

Le contexte utilisé pour la génératin des SA ne permet pas de prendre en compte que les éléments appartenant aux SAD. Le schéma contextuel de l'allemand illustre bien ce propos (annexe C). Si nous observons la composition des éléments apparaissant à la position (3) du schéma (les fins libres), nous y voyons types d'éléments :

- des marqueurs de fin de SAD (*nicht, pronoms*)
- des pronoms sujets dus à la présence d'un adverbe préposé au verbe.

tout ce qui ne relève pas
et qui ne lui pose pas
mais qui ne devrait pas
et comme il n'y a pas
ce qui ne correspond pas
quand elle ne se limite pas
si on ne se défend pas
et il n'y a pas
mais on ne savait pas
c'est pour le laisser
pour y célébrer
tout en affirmant
tout en se passionnant

TAB. 6.8 – Exemple de SAD français.

- des SAF fréquents comme *gehen, sein, thun*
- des adverbes (*nieder,*)
- des marqueurs de fin de proposition (*ab, ein, auf*)

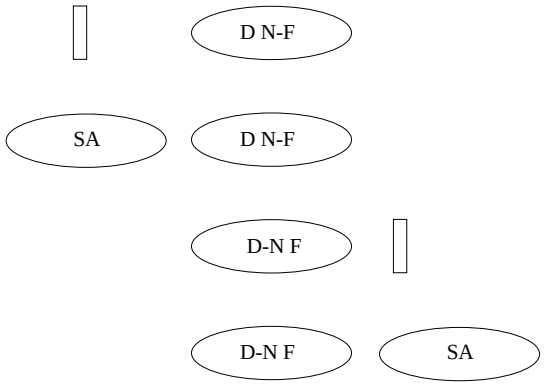
Nous voyons que la classe des fins de SAD et des fins de proposition a fusionné, résultat attendu car le contexte ne permet pas de distinguer les deux catégories. Le même résultat est obtenu avec les marqueurs de début de SAD et de proposition (nous ne pouvons distinguer les débuts de SA des débuts de proposition).

D'une manière générale, la génération des SA permet une identification de toutes les catégories intervenant au niveau propositionnel. La distinction entre ces différentes catégories ne pourra se faire qu'en utilisant des contextes comprenant des SR et SSub, afin d'identifier les éléments du niveau propositionnel des éléments du niveau des SA.

6.4.5 La génération des structures SR

La technique utilisée pour la catégorisation de SR est similaire à celle décrite précédemment. La seule différence est l'utilisation des SA trouvés précédemment : ils peuvent servir, de la même manière que les ponctuations de délimiteurs de SR. Rappelons que les SA trouvés sont marqués dans le corpus et ainsi identifiables par les traitements suivants. Nous commençons par traiter les structures morphémiques, puis les couples morphologiques, enfin les classes lexicales. La sélection de ces structures est différente (celle des SA se basait sur la position absolue de certains couples) Pour trouver ces structures, nous sélectionnons tous les éléments qui apparaissent dans ces contextes décrits par la figure 6.9.

Prenons l'exemple de la génération des couples morphologiques en français. Cette opération de sélection nous donne une liste de couples morphologiques prototypiques des SR (tableau 6.10). Cette construction sélectionne les couples qui ont une structure de couples morphologiques. Nous voyons que nous n'obtenons pas seulement des structures nominales, mais aussi verbales (*été N-é*, *à*



TAB. 6.9 – Les SA sont intégrés au contexte pour la découverte des SR. Ils servent de délimiteurs de SR au même titre que les ponctuations.

N-ir).

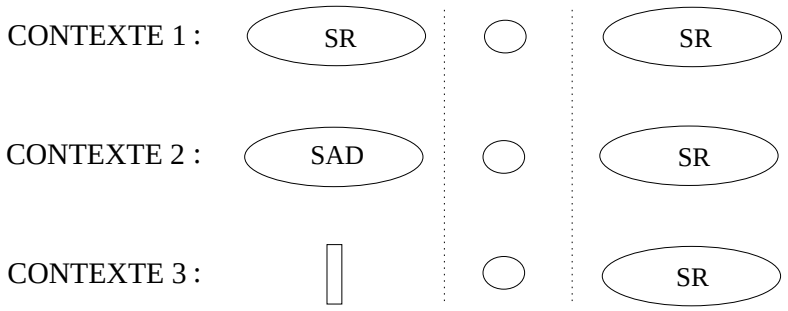
des N-es
cette N-e
à N-ir
les N-es
leurs N-es
de N-ir
le N-at
la N-ue
des N-tions
été N-é

TAB. 6.10 – Quelques couples morphologiques considérés comme SR.

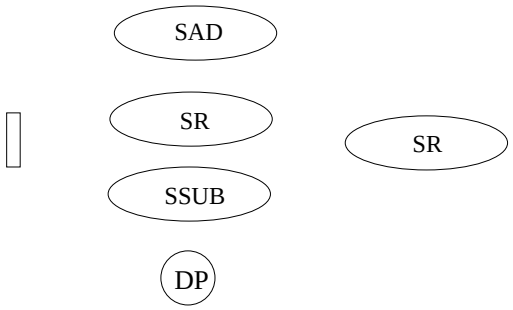
Une fois ces couples prototypiques de SR obtenus, nous appliquons un algorithme de génération de SR similaire à celui des SA, à la différence près que, là où les SA utilisent une ponctuation pour délimiter le syntagme, nous doublons le contexte en la remplaçant par un SA. Il est très important de noter que les contextes avec ponctuation et avec SA ne sont pas complémentaires mais tous les deux nécessaires.

Si nous utilisons seulement le contexte constitué des ponctuations comme délimiteur de SR, la catégorisation produit un résultat assez médiocre. En effet le contexte de la figure 6.12 n'est pas assez contraignant (des éléments comme des verbes ou adjectifs selon les langues peuvent apparaître assez souvent dans un tel contexte).

Il faut donc ajouter d'autres contraintes à ce contexte. Ceci est réalisé en ajoutant différents contextes (figure 6.11). Il est donc indispensable de combiner les trois contextes, le troisième servant à sélectionner des candidats, le premier et le second à éliminer les mauvais : seuls les éléments du troisième contexte apparaissant au moins une fois dans les deux premiers contextes sont retenus. Le



TAB. 6.11 – Trois sortes de délimiteurs sont utilisés pour la recherche des débuts de SR : la ponctuation, les SA, et les SR.



TAB. 6.12 – Les éléments pouvant théoriquement s'intercaler entre une ponctuation et un SR : on peut trouver tous les types de syntagmes, ainsi que des débuts de propositions (DP).

premier contexte n'est utilisé qu'une fois certaines structures de SR découvertes. La diversité des délimiteurs assure (la plupart du temps) que les éléments ainsi catégorisés sont bien des marqueurs de frontière de SR. En théorie, un élément apparaissant dans ces trois contextes peut ne pas être un début de SR (un élément polycatégoriel par exemple, qui serait SA et SSUB), mais en pratique, nous obtenons bien un début de SR.

Au niveau du SR, la recherche des éléments intercalés (position 2) n'est pas réalisée, ces éléments correspondant le plus souvent à un SSub en composition interne.

Cet algorithme ne produit pas que des SR théoriques. Comme nous l'avons dit, la découverte des SA n'est pas totale, certaines structures ne sont pas incorporées. Ainsi, la structure française *de N-ir* (infinitif deuxième groupe) est considérée comme SR et non comme SA, alors que la structure *de N-er* (infinitif premier groupe) est un SA. Ceci n'est pas dû à une différence entre les deux groupes, mais à un silence de la catégorisation des SA. De tels couples peuvent alors apparaître dans cette phase de catégorisation et sont donc considérés comme SR. La catégorisation des syntagmes n'est pas absolument fiable, mais cela n'empêche pas une bonne catégorisation en marqueur de début ou de fin, et donc une bonne construction des syntagmes. Nous reviendrons sur le problème de la catégorisation des syntagmes dans la section 6.6. La figure 6.13 montre le schéma contextuel obtenu pour le français. Certains éléments (*ainsi, aussi*) sont catégorisés début de SR alors qu'ils correspondent à des fins de SA. Ces erreurs sont dues au contexte 2 et au fait que ces éléments n'ont pas été catégorisés précédemment comme fin de SA.

6.4.6 La génération des structures SSub

Une fois une structure (SA ou SR) générée, nous pouvons nous intéresser à ces Syntagmes Subordonnés (SSub). Nous cherchons des syntagmes dont les éléments pris en considération doivent comporter un noyau lexical. Nous retrouvons en fait les trois types de structures de la section 6.4.2 :

- D-N ou N-F
- D-N F ou D N-F
- Classes lexicales

Le régissant peut être un SA ou un SR. Une fois certains SSub trouvés, la recherche des SSub ayant un SSub comme régissant peut se faire.

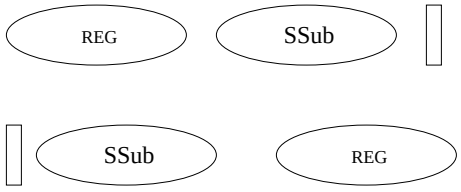


FIG. 6.9 – Contextes utilisés pour la génération des SSub. La structure régissante (Reg) peut être soit un SA soit un SR.

			-ONS
			-CATIONS
			-S
	À		-T
AVEC	UN		-AT
TOUTES	LES		-RE
QUE	NOUS		-ER
CONTRE	ÉTÉ		-RE
SUR	UNE		-ES
DANS	DE		-IONS
FAIT	SES		-EMENT
DE	LEURS		-TION
PAR	LA		-ENTS
PLUS	SA		-UE
EN	AUX		-ENCE
ET	SON	NOYAU	-ATION
EU	LE		-IR
AUSSI	D'UNE		-ÉS
DEVANT	EN		-ON
PEU	SA		-EURS
COMME	SANS		-IER
POUR	CETTE		-ÉES
L'UN	AU		-E
LOIN	SE		-ITÉ
ENCORE	LEUR		-ION
AINSI	DU		-ENT
À	DES		-É
			-EMENTS
			-TIONS
			-IE
			-ATIONS

TAB. 6.13 – Schéma contextuel des SR français.

mais aussi le classement
lors de la con-férence
et sur le territoire
tant que des élections
de ne pas en dire
qui a été frappé
pendant plus d'un siècle

TAB. 6.14 – Exemple de SR français. On trouve aussi bien des groupes nominaux que verbaux. Nous retrouvons toutes les structures non étiquetée SA, de structure [*D N-F*].

Le contexte mis au point (figure 6.9) n'est pas très contraignant et différents éléments peuvent apparaître. Le problème se pose en particulier lorsque nous recherchons des SSub de SR. Nous nous sommes aperçu durant notre travail que les SSub d'une langue étaient souvent de nature différente de leur régissant. Si le régissant est de structure [D-N F] ou [D N-F], alors le SSub est de nature [D-N], [N-F] ou classe lexicale. Nous allons donc restreindre la recherche des syntagmes à des structures différentes de celles de leur régissant. Ainsi seuls les syntagmes de modèle [D-N], [N-F] ou classe lexicale sont pris en compte dans la recherche des SSub français (les SA et SR correspondent au modèle [D N-F]). Sans cette contrainte, il devient impossible de différencier les SSub de SA des actants (de nature SR). Nous ne considérons donc pas comme SSub un SR dépendant d'un autre SR ou d'un SA.

Les SSub qui partagent une même nature morphologique que leur régissant sont donc très difficiles à différencier de leur régissant. En pratique la génération de ces deux structures est réalisée pendant la génération du régissant, et notre méthode ne produit donc pas de SSub. En cas de partage d'un même modèle morphologique, la distinction entre régissant et SSub peut se réaliser si les SSub possèdent une contrainte positionnelle (section 4.8.2) comme la structure génitive allemande (de nature [*D N-F*] comme leur régissant, mais avec une morphologie légèrement différente) ou les adjectifs/adverbes (de nature lexicale ainsi que leur régissant) du vietnamien.

Une autre solution serait de mettre au point un contexte plus contraignant, mais nous n'avons pas réussi (l'ajout d'autres structures comme dans le cas du SR (figure 6.13) ne donne pas de meilleur résultat). Nous voyons que la génération des SSub mélange contexte distributionnel et critère morphologique. Le tableau 6.15 donne le résultat de la génération des SSub de SA en français. Si aucune contrainte n'était imposée sur le modèle des SSub, la liste des SSub intègre les SR de la langue.

SA	N-ement
	N-er
	N-é
	N-és
	N-ées
	N-ir

TAB. 6.15 – Les SSub de SA français. Le modèle morphologique pris en compte est [N-F]. Le résultat correspond aux structures adverbiales, mais aussi capture les séquences verbales. Aucun SSub n'est trouvé pour le contexte gauche du SA.

Dans les langues morphologiques, les structures d'accord (section 3.3.2) peuvent construire des structures comprenant le régissant et son subordonné. Cette opération a été implémentée (tableau 6.16). Le résultat est une structure comprenant deux syntagmes, dont il est parfois difficile de distinguer le régissant du subordonné pour certaines langues.

Structure	couple d'accord
ces dernières années	*ces N-es N-es*
les par-ten-aires con-ventionn-els	*les N-s N-es*
des banques centrales	*des N-es N-es*
en quelque sorte	*en N-e N-e*
la caisse nationale	*la N-e N-e*
de banques centrales	*de N-s N-es*
la semaine dernière	*la N-e N-e*
la politique monétaire	*la N-e N-e*
les pouvoirs publics	*les N-s N-s*
des affaires étrang-ères	*des N-s N-s*

TAB. 6.16 – Structures de deux syntagmes générées grâce aux structures d'accord.

6.4.7 Le résultat de la catégorisation

Une fois cette catégorisation effectuée, quels résultats obtient-on ? Le premier est la construction de la *table des catégories* (tableau 6.17). Cette table recense la liste des catégories possibles pour un mot. Il suffit de parcourir les schémas contextuels, et pour chaque catégorie du schéma (marqueur de début, fin) recenser les éléments apparaissant dans celles-ci. L'évaluation des tables est donnée à la section 6.5.

Mot	N	FSAD	FSAF	FSR	DSAD	DSAF	DSR	DSSub
elle					√			
je					√			
comme					√		√	
d'où							√	
car					√		√	
leur					√		√	√

TAB. 6.17 – La table de catégorisation. Quelques éléments français.

En l'état actuel, les catégories traitées sont :
- Élément comprenant un noyau syntagmatique (N)
- Début de SAD ou de Proposition (DSAD)
- Fin de SAD ou de proposition (FSAD)
- Début de SAF ou de Proposition (DSAF)
- Fin de SAF ou de Proposition (FSAF)
- Début de SR (DSR)
- Fin de SR (FSR)
- Début de SSub (DSSub)

La distinction entre début/fin de proposition et de SA ne peut se faire qu'en intégrant la structure propositionnelle. Dans le cas où aucun SR ne peut s'intercaler entre un DP et un SA (langues VSO par exemple), la discrimination est

très délicate. La situation est la même dans le cas des langues SOV (turc), où la discrimination entre les Fins de Proposition (FP) et les Fins de SAF. D'ailleurs, aucun élément turc n'est considéré comme FP (en se référant à une grammaire turque). La figure 6.10 montre les contextes utilisés pour discriminer les débuts de proposition des débuts de SA (par exemple une conjonction d'un pronom sujet en français).

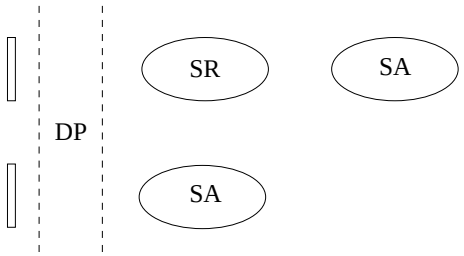


FIG. 6.10 – Discrimination entre Début de Proposition (DP) et Début de SAD (DSAD). La connaissance des SA et des SR est nécessaire.

Le deuxième résultat est la segmentation partielle du corpus en syntagmes. À la sortie de la chaîne de traitement, le corpus a été segmenté en SA(D|F), et SR. Cette segmentation n'est pas la segmentation finale, mais celle produite lors de la catégorisation des éléments. Voici le résultat sur la première entre-ponctuations de *français01* :

```
. quatre cents spéci-alistes SR-se sont re-trouv-és-SR SR-le mois
dernier-SR à lyon
```

6.4.8 La segmentation du corpus en syntagmes

Le résultat de la catégorisation nous fournit donc deux choses : la table de catégorisation et les schémas contextuels. À partir de ces deux ressources, nous allons segmenter les corpus en syntagmes. Nous sommes ici dans une phase non plus de découverte mais d'analyse. Notre analyseur est très rudimentaire. Il est basé sur des expressions régulières, et n'utilise que ces contextes immédiats. Les résultats sont assez bons dans des langues comme l'anglais ou le français. Ils se dégradent avec une langue comme l'allemand, où les éléments propositionnels sont beaucoup plus délicats à gérer. Dans cette langue, la prise en compte, dans le processus d'analyse, du niveau propositionnel est indispensable.

Voyons comment les syntagmes sont analysés. Certains éléments sont faciles à gérer : les débuts et fins sûrs sont collés au mot suivant ou précédent. Un élément est considéré comme début sûr ou fin sûre si toutes ces catégories sont des catégories de début ou de fin. *Comme la structure propositionnelle n'a pas été intégrée explicitement au traitement, il n'est pour l'instant pas fait de différence entre marqueur des structures syntagmatique et propositionnelle.* On colle donc les débuts et fins de syntagmes ainsi que les débuts et fins de proposition à l'élément suivant ou précédent.

Pour les éléments polycatégoriels, il est nécessaire de recourir aux schémas

contextuels. Les schémas contextuels sont transformés en expression régulière. Nous considérons trois éléments dans le schéma : les marqueurs de début (MD), les marqueurs de fin (MF), et les couples morphologiques (CM). Plusieurs expressions régulières sont générées à partir de ces données :

- (MD)* CM (MF)*
- (MD)+ X (MF)+
- (MD){2,} X
- X (MF){2,}

Le premier modèle correspond à une séquence comprenant un couple morphologique, plus un nombre quelconque de marqueurs de début et de fins. Le deuxième modèle correspond à une séquence d'un mot compris entre au moins un marqueur de début et au moins un marqueur de fin. Le troisième correspond à un mot précédé d'au moins deux marqueurs de début. Le dernier modèle est le symétrique du précédent. On ajoute aussi les deux contextes suivants :

- Ponctuation D
- F Ponctuation

Si un élément D peut être un marqueur de début et qu'il se trouve après une ponctuation, il est catégorisé comme début et est collé au mot suivant. Idem pour les marqueurs de fin. *Les éléments qui ne se rencontrent pas dans ces contextes ne sont pas traités.* Nous ordonnons le traitement en commençant (toujours) par les structures SA, puis les structures SR. Malgré sa simplicité, ce processus d'analyse assure généralement un bon contexte pour les éléments polycatégoriels. Toutes les séquences qui correspondent à ces modèles sont regroupées en syntagmes. Ces expressions régulières ne tiennent pas compte des règles de structuration interne à un type de syntagme (par exemple tous les débuts sont considérés de manière similaire), mais elles produisent de très bons résultats. Nous n'avons pas à imposer nous même l'ordre dans les séquences de marqueurs de frontière : le corpus le fait pour nous. Voici un exemple de segmentation sur le corpus *français01* (les syntagmes sont mis entre crochets) :

. quatre cents spécialistes [se sont retrouvés] [le mois] dernier [à lyon]
, invités conjointement [par la direction] [de l'architecture] [et de l'urbanisme] [du ministère] [de l'équipement]
, [du logement]
, [des transports] [et de la mer] [et par la fédération] nationale [des agences] d'urbanisme (fnau) [pour réfléchir] [sur l'avenir] [de la planification] urbaine

Ce corpus segmenté sert d'entrée au processus de catégorisation des syntagmes (section 6.6). La mise en syntagme est facile dans les langues où les éléments catégorisés comme début (fin) ne sont pas catégorisés comme fin (début) d'une autre structure. Il suffit alors de coller systèmatiquement ces éléments au suivant ou au précédent. C'est généralement le cas en français, anglais, turc, swahili. Dans une langue comme l'allemand ou des éléments peuvent assez souvent être marqueurs de début et de fin, il est nécessaire d'utiliser un analyseur plus performant, sinon le nombre d'occurrences d'éléments non traités est assez important. Nous voyons ici la différence entre un travail de découverte et d'analyse. *Nos algorithmes (de découverte) nous ont fourni des renseignements sur la catégorie d'un élément, mais savoir reconnaître chaque occurrence de cet*

élément est un problème d'analyse, problème non central à ce travail.

La mise au point de cet analyseur permet une certaine généralisation des structures. Cet analyseur peut reconnaître des syntagmes qui n'apparaissent pas dans le corpus d'apprentissage. Ainsi la séquence française :

il le leur N-a

ne se trouve pas dans notre corpus *français01* sur lequel les algorithmes ont opérés la catégorisation, mais appartient au corpus *français02* (*il le leur livra*). Lorsque nous segmentons ce corpus, cette séquence est bien reconnue comme étant un syntagme de la langue.

Notre objectif principal n'étant pas une mise au point d'une procédure d'analyse, cet analyseur n'a pas été amélioré, malgré sa rusticité. Les adresses Web des corpus segmentés sont données en annexe C.

6.5 Évaluation des résultats

There are lies, damn lies and statictics. (Mark Twain)

Voici venir le temps des évaluations. La comparaison avec d'autres travaux est délicate, puisque aucun travail similaire n'a été réalisé jusqu'à présent (travail sur des données brutes, et surtout multilingue). Nous allons donc évaluer notre travail selon *nos propres critères* et notre propre jugement en essayant de les expliciter. Il existe plusieurs manières d'évaluer ce travail. Nous en proposons quatre :

- la table de catégorisation
- la couverture de la catégorisation
- la qualité des syntagmes obtenues
- la segmentation du corpus

Une évaluation intéressante ne pourra se faire que lorsque toutes les structures et les catégories seront pris en compte. Les tables de catégorisation, les schémas contextuels, ainsi que les corpus segmentés des différentes langues sont donnés en annexe C.

La table de catégorisation L'évaluation des tables de catégorisation utilise deux critères : le silence et l'erreur. Le silence correspond au nombre de mots de la table n'ayant pas été totalement catégorisés par notre algorithme. L'erreur correspond au nombre de mots qui ont été incorrectement catégorisés. Nos algorithmes actuels ne nous permettant pas de distinguer entre marqueur de frontière de syntagme et de proposition, nous avons fusionné les différentes catégories de marqueur de début (de SA, SR, SSub, et P) et de fin. Nous pouvons donc considérer que nous avons trois catégories fiables : noyau, marqueur de début et marqueur de fin. Elles suffisent à réaliser une segmentation en syntagmes (en prenant comme convention que les marqueurs de frontière de proposition appartiennent au syntagme voisin). Ce sont donc ces trois catégories que nous avons retenues pour évaluer la catégorisation. Une évaluation plus fine sera possible lorsque toutes les catégories pourront être traitées (cela nécessite la catégorisation des syntagmes qui n'a pas été réalisée).

Langue	nb mots catégorisés	Erreur	Silence
français	113	2 (2%)	25 (22%)
allemand	157	5 (3%)	16 (10%)
anglais	78	6 (8%)	24 (30%)

TAB. 6.18 – Évaluation des tableaux de catégorisation.

Le tableau 6.18 présente les taux d'erreur et de silence pour trois langues : français, allemand et anglais. Détaillons les résultats du français. Les silences correspondent majoritairement (10) à des auxiliaires ou modaux (avoir, être, devoir) catégorisés comme marqueurs de début, et non comme élément lexical. D'une manière générale, tous les déterminants et toutes les prépositions (de la liste) sont identifiés comme marqueurs de début. Le fait de prendre en considération ou non les catégories très rares fait passer le silence de 25 à 33. Les catégories suivantes représentent environ 1 à 2% des occurrences des mots dans le corpus *français03* :

- contre (SAD, SR)
- son (SR)
- entre (SAD)
- une (pour "TF1", SR)
- car (SR)

Voyons les résultats obtenus sur l'allemand. Un certain nombre d'erreurs (7) correspond à des verbes à l'infinitif qui ont été considérés comme fin de SAD alors qu'ils correspondent à des SAF (*thun, gehen, hören*). Les silences sont dus à quelques prépositions non catégorisées comme fin de proposition.

Les silence sur l'anglais proviennent principalement (9 cas) d'auxiliaires (be, have, would) catégorisés comme marqueurs de début, ainsi que des éléments pouvant être préposition (début de SR) et particule verbale (fin de SAD) qui ne sont pas catégorisés FSAD (nous avons systématiquement considérés chaque préposition comme pouvant être une fin de SAD, ce qui correspond à 10 silences). Les erreurs proviennent des éléments *go*, *came*, *take*, *aunt*, *huck* et *tom* catégorisés comme marqueurs de début.

Remarques générales Un mot est toujours assigné à sa catégorie principale. Si un élément correspond à deux catégories assez fréquentes (par exemple le mot allemand *meine* qui peut être SA ou DSR), alors la catégorisation est aussi correcte. Les silences proviennent essentiellement de verbes irréguliés catégorisés comme marqueur de frontière et non noyau (on pourrait considérer ces silences comme des erreurs, mais cela ne nous semble pas justifié). Les prénoms sont souvent catégorisés comme marqueur de début (anglais *Tom*, allemand *Halef*, espagnol *Moises*).

La couverture de la catégorisation Une estimation intéressante est celle de la couverture de la catégorisation. Elle correspond au nombre de mots du corpus qui ont été catégorisés. Certains éléments caractéristiques ne sont pas

catégorisés (all. *empor*, *her*). La centaine de mots catégorisés correspondn à 40% des mots du corpus. On remarque que nous obtenons une même estmation avec les trois langues. Nous ne savons si cela est au hasard ou bien si ce fait correspond à une propriété des trois langues.

Langue	nb mots catégorisés	taille du corpus (mots)	couverture (%)
français	113	263627	43%
allemand	157	152036	43%
anglais	78	115187	43%

TAB. 6.19 – Couverture de la catégorisation des mots grammaticaux. Les mots catégorisés représentent plus de 40% du corpus.

Une autre estimation intéressante est celle de la couverture de la mise en syntagmes. Pour cela nous comptons le nombre de syntagmes obtenus (tableau 6.20). Les syntagmes singleton correspondent à des mots sans marqueur de début ou de fin. Nous voyons donc que plus de 60% des syntagmes sont composés de plus d'un mot. On notera une fois encore la similitude des résultats entre ces trois langues. Nous ne pouvons dire pour l'instant si cette similitude est un hasard ou se retrouve aussi dans les autres langues. Des estimations sérieuses devraient être faites mais sur des corpus où la construction des syntagmes est

Langue	nb mots	nb syntagmes segmentés	nb syntagmes singleton
français	263627	147866	47459 (32%)
allemand	152036	85237	27319 (32%)
anglais	115187	65921	23188 (35%)

TAB. 6.20 – Couverture de la mise en syntagmes.

La qualité des séquences L'estimation la plus parlante est celle qui concerne la qualité des syntagmes contruits. Si nous avons déjà dit que cette phrase nécessite la mise en place d'un analyseur plus perfectionné, les résultats obtenus montrent que les syntagmes sont assez faciles à construire. Nous avons évalué la liste des SAD générés en français, anglais et allemand, c'est à dire vérifié si les séquences correspond bien à notre définition du syntagme. Ainsi les séquences suivantes ne sont pas validées comme étant des SAD français (certaines correspondent cependant à des syntagmes bien formés) :

- *pour qui elle*
- *pour ne pas*
- *est le premier pas*
- *en carburant*
- *tout petits pas*
- *pour le brigadier*

Les séquences qui ne correspondent pas à un syntagme bien formé (*pour qui elle*) ne sont pas dues à une mauvaise catégorisation, mais à une erreur générée par

notre analyseur. Nous avons nous même réalisé la validation des langues allemande et anglaise. Un taux d'erreur peut varier fortement suivant la convention utilisée. Nous trouvons un taux d'erreur de 18% en anglais. Mais près de 50% de ces erreurs correspond à une catégorisation du mot *tom*[75] : 132 séquences sont considérés comme incorrectes (ce qui fait passer le nombre d'erreurs de 151 à 283). Si nous considérons le mot *tom* comme un début de SAD (de manière similaire à un pronom), alors le taux d'erreur passe à 10%. Les autres erreurs sont majoritairement dues au mot *to* (134 erreurs), classé comme FSAD, et qui correspond en fait à un début de SR. Ceci est dû à un mauvais fonctionnement de notre "analyseur" qui privilégie les SA, puisque notre catégorisation de l'élément est correcte (fin de SAD et début de SR). Si nous éliminons ces deux erreurs, nous obtenons un taux d'erreur de 1%. Les principales erreurs allemandes proviennent des éléments qui peuvent être fin de proposition ou de SAD mais aussi début de SR, comme *ein* (249) et *mit* (236). Le taux d'erreur allemand est plus levé à cause des séquences correspondant aux marqueurs de début de proposition du type *aber als ich*. Il retombe à 967 (13%) si nous considérons ces séquences comme étant des SAD. Encore un exemple de l'importance de la prise en compte de la structure propositionnelle. La plupart des erreurs proviennent donc d'éléments qui apparaissent à la fin d'une structure SA ou Proposition mais aussi au début de la structure SR. Nous voyons donc qu'un petit nombre d'éléments peut parfois générer un nombre important d'erreurs. Le faible taux d'erreur en français est dû au fait qu'il n'existe pas d'élément (sauf *pas*) catégorisé comme marqueur de début d'une structure et marqueur de fin d'une autre.

Langue	Effectif	SAD correct
français	2837	97%
allemand	7019	81%
anglais	1502	81%

TAB. 6.21 – Évaluation des SAD générés.

Langue	Erreur
français	3%
allemand	18%
anglais	18%

TAB. 6.22 – Évaluation des SR générés (faite sur les 1000 premiers Sr du corpus).

La segmentation en syntagmes On nous a souvent demandé d'évaluer notre segmentation en syntagmes en les comparant à d'autres résultats. Deux objections à cela. D'une part il n'existe pas de corpus segmenté en syntagmes (en

[75]Le corpus contient la nouvelle de Marc Twain : *les aventures de Tom Sawyer*.

prenant notre définition du syntagme comme référence). D'autre part, n'ayant pas réalisé un analyseur (il est parfois difficile d'appeler *analyseur* notre "segmenteur syntagmique"), la comparaison serait sans intérêt (et sans doute peu flatteuse pour nous). La qualité de la segmentation en syntagmes dépend de la langue. La segmentation est assez facile pour une langue comme le français ou le swahili où les marqueurs de début ne se trouvent pas aussi comme marqueurs de fin. Elle est plus délicate dans une langue comme l'allemand ou de nombreux éléments apparaissent aussi bien comme marqueurs de début et de fin. Pour ces langues, la mise en place d'un analyseur plus perfectionné est nécessaire pour obtenir de bons résultats.

6.6 La catégorisation des syntagmes

Suivant le principe développé à la section 4.4.2, la catégorisation d'un élément ne peut se faire qu'en l'intégrant dans une structure supérieure. La catégorisation des syntagmes doit donc se faire en travaillant au niveau de la structure propositionelle ou des couples de syntagmes. Notre algorithme de catégorisation nous propose déjà une catégorisation en SA et SR, mais comme nous l'avons noté, cette catégorisation n'est pas fiable. Elle se base sur des critères morphologiques qui ne possèdent pas assez de contraintes pour permettre une catégorisation correcte. En particulier, si la catégorisation des SA génère bien des SA, la génération des SR "ramasse" le reste des structures non catégorisées comme SA. Certains SA oubliés deviennent donc des SR. Un fois le corpus segmenté en syntagmes, il est nécessaire de reprendre leur catégorisation en utilisant les couples de syntagmes. Aucune implémentation n'a été réalisée. Une idée d'algorithme serait d'utiliser les SSub prototypiques des structures pour les catégoriser.

6.7 La catégorisation interne au syntagme

Une fois le corpus segmenté en syntagmes, nous pouvons étudier la structure de ceux-ci. L'étude n'est pas effectuée sur les schémas contextuels, mais sur les séquences construites lors de la segmentation en syntagmes, particulier les syntagmes comprenant le plus d'éléments. L'étude de la structure des syntagmes consiste à étudier les positions relatives des éléments dans un syntagme. Pour l'instant les éléments sont catégorisés en deux classes : marqueur de début et marqueur de fin. Mais ces marqueurs possèdent généralement des contraintes quant à leur positionnement dans ce syntagme. Cette étude permet donc d'ordonner les éléments dans un syntagme. L'étude des syntagmes du tableau 6.23 montre que dans une séquence de marqueurs de fin, l'élément *nicht* se positionne toujours en dernière position. Tous les marqueurs de frontière ne sont donc pas équivalents : la description d'un syntagme en terme de début, noyau, fin est donc insuffisante. Cette propriété est intrinsèque au syntagme, c'est à dire qu'elle ne dépend pas de la structure dans laquelle le syntagme s'insère. Ces propriétés sont donc étudiables au niveau syntagmatique. Mais ce n'est pas le cas de toutes les propriétés du syntagme. Il est parfois nécessaire d'intégrer le syntagme dans sa structure supérieure pour comprendre certaines règles de

construction. Ainsi, l'étude de la position du pronom sujet en allemand ne peut se faire qu'en intégrant le syntagme dans sa structure propositionnelle. L'étude interne du syntagme n'a conduit à aucune réalisation informatique. Elle est bien sûr indispensable pour obtenir une bonne analyse syntagmatique du corpus.

Début	Noyau + affixes	Fin
du	vermuthest	es auch *nicht*
aber ich	werde	auch *nicht*
ich	habe	dich *nicht*
warum	sollte	ich dich *nicht*
wir	werden	dich *nicht*

TAB. 6.23 – Dans la structure SAD allemande, le marqueur de fin *nicht* se trouve toujours en dernière position des séquences de marqueurs de fin.

6.8 Ce qu'il reste à faire

Beaucoup de choses bien sûr. Voici un début de liste :

La prise en compte des différentes structures Nous avons vu dans la section 6.4.2 que les algorithmes de catégorisation utilisaient des structures prototypiques pour amorcer cette catégorisation. Ces structures pouvaient prendre trois formes :

- structure morphémique
- couple morphologique
- classe lexicale

Ces trois types nécessitent trois programmations différentes, à moins de trouver un formalisme qui homogénéise le traitement (il est sans doute possible d'unifier les deux premiers types). Le tableau 6.24 montre l'état actuel de l'implémentation. Le traitement de la structure morphémique des SSubF (SSub marquant une fin de SR ou SA) permettrait par exemple l'intégration de la structure adverbiale et adjectivale en français (de structure *N-F*).

	SAD	SAF	SR	SSubD	SSubF
structure morphémique		√			
couple morphologique	√	√	√	√	
classe lexicale					

TAB. 6.24 – État actuel de la couverture des structures prises en compte dans la réalisation informatique.

La découverte des structures composées Nous avons mis au point des algorithmes concernant le niveau syntagmatique. Ce niveau est suffisant pour catégoriser tous les éléments syntagmatiques correctement (par exemple, aucun

marqueur de fin de SR n'est en français). Mais il existe quand même des erreurs. Prenons le cas des SR français. Parmi les marqueurs de début, nous trouvons *notamment*. Cet élément possède toutes les caractéristiques d'un marqueur de début de SR. Seule la connaissance des structures SSub du SA français permettrait de le catégoriser correctement.

Cette découverte des structures composées permettrait surtout la prise en compte du niveau propositionnel. *L'intégration de la structure propositionnelle est une étape incontournable.* Toute tentative d'amélioration des algorithmes qui n'intégrerait pas la proposition est sans intérêt. L'amélioration serait quantitative mais pas qualitative. Certains erreurs ne peuvent être évitées si les contextes restent au niveau du syntagme. Par exemple, lors de la construction des SAD allemands, certains éléments sont catégorisés comme marqueurs de fin de SAD alors qu'il sont en réalité des SAF (comme les verbes très fréquents : *thun*, *habe*). Seule la connaissance de la structure du niveau propositionnel peut éviter cette erreur. Et seul ce niveau peut nous permettre un traitement des actants. Cette construction du niveau propositionnel peut sans doute se réaliser à partir des structures de SA : SAD et SAF, et en essayant d'intégrer des SR.

Vers un vrai système d'apprentissage ? Les programmes n'ont été écrits que dans un but de validation des concepts. Le propos n'était pas de réaliser un système performant. Dans cette optique, l'ensemble de la chaîne de traitement est à revoir. Comme nous venons de le dire, le processus doit se centrer sur la structure propositionnelle. L'intégration des ressources lexicales doit aussi être effectuée. La question qui se pose est de savoir quelles sont les autres connaissances que nous n'avons pas recensées nécessaires à la construction d'un tel système ? Des attendus sur des éléments comme les pronoms (un universel des langue selon [Greenberg, 1963]), éléments difficiles à manipuler, mais pourtant assez faciles à découvrir, sont-ils nécessaires ? Mais ces résultats, aussi partiels soient-ils, montrent bien que la langue possède assez d'indices formels pour permettre un amorçage d'un système d'apprentissage. Reste la question de savoir si la mise au point d'un tel système est nécessaire ou utile. Ce type de travail a essentiellement un intérêt théorique (c'est, en tout cas, notre point de vue). Un travail d'analyse peut profiter de ces résultats non pas en utilisant les sorties des algorithmes, mais en intégrant les concepts linguistiques que ce travail a mis en évidence.

Cinquième partie

Conclusion

Chapitre 7

Mais, à quoi ça sert ?

Sommaire

Cette question nous a trop souvent été posée lors d'exposés oraux de ce travail pour ne pas essayer d'y répondre. Cette réponse servira de conclusion à ce travail. Mais avant d'y répondre, résumons le travail effectué.

7.1 Retour sur le travail accompli

Les résultats Nous avons commencé notre introduction générale par une question :

> Que peut-on apprendre sur la structure d'une langue à partir d'un texte écrit dans cette langue, et ceci sans connaissance particulière sur celle-ci et avec l'aide (disons l'utilisation) d'un ordinateur ?

La réponse semble être : beaucoup de choses mais pas tout. Il nous aurait été difficile de seulement esquisser une réponse il y a trois ans. Revenons d'abord sur nos données. Travailler avec des *textes bruts*, sans annotation et sans lexique a très fortement orienté notre démarche. L'idée de partir de ces données nous a été donnée par Zellig Harris, qui a eu une grande influence la première année. Partir de telles données nous a isolé (méthodologiquement) du reste de la communauté travaillant dans le domaine de l'inférence grammaticale[76]. Les méthodes utilisant soit des corpus annotés, soit des lexiques, soit des corpus construits artificiellement se sont révélées totalement inadéquates sur nos données. Une autre propriété nous différencie légèrement des autres travaux : notre approche multilingue. Nous reviendrons sur ce fait, très important méthodologiquement, dans la section 7.4. Nous travaillons toujours en ayant en tête la

[76] Cette communauté nous semble être celle qui se rapproche le plus de notre travail.

question : en quoi tel fait peut être utile en vue d'un traitement automatique des langues ? Se placer dans le cadre d'un traitement automatique nous fait préférer un type d'information particulier : la forme. Cette contrainte limite nécessairement notre champ d'investigation, et il peut paraître à certains que ce sont de mauvaises conditions de travail. Toutefois, ce sont celles que nous avons choisies.

Voyons maintenant quels ont été les résultats produits. Nous pouvons les classer en quatre points :

- Segmentation des mots en morphèmes.
- Identification des marques structurelles de la langue
- Construction (validation) d'une hiérarchie structurelle des langues à partir de ces marques.
- Algorithmes de génération de la structure.

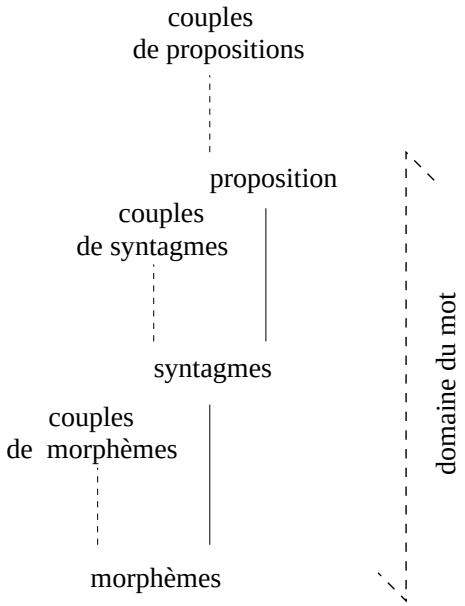
La procédure de segmentation des mots est directement inspirée de l'algorithme de Harris. Elle permet l'identification de l'unité de base de la structure : le morphème. Mais la partie la plus intéressante de ce travail concerne l'élaboration de la structure formelle. Au début de ce travail, deux solutions (au moins) s'offraient à nous : soit partir d'une structure déjà établie par un linguiste, soit construire notre propre structure. Nous avons opté pour la deuxième solution, ceci pour deux raisons. La première est d'ordre pratique : *Connaître l'existence d'un objet (morphème, syntagme, proposition) ne suffit pas à son identification dans un texte.* Prenons pour illustrer ce propos le cas du morphème : c'est un segment que l'on doit trouver dans toutes les langues (la première articulation de Martinet.) Mais cela ne nous indique en rien comment le trouver à partir d'un texte[77]. Le propos peut être étendu à toutes les autres structures (syntagme, proposition). *Notre principal travail a donc consisté à faire émerger automatiquement les marqueurs d'une structure dans un texte. Puis une fois la structure identifiée (dans notre cas le syntagme et la proposition), mettre au point un algorithme permettant de générer cette structure en partant de ces marqueurs.* Il faut trouver quelles sont les marques formelles qui permettent une identification de ces structures, en partant de l'hypothèse (utilisée dans d'autres travaux et ici validée) que ces structures sont effectivement marquées formellement. Devant ce fait, il nous a paru préférable d'introduire les structures au fur et à mesure que nous identifions de nouvelles marques. Le travail s'est donc réorienté vers la recherche de marqueurs qui permettent une correspondance entre un texte et une structure linguistique. L'émergence de ces éléments a été aidée par la prise en compte d'une conception théorique de la langue : la langue (plus exactement la parole ou l'écrit) est un objet *linéaire*. Cet aspect de la langue nous a conduit à limiter les différentes structures à rechercher. En particulier, nous avons utilisé le fait que ces marqueurs caractéristiques les délimitaient et correspondaient donc à des *marqueurs de frontière*. Ceci a considérablement facilité notre travail : au lieu de rechercher des régularités dans tout le corpus (travail long et fastidieux), nous nous sommes focalisé sur les éléments apparaissant en début et fin d'entre-ponctuations.

[77] Ainsi les universaux de Greenberg ne peuvent nous aider dans notre travail, si ce n'est a posteriori dans une phase de validation, nos résultats ne devant pas les contredire.

Cette recherche a permis la mise au point d'une structure hiérarchique composée des niveaux suivants :

- le morphème
- le syntagme
- la proposition

Le morphème est l'élément de base de la structure. Il est considéré comme indécomposable. Le syntagme et la proposition sont générés grâce à leurs marqueurs de frontière. Chacune de ces structures peut se combiner pour former soit une séquence de même nature, soit une unité supérieure (figure 7.1). On notera l'absence des segments classiques tels que le *mot* et la *phrase*.



TAB. 7.1 – La hiérarchie structurelle retenue.

Toutes ces hypothèses sur la structure des langues ont été testées plus ou moins complètement sur une vingtaine de langues, soit manuellement, soit automatiquement (les algorithmes ne couvrent pas à l'heure actuelle tous les phénomènes décrits). Du point de vue opératoire, des résultats corrects ne peuvent être espérés qu'en centrant le processus de génération des structures sur la proposition, ce qui nous semble réalisable (en partant de la structure SA). La différence la plus importante entre les travaux réalisés précédemment et le nôtre concerne la méthode employée pour catégoriser les mots et morphèmes. Les algorithmes classiques de catégorisation des mots utilisent des techniques de *clustering* qui ne permettent pas une polycatégorisation des mots (section 3.4). Pour parvenir à cette polycatégorisation, nous avons centré notre algorithme sur la notion de *contextes*, en décomposant la catégorisation en deux étapes :

1. la construction des contextes *appropriés* pour chaque catégorie théorique.
2. la catégorisation des mots et morphèmes grâce à ces contextes.

La construction des contextes avant une quelconque catégorisation est réalisable car toutes les structures des langues que nous utilisons possèdent des marqueurs de frontière dits *prototypiques*. La construction des contextes s'appuient sur ces marqueurs facilement identifiables. Seul un *a priori* sur les structures théoriquement possibles permet la construction de ces contextes. Mais la généralité de ces structures permet la prise en compte d'un nombre de langues très grand. Les algorithmes décrits ici sont à notre connaissance les seuls permettant une catégorisation *formelle multilingue* à partir de corpus non annotés et permettant un traitement multilingue. Dans notre mise en œuvre (aussi bien pour la segmentation que pour la génération des structures syntagmatiques et propositionnelles), nous commençons toujours par rechercher des marqueurs prototypiques. Puis nous nous servons de ceux-ci pour étendre notre connaissance (identification de nouveaux morphèmes ou marqueurs de syntagmes et propositions).

Voyons maintenant quelles sont les structures que notre méthode peut appréhender. Les structures manipulées sont le syntagme, la proposition et les couples de ces deux structures. Il nous semble qu'une approche formelle peut construire les deux unités élémentaires que sont le syntagme et la proposition (comme le montre en partie le travail de [Vergne, 1999]). Les couples de propositions nous semblent eux aussi assez faciles à construire : la plupart des langues utilisent des marquages assez précis (mais aucune implémentation n'a été réalisée pour étayer ce propos). Reste les couples de syntagmes. Nous retrouvons là un des problèmes majeurs de l'analyse syntaxique. Si la relation entre certains couples de syntagmes peut être marquée formellement (comme la structure génitive allemande), ce n'est pas le cas pour toutes les relations. On notera que la segmentation en proposition permet l'élimination de certaines relations entre syntagmes (comme il est dit dans la section 4.8.3, deux syntagmes appartenant à deux propositions ne peuvent être en relation de dépendance). Certaines relations entre syntagmes peuvent être détectées grâce aux éléments lexicaux (chapitre 5) de ces syntagmes (chapitre 5), mais la détection de ces relations devient alors contingente au corpus utilisé. Est-ce que d'autres structures peuvent être générées en utilisant cette méthode (la relation anaphorique par exemple) ? La question reste posée (voir paragraphe suivant).

Nous n'avons pas réalisé un système informatique qui permette une génération automatique d'une grammaire, mais ces algorithmes montrent que les marques formelles que nous avons utilisées fournissent un excellent système d'amorçage pour un tel système.

Un intérêt méthodologique Nous nous sommes trop souvent abrité derrière l'objectif méthodologique de ce travail pour justifier de la mauvaise qualité de nos propres résultats opératoires, pour ne pas revenir dessus. Nous pouvons dire que la question qui a guidé notre travail n'était pas *comment ?* mais *avec quoi ?* Notre problème n'était pas de savoir quelle était la meilleure façon d'utiliser telle ou telle ressource, mais de savoir quelles ressources utiliser pour découvrir la structure formelle des langues. Il s'en est suivi un travail de recensement de ces ressources qui, nous le voulions, devaient être formelles. Dans un deuxième temps seulement s'est posée la question de savoir comment les utiliser. Mais une

réponse imparfaite à cette question nous suffisait, si elle permettait de découvrir d'autres ressources. Ainsi, si notre segmentation des mots n'est pas aussi bonne qu'elle pourrait l'être, elle est néanmoins suffisante pour mettre à jour les structures morphologiques des langues. L'important était pour nous de savoir qu'il fallait utiliser la ressource morphologique. Cette utilisation peut paraître triviale, elle est pourtant assez peu utilisée dans les travaux en inférence grammaticale (on trouvera cependant un exemple dans [Brill, 1993]).

Ce travail présenté ici ne doit pas être jugé sur ses résultats opératoires sur telle ou telle langue, mais sur la liste des ressources utilisées. Le point le plus intéressant de ce travail est de savoir comment cette liste a été construite. La notion centrale qui nous a guidé est celle de *structures marquées aux frontières*. Nous utilisons une propriété commune à toutes les langues : *sa linéarité*, déjà notée dans [de Saussure, 1972]. Si cette propriété est connue depuis longtemps, il nous semble qu'elle est souvent utilisée implicitement. Tout notre travail repose sur la prise en compte de cette linéarité, et cherche à répondre à cette question : *quelles sont les structures que nous pouvons construire avec un objet possédant cette propriété* ? La langue est un objet linéaire composé de segments dont les débuts et/ou les fins sont identifiables formellement. Ce point de vue nous a permis une étape supplémentaire dans la formalisation de la méthode distributionnelle. Ensuite, il restait juste à trouver quels étaient ces segments et par quels éléments ils étaient marqués. Cette approche a permis une prédiction des structures possibles en théorie. Nous avons vu que notre unité la plus haute était la proposition. Il reste à monter encore plus haut dans la hiérarchie. [Lucas, 1995] montre que cette notion de structures marquées aux frontières s'applique à des unités beaucoup plus grandes, qui vont jusqu'au niveau du livre.

Vouloir offrir une méthode permettant de construire une théorie générale est peut être plus ambitieux qu'offrir cette théorie. Trop ambitieux diront certains. Au moins essayons. Cela a déjà été essayé dans les années quarante diront d'autres. Oui, mais nous sommes maintenant en possession de l'outil qui manquait à ce travail : l'ordinateur. Bien sûr l'utilisation de cette méthode contraint la théorie mise au point, et cette méthode ne permet pas de traiter de tous les phénomènes linguistiques. Seuls les phénomènes formels sont pris en compte. Une des difficultés de ce travail a consisté à dompter le corpus. Rechercher des régularités formelles. Mais quelles régularités ? Il en existe beaucoup, et l'on peut vite se laisser déborder. Il a fallu ordonner cette recherche. Le critère a été facile : travaillons d'abord sur les éléments (mots et morphèmes) fréquents du corpus. Deux raisons a ceci. Premièrement, plus un élément est fréquent, plus nous possédons de renseignements sur sa distribution. Deuxièmement, plus un élément est fréquent, plus les retombées sur les autres structures du corpus sont grandes. En bref, nous n'allons pas commencer par étudier les structures peu fréquentes du corpus. Ces éléments ne sont pris en compte (ne peuvent être pris en compte) que lorsque les structures fréquentes sont identifiées. Nombre de ces structures restent donc à étudier. Ce travail offre donc une vue très partielle des structures des langues. Il n'en est qu'à son début. Tous les phénomènes linguistiques ne peuvent être pris en compte par cette méthode. Signalons par exemple les phénomènes d'ellipse. Mais si ce travail n'a pas traité toutes les structures

formellent marquées à ses frontières se trouvant dans tel ou tel corpus, il nous semble qu'il met à notre disposition de bons moyens pour les traiter, en appliquant la même méthodologie que celle utilisée pour les structures décrites.

7.2 Les retombées en linguistique

Un travail de validation Ce travail valide un certain nombre de concepts et de méthodes en linguistique. Commençons bien sûr par la méthode distributionnelle. Il nous semble que cette méthode est validée pour deux raisons : d'une part, cette méthode est en pratique opérationnelle et est très adéquate pour ce genre de travail. D'autre part, les résultats fournis correspondent à des connaissances linguistiques déjà connues, résultats qui valident dans une certaine mesure cette méthode. Donc si ce travail n'a pas abouti à la découverte de nouveaux concepts, il a permis la validation expérimentale de concepts connus, comme le syntagme et la proposition. Nous avons pu vérifier l'adéquation de ces concepts à l'objet par une méthode se basant sur l'observation de cet objet.

Durant ce travail, nous avons retrouvé différents faits linguistiques connus. Le premier concerne la typologie des langues. Si l'on considère la typologie donnée par [Tesnière, 1959, page 33] basée sur "le sens du relevé linéaire", nous pouvons réinterpréter les notions de langues centrifuges par langues qui privilégient les marqueurs de début, et les langues centripètes par les langues privilégiant les marqueurs de fin. De plus, nos résultats permettent d'affiner cette classification, puisque nous possédons deux niveaux où les marqueurs de frontière existent : le syntagme et la proposition. Ainsi selon Tesnière, le français est centrifuge et l'allemand centripète. Or cette distinction est beaucoup plus pertinente au niveau de la proposition[78] qu'au niveau du syntagme (les deux langues utilisant des prépositions et des déterminants). En plus des deux niveaux, nous pouvons prendre en considération la position d'un élément (syntagme ou proposition) subordonné relativement à son régissant. Toute la combinatoire possible peut alors servir de critère de classification (Certaines langues peuvent favoriser les marqueurs de début d'un niveau et les marqueurs de fin d'un autre niveau). Comme Tesnière, nous ferons la différence entre une classification se basant sur des critères typologiques (formels) et une classification se basant sur un critère génétique. La notion d'agglutination peut aussi être étudiée en observant la quantité de morphèmes libres ou liés qui composent un syntagme. Le degré d'agglutination peut être calculé comme suit : le rapport entre le nombre de morphèmes grammaticaux libres et le nombre de morphèmes grammaticaux liés à l'intérieur d'un syntagme. La même opération peut se faire au niveau de la proposition. D'une manière générale, il serait intéressant de prendre en compte cette structure formelle dans le domaine de la linguistique comparative, cette hiérarchie structurelle offrant un bon cadre pour une étude comparative des langues.

[78] Nous reprenons l'exemple de Tesnière. Il nous semble que la proposition allemande est assez "neutre" : aussi bien centrifuge que centripète.

Le déchiffrement de langues Il serait intéressant de savoir quelles utilisations de ce travail feraient les linguistes qui travaillent sur le déchiffrement de langues anciennes. La question reste posée.

7.3 Les retombées en Traitement Automatique des Langues

La hiérarchie Le résultat le plus immédiat concerne les unités ainsi définies : morphème, syntagme, proposition. Ces unités ne sont pas nouvelles. La caractéristique le plus frappante n'est pas dans les unités sélectionnées mais dans celle qui ne l'est pas : le *mot*. Comme Martinet le note, cette notion n'est pas pertinente en linguistique générale, dépendant trop de la langue étudiée. Il ne faut pas pour autant rejeter ce segment : il offre un excellent point de départ à un traitement de l'écrit (meilleur que la lettre ou l'entre-ponctuation par exemple). Il est à ce point excellent qu'il a occulté la "vraie" structure linguistique de longueur similaire : le syntagme. Mais il ne demeure pas moins qu'un point de départ. Revenons aux unités décrites dans notre hiérarchie . Elles ne sont pas nouvelles. Nous avons essayé d'en donner une définition aussi formelle et complète que possible. Il nous semble que la définition de la proposition reste encore à approfondir. Il est à signaler que les segmenteurs proposés (de syntagmes et de propositions) sont plus facile à mettre au point qu'un analyseur complet. La segmentation ne semble demander que l'identification du verbe de la proposition ainsi que du premier actant réalisé (identification réalisée par [Giguet and Vergne, 1997]). De tels outils seraient sans doute très appréciés dans une boite à outils en TAL.

L'utilisation de la hiérarchie La retombée la plus immédiate de ce travail doit être la prise en compte des différentes unités linguistiques utilisées dans ce travail : morphème, syntagme, et proposition. On retrouve déjà certaines de ces unités dans nombre de travaux. Les morphèmes, par exemple, sont utilisés par certains analyseurs, pour déterminer la catégorie d'un mot inconnu : ce que l'on nomme les "guessers" ([Chanod and Tapanainen, 1995]). Mais nous voyons qu'ils ne sont utilisés que comme roue de secours (quand un mot manque dans le lexique). Une utilisation plus intéressante est celle développée par [Vergne and Giguet, 1998], où la ressource morphologique est directement intégrée dans le processus d'analyse. On trouvera dans [Giguet, 1996] une utilisation dans le diagnostic de langues, qui montrent que la connaissance des affixes et mots grammaticaux des langues fourni un meilleur résultat que les autres techniques qui utilisent des trigrammes (séquence de trois lettres). Cette unité ne semble pas être utilisée dans les travaux en génération automatique de grammaire, à l'exception de [Brill, 1993]. La notion de syntagme est aussi largement utilisée [Argamon et al., 1998]. Si l'utilisation explicite de ces segments n'est pas nécessaire pour obtenir de bons résultats, les meilleurs sont toujours obtenus par les systèmes mis au point en les prenant en compte. Ainsi [Giguet and Vergne, 1997] qui manipule explicitement la notion de syntagme, fournit le meilleur étiqueteur du français. La notion de proposition est moins

utilisée semble-t-il en TAL (du moins explicitement, mais on trouve souvent les notions d'actants, ou de structures prédicatives). On trouvera une illustration du niveau propositionnel dans [Giguet, 1998] dans le cadre de l'analyse syntaxique.

Tous ces traitements utilisent comme segment de base le mot. Il serait intéressant de voir les avantages qu'apporterait une segmentation des séquences de mots en syntagmes et en propositions dans les traitements automatiques. La première couche décrite dans [Giguet and Vergne, 1997] est assez similaire à une segmentation en syntagmes. Si la segmentation en syntagmes est assez facilement réalisable (l'opération est plus facile qu'un étiquetage, et demande moins de ressources), la question est de savoir si une segmentation en propositions est aussi facilement réalisable ? Une segmentation en propositions ne requiert pas la mise en relation de tous les syntagmes, mais pourrait peut-être aider à cette opération. On en trouvera un exemple dans [Rosmorduc, 1994, page 130], qui segmente un texte égyptien en propositions avant d'effectuer une analyse de ses éléments. Cette segmentation est facilitée par l'existence de "*marqueurs d'initialité* qui existent dans la langue égyptienne". Nous étendons cette remarque à toutes les langues.

Intégration des structures supérieures Les différents analyseurs travaillent au niveau de la phrase. On renvoie le lecteur aux travaux de [Lucas, 1995], où des notions structurelles très similaires (marqueur de début et de fin) sont appliquées à des niveaux très supérieurs à la phrase. Il serait intéressant d'étudier les retombées de l'intégration de ces structures supérieures dans un processus d'analyse.

Les ressources formelles Ce travail illustre les intérêts et les limites des ressources formelles dans un processus d'analyse de textes. Le résultat opératoire le plus immédiat est que les ressources formelles permettent la mise en relation des mots grammaticaux des langues, et dans certains cas, de mettre en relations certains syntagmes d'une proposition. Cette mise en relation est suffisante pour produire une segmentation du texte en syntagmes et en propositions. À noter qu'étiqueter un texte est plus difficile que le segmenter en syntagmes, puisque les étiquettes généralement utilisées sont plus fines que celles nécessaires à une mise en syntagme (marqueurs de frontière).

L'inférence grammaticale Resituons les résultats obtenus avec les autres travaux en inférence grammaticale. Commençons par les différences. Il en existe trois principales :

- utiliser des données brutes
- ne pas utiliser de ressources spécifiques à une langue donnée
- avoir une approche multilingue

Si certains de ces critères se retrouvent dans certains travaux individuellement, nous n'avons retrouvé la combinaison des trois dans aucun autre travail. Comme la section 7.4 le montrera, ces pré-requis que nous nous sommes donnés, n'ont pas été une entrave à notre travail. Mais il a fallu trouver dans la langue les indices qui permettaient ce travail. Ces différences avec les autres travaux font

que toute comparaison est délicate. Nous noterons aussi une différence méthodologique. Nous ne concevons pas les travaux en inférence grammaticale comme ayant pour objectif la génération automatique d'outils d'analyse, mais d'un point de vue plus théorique : cette tâche de découverte ne peut se faire qu'en utilisant des propriétés fondamentales (des structures) des langues. Ce point de vue nous rapproche plus de travaux comme ceux de [Brent, 1996], où la question de l'acquisition par les enfants de leur langue maternelle est centrale. Une question intéressante est de savoir comment l'enfant amorce cet apprentissage (problème du "bootstrapping"). On trouvera dans [Finch, 1993, pages 77-79], les différentes hypothèses émises à ce sujet. Quatres pistes sont données :

- l'amorce distributionnelle
- l'amorce syntaxique
- l'amorce sémantique
- l'amorce prosodique

[Pinker, 1984] pour sa part opte pour une interaction entre l'approche syntaxique et sémantique , en jugeant l'amorce distributionnelle irréaliste[79]. Un reproche que fait Pinker à l'hypothèse distributionnelle est que l'enfant se serait comment choisir parmi toutes les régularités possibles :

> The properties that a child can detect in the input -such as the serial positions and adjacency and co-occurrence relations among words- are in general linguistically irrelevant. [Pinker, 1984, page 55]

Notre travail semble montrer qu'une *amorce* purement distributionnelle est envisageable, mais nous parlons uniquement de l'amorce de l'apprentissage. Au lieu de dire que les critères extra-linguistiques peuvent servir à l'amorce d'un système d'apprentissage d'une grammaire, et qu'ensuite les critères distributionnels son utilisés ([Finch, 1993, page 75]), nous pensons que l'inverse est tout aussi envisageable. Notons que notre étude a porté uniquement sur des textes écrits. Or l'acquisition d'une langue par un enfant se fait de manière orale. Notre travail est-il transposable à une étude du corpus oral ? Certains travaux [Abney, 1992], [Wanner and Gleitman, 1982] mettent en parallèle structures syntaxique et prosodique. Un travail intéressant serait d'appliquer notre méthode à un corpus oral. De manière similaire, nous partirions des segments de la strate orale (sans doute syllabe, groupe prosodique), et essayerions de construite la strate grammaticale. Si cette dernière correspond à notre strate, alors nous aurions une validation de celle-ci. Nous pensons donc que les hypothèses de l'amorce distributionnelle et de l'amorce prosodique ne s'opposent pas, mais plutôt se confortent l'une l'autre.

Une autre question intéressante est de savoir si l'enfant possède déjà la connaissance de la hiérarchie (morphème, syntagme, proposition), et donc "n'a plus" qu'à l'instancier pour sa langue, ou bien, s'il ne la connaît pas et qu'il doit la détecter.

[79] On notera que les travaux privilégiant l'approche sémantique basent trop souvent leurs réflexions sur des phrases artificiellement simples (les fameuses phrases de trois mots). Or les énoncés auxquels l'enfant est soumis sont autrement plus complexes.

7.4 Le travail multilingue

Quel est l'intérêt de travailler sur plusieurs langues à la fois, si ce n'est d'augmenter la difficulté du travail. Voyons d'abord pourquoi nous avons travaillé sur plusieurs langues. Essayant d'appliquer bien sagement les idées de Harris, nous voulions ne prendre en compte que des critères formels dans notre étude. Or travaillant sur le français, nous nous sommes aperçu que notre connaissance de cette langue, ainsi que les attendus que nous avions sur ces structures nous empêchaient de travailler uniquement avec les critères formels. Notre solution a été de travailler sur des langues que nous ne connaissions pas. Dans cette configuration, seuls les critères formels sont utilisés. Aucune considération du sens ne peut être prise en compte. Nous pouvons donc considérer ce travail multilingue comme étant une contrainte liée à la méthode.

Mais cette contrainte a eu deux effets bénéfiques. Premièrement, les structures manipulées sont multilingues. Il n'était pas évident, au début de ce travail, que toutes les langues partageaient une même structure. La retombée la plus immédiate a été l'abandon du mot comme unité linguistique au profit du syntagme.

Le deuxième effet n'est pas apparu immédiatement, mais à la fin de ce travail. L'intégration du niveau propositionnel en est le meilleur exemple. Ce niveau a été intégré à la hiérarchie parce qu'il facilitait énormément le travail de découverte des structures allemandes (section 4.7). En confrontant ce niveau avec les autres langues, nous nous sommes aperçu qu'il était bénéfique dans le traitement de toutes les langues. Il existe des structures formelles très marquées dans certaines langues, et plus discrètes dans d'autres. Dans le premier cas, la connaissance de ces structures est nécessaire pour manipuler ces langues. L'importation de ces structures vers d'autres langues a généralement des retombées positives sur le traitement de ces dernières. Nous pouvons prendre aussi l'exemple du syntagme. Notre étude a commencé par les langues européennes dans lesquelles le syntagme est très fortement caractérisé. Lors de l'étude du vietnamien, cette structure lui a été appliquée bien qu'elle ne soit pas très caractérisée dans cette langue. Si notre étude avait commencé par le vietnamien, le syntagme n'aurait pas été introduit aussi vite.

Un problème se pose dans une telle étude : comment valider les résultats dans une langue que l'on ne comprend pas ? Cette validation est assez facile à réaliser pour l'opération de segmentation et de construction des syntagmes, les informations formelles étant très présentes (un lexique de la langue suffit dans la plupart des cas). Pour les structures supérieures (construites par une mise en relation des syntagmes), la validation est beaucoup plus délicate et nécessite un locuteur de la langue.

Une question intéressante est de savoir jusqu'où un travail multilingue peut conduire, c'est-à-dire à quel moment doit-on prendre en considération les spécificités de la langue étudiée ? Pour répondre à cette question, une étude complète d'une langue doit être réalisée grâce à cette méthode, ce qui n'a pas été fait.

Annexes

Tous les algoritmes, données, résultats sont accessibles à partir de la page :

`www.info.unicaen.fr/~dejean/these/`.

Annexe A

Détail des corpus utilisés

Langue	nom	type	taille (mots)
allemand	*allemand*	roman	150666
anglais	*anglais*	Tom Sawyer (roman)	40479
	anglais	From earth to moon (roman)	73633
arabe	*arabe*	le Coran	81224
chinois	*chinois*	la Bible	??
coréen	*coréen*	le Nouveau Testament	76780
espagnol	*espagnol*	le Pentateuque	199920
français	*français01*	Le monde	266047
	français02	la Bible	767223
	français03	Le monde (étiqueté)	168511
indonésien	*indonésien*	le Coran	68581
italien	*italien01*	évangile selon Saint Jean	17283
	italien02	journal	50985
japonais	*japonais*	le Nouveau Testament	??
polonais	*polonais*	le Nouveau Testament	173866
quechua	*quechua*	évangile selon Saint Jean	27245
russe	*russe*	textes administratifs	57578
swahili	*swahili*	le Nouveau Testament	128273
turc	*turc01*	le Nouveau Testament	129909
	turc02	rapport technique	33001
vietnamien	*vietnamien*	le Nouveau Testament	93861

Les résultats sont obtenus avec la commande Unix *wc* qui a une définition du mot très proche de la nôtre. Le comptage des mots peut varié si l'on prend ou non en compte les signes de ponctuations. Aucune valeur n'a été donnée pour les corpus japonais et chinois. Les corpus sont accessibles à l'adresse suivante : `www.info.unicaen.fr/~dejean/these/donnees/corpus/`

Annexe B

Les outils et programmes

B.1 Les outils

La tokenisation en mots des corpus a été écrite en `Flex`. Dans un premier temps, les algorithmes de segmentation ont été développés en `C++`. Puis,

Le langage de programmation `Perl` s'est révélé très adapté à notre travail sur corpus, permettant un maquettage rapide des algorithmes grâce aux expressions régulières. Les autres outils correspondent aux commandes `Unix`, principalement la commande de tri `sort`. La commande `match` écrite en `Perl` nous a servi d'outil d'observation des données. Elle permet la visualisation des concordances d'expressions régulières.

B.2 les programmes

Le listing des différences programmes est donné à l'adresse :

`www.info.unicaen.fr/~dejean/these/programmes/index.html`

Annexe C

Résultats obtenus sur différentes langues

Pour chaque langue est donnée :
- la liste de morphèmes prototypiques (suffixes et préfixes)
- la liste complète des morphèmes
- le schéma contextuel des SA
- le schéma contextuel des SR

Nous rappelons que tous les résultats ont été obtenus avec les mêmes algorithmes et les mêmes paramètres. Les morphèmes jugés incorrects sont en *italique.* Les résultats des langues étrangères ont été, en partie, validés grâce à des grammaires et dictionnaires de ces langues. L'ouvrage de référence dont nous nous sommes servi est [Malherbe, 1995] qui donne une liste des mots les plus courants (noms, pronoms, verbes, adjectifs, adverbes) pour 171 langues. Même si la description de ces langues est très sommaire, cet ouvrage permettait une validation (ou non) très rapide, en particulier pour les structures SA grâce à la liste des pronoms.

C.1 allemand

Liste de morphèmes prototypiques

suffixes

-ige -liche -ere -er -tet -es -end -ung -lich -el -en

préfixes

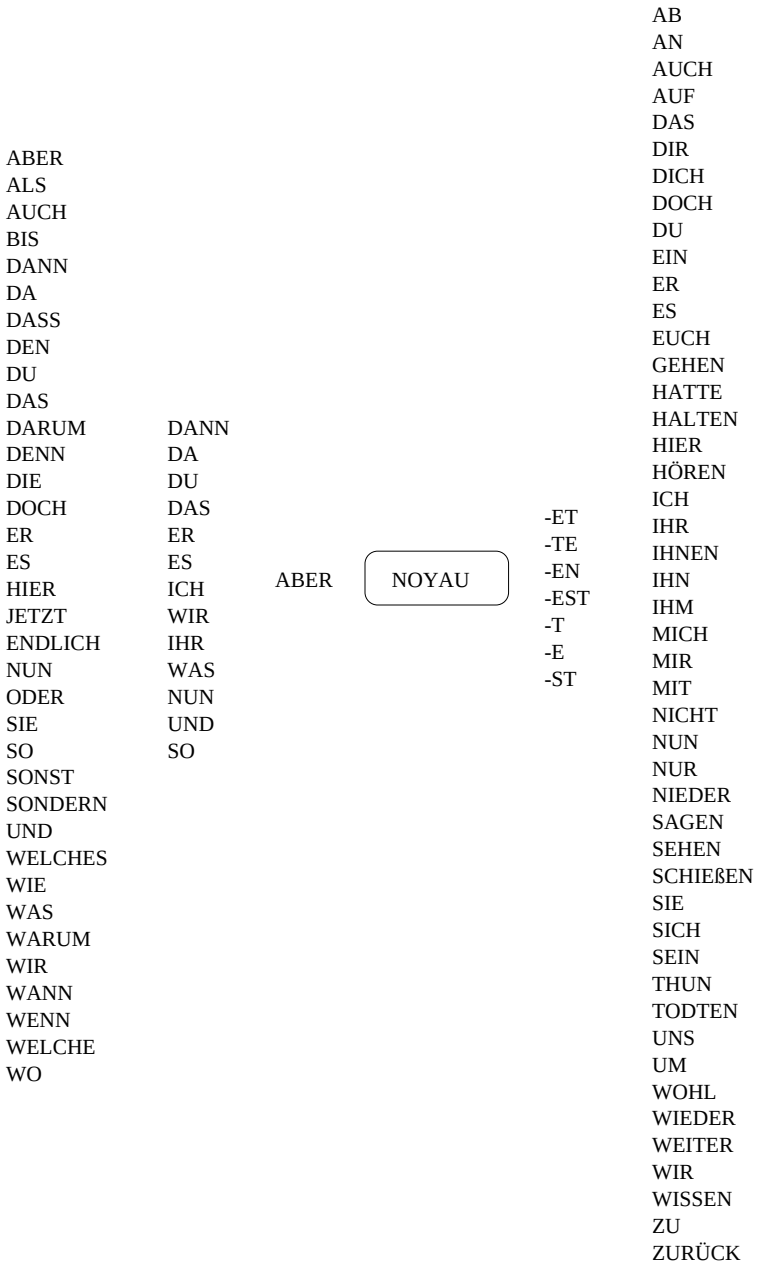
wi- un- be- ge- ver- *ma-* *le-* *me-* *ne-* über- er- nach- auf- *sch-* her- hin-

Liste complète des morphèmes

-lich -ern -st -ung -ste -ige -te -test -igen -liche -e -ten -iger -iges -eren -ter -est -eten -*tes* -tet -n -eres -ig -el -em -tete -en -lichen -t -end -er -licher -eses -et -sten -ere -ete wi- un- be- ge- ver- *ma-* *le-* *me-* *ne-* über- er- nach- auf- *sch-* her- hin-

Aucun nouveau préfixe n'est généré.

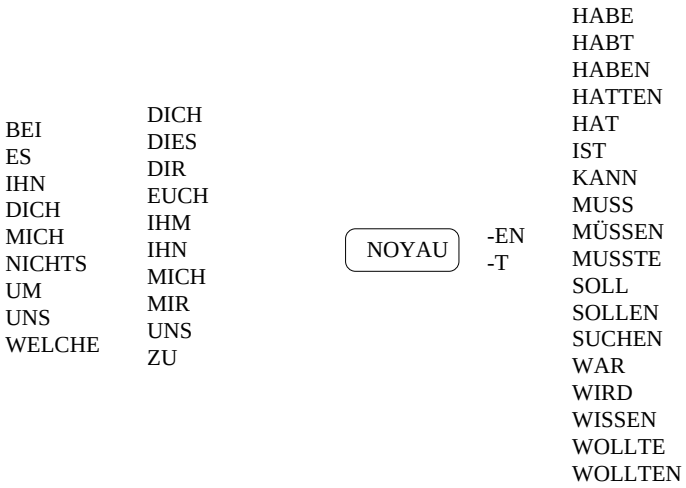
Schéma contextuel des SAD



analyse des résultats L'allemand illustre parfaitement le fait qu'un élément peut apparaître dans plusieurs positions du schéma. En particulier les pronoms sujets (*ich, du, er, es, sie, wir, ihr, das*) sont aussi bien des marqueurs de début que de fin. La liste des terminaisons verbales liées est complète. La catégorie des marqueurs de fin correspond à une fusion partielle des trois catégories : FSAD, SAF et FP. De même la catégorie des marqueurs de débuts correspond aux catégories des DP (conjonction, adverbe) et des DSAD (pronom

sujet). Le schéma contextuel permet une identification correcte des structures SAD à 82%. Les principales erreurs proviennent non pas d'une mauvaise catégorisation, mais d'une mauvaise analyse : certaines prépositions dans le contexte [SAD préposition] sont identifiées comme fin de SAD.

Schéma contextuel des SAF



analyse des résultats Les SAF correspondent aux verbes terminant les propositions allemandes, incluant la structure classique *zu N-en* mais aussi la structure [*préposition pronom verbe*] comme :

bei dir bleiben

la séquence maximale reconnue par ce schéma étant du type [*préposition pronom verbe auxiliaire*] :

bei dir gesehen habe

Se pose ici la question de avoir si nous considérons ce type de séquences comme un syntagme simple ou bien un couple de syntagmes formé des éléments *bei dir gesehen* et *habe*. Une étude plus spécifiques de ces séquences verbales est nécessaire pour apporter une réponse. Nous voyons que le nombre de prépositions identifiées est assez faible. La catégorie des fins correspond aux différents auxiliaires (*haben*, *sein*, *werden*,...).

L'identification des SAD et SAF allemands permet d'avoir un bon aperçu de la structure propositionnelle. Nous voyons que la structure SAF bruite la structure SAD (des SAF apparaissent comme fin de SAD). Encore une fois, la prise en compte de la structure propositionnelle dans son ensemble permettrait de meilleurs résultats.

Schéma contextuel des SR

	ALS		
	AM		
	AUS		
	AUCH		
	DEIN		
	DREI		
	DIESE		
	DIE		
	DEM		
	DEN		
	DAS		
	DER		
HALEF	DES		
AUF	DEINE		
DENN	DIESEM		
WIE	DIESER		-E
NICHT	EINE		-EN
ZWISCHEN	EINEM		-ET
BEI	EINEN		-ER
ERST	EINER		-ES
VOR	EIN		-ERE
MEHR	EINIGE		-IG
ÜBER	ES		-IGE
DURCH	EURE	NOYAU	-IGEN
HINTER	GANZ		-IGER
AN	IHRE		-HEIT
OHNE	IM		-KEIT
ABER	IN		-N
BEREITS	JETZT		-M
DORT	KEINE		-S
DA	MEINE		-ST
ALSO	MEIN		-STE
NUR	MEINEN		-STEN
UM	NOCH		-T
UNTER	NUN		-UNG
HEUTE	SO		
GEGEN	SEHR		
FÜR	SEINEM		
IST	SEINEN		
	SEINER		
	SEINE		
	SOFORT		
	UND		
	UNSERE		
	VOM		
	VON		
	VIELE		
	ZWEI		
	ZU		
	ZUM		
	NACH		

analyse des résultats Aucun marqueur de fin de SR n'est identifié. La position (1) comprend des marqueurs de début de SR (préposition), mais aussi des marqueurs de début de proposition (conjonction, adverbe). La position (3) comprend essentiellement des déterminants et des prépositions. Nous trouvons

aussi certains adjectifs.

schéma contextuel des SSub Aucun élément n'a été identifié comme SSub avec les algorithmes actuels. La génération des structures d'accord permet l'identification de certaines constructions [*déterminant adjectif substantif*].

C.2 anglais

Liste de morphèmes prototypiques

suffixes

-ance -ence -age -able -ture -ate -er -n't -*ight* -ment -est -ly -er's -ings -ations -ers -ness -ous -ed -ing -ish -th -al -ow -ic -ation

préfixes re- *the-* un- in- pro- per- *sha- sho- da- de- du-* dis- *do- for- gra- ha- hu- hi- ho- ju- jo- la- le- li- lo- ma- me- mu- mi- mo-* con- *va- vi- ne- ni- no-*

Nous voyons donc que la génération des préfixes est très mauvaise (80% d'erreurs). Mais ces erreurs ne gênent aucunement la construction des couples morphologiques.

Liste complète des morphèmes -ance -ence -age -able -ture -ate -er -n't -*ight* -ment -est -ly -er's -ings -ations -ers -ness -ous -ed -ing -ish -th -al -ow -ic -ation -e -s re- *the-* un- in- pro- per- *sha- sho- da- de- du-* dis- *do- for- gra- ha- hu- hi- ho- ju- jo- la- le- li- lo- ma- me- mu- mi- mo-* con- *va- vi- ne- ni- no-*

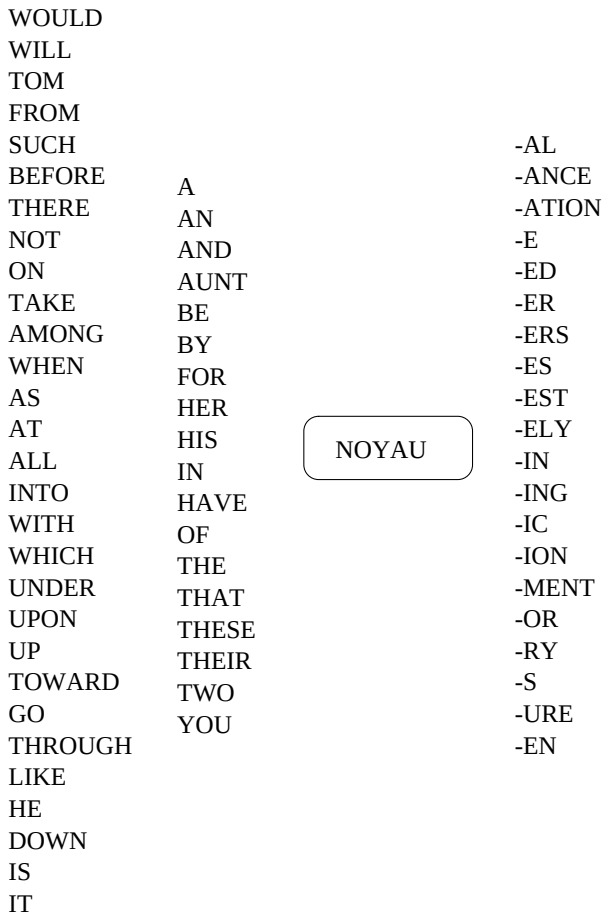
Les seuls nouveaux morphèmes sont -*e* et -*es*.

Schéma contextuel des SAD

analyse des résultats Le faible nombre de marqueurs de début (les conjonctions en particulier) s'explique par la faible variation morphologique du système verbal anglais. Il faudrait effectuer cette génération en utilisant la notion de *classe lexicale* 6.4.6. Néanmoins, nous obtenons un schéma contextuel assez représentatif du syntagme verbal anglais (en particulier la catégorie des marqueurs de fin est bien détectée). On noera que la forme négative du groupe verbal (*don't, didn't,...*) n'est pas reconnue.

Schéma contextuel des SAF Aucun SAF n'a été généré.

Schéma contextuel des SR



analyse des résultats On trouve un certain nombre d'erreurs parmi les marqueurs de début, en particulier des verbes fréquents (*would, will*). Une meilleure couverture des SAD permettrait d'éviter ce type d'erreur. On trouve aussi des prénoms (*tom*). Quelques élément de SAD se retrouvent aussi, mais ils sont identiquement catégorisés comme marqueur de début (*he, it*). Les affixes correspondent aux terminaisons nominales et adjectivales de l'anglais.

C.3 coréen

Liste de morphèmes prototypiques

suffixes

-가 -에 -를 -하여 -나 -에서 -리 -도 -만
-은 -을 -려 -는 -느냐 -매 -신 -아 -의
-케 -이 -와 -로 -니 -한 -할 -에게 -다
-고자 -야 -기 -이요 -며 -면 -고 -어
-라 -지 -히 -던 -과

préfixes Aucun préfixe n'est trouvé.

Liste complète des morphèmes

-에 -니 -게 -대로 -에게 -아 -러 -의 -여 -이나
-들이 -도 -느냐 -이 -다 -케 -이니 -이요 -가 -하고
-까지 -지 -려 -되 -으로 -고 -야 -를 -며 -면
-신 -로 -리 -서 -에게로 -라 -에서 -과 -하는 -던
-어 -와 -에게서 -더러 -라' -만 -기 -이라 -에는 -는
-고자 -한 -들을 -할 -은 -을 -매 -나 -하여 -보다 -히

analyse des résultats Nous voyons que la segmentation génère des éléments composés d'un nombre pairs de caractères. Nous retrouvons bien le principe de codage utilisant 2 octets pour cet alphabet. Le même algorithme est utilisé pour les systèmes d'écriture européens. La seule différence se situe dans la liste des signes du systèmes. La segmentation se fait en lettres ou couples de lettres (certains couples d'octets correspondent à des combinaisons de deux lettres). La validation est très superficielle : elle ne concerne qu'une demi douzaine de morphèmes. Mais nous retrouvons bien les caractéristiques d'une langue agglutinante (40 éléments prototypiques). De plus la liste de morphèmes prototypiques génèrent bien de nouveaux morphèmes, ce qui est une caractéristique d'une bonne segmentation.

Schéma contextuel des SAD Non traité

Schéma contextuel des SAF Non traité

Schéma contextuel des SR Non traité

C.4 français

Liste de morphèmes prototypiques

suffixes

-era -ez -ance -ence -age -*che* -ologie -able -elle -isme -*ine* -ienne -ière -aire -ture -resse -euse -ante -iste -ette -ique -er -ement -eau -es -*and* -ard -ing -e-t-il -aux -eux -ation -isé -ité

préfixes

anti- auto- en- ex- re- trans- uni- in- par- per- *pla-* pro- *qu'a-* saint– *s'a-* *sou-* *sta-* *da-* *de-* d'- dis- dé- *do-* *du-* *gen-* jean– l'- mont- con- *ver-* *bou-* *n'a-*

Liste complète des morphèmes

-a -able -ables -age -ages -aient -aire -aires -ait -ance -ances -ant -ante -antes -ants -ard -ateur -ateurs -ation -ations -aux -e -e-t-il -eau -elle -ement -ements -ence -ent -era -erait -eront -ette -eur -eurs -euse -eusement -euses -eux -ez -ie -ienne -ier -iers -ing -ion -ions -ique -iques -isation -iser -isme -iste -istes -isé -isée -ité -ités -ière -ières -ologie -ons -resse -s -ture -é -ée -ées -és anti- auto- *bou-* con- d'- *da-* *de-* dis- *do-* *du-* *dé-* *en-* ex- *gen-* in- jean– l'- mont- *n'a-* *par-* *per-* *pla-* pro- *qu'a-* re- *s'a-* saint– sou- *sta-* trans- *uni-* *ver-*

Schéma contextuel des SAD

OÙ	JE	LE		-ONS	
ET	IL	EN		-AIENT	
MAIS	ELLE	S'Y		-ANT	
CAR	ON	NE		-RA	
COMME	QU'ON	LES		-E	
DONT	QUI	NOUS	NOYAU	-ER	PAS
SI	NOUS	N'EN		-ENT	
QUAND	ILS	LEUR		-ERA	
S'IL	ELLES	SE		-IONS	
CE	EN	Y		-AIS	
CEUX	POUR	LUI		-IT	
CELA	EST	S'EN		-AIT	
		N'Y			

analyse des résultats La description du schéma est donnée à la section 6.4.4.

Schéma contextuel des SAF Aucune structure SAF n'a été générée.

Schéma contextuel des SR

			-ONS
			-CATIONS
			-S
	À		-T
AVEC	UN		-AT
TOUTES	LES		-RE
QUE	NOUS		-ER
CONTRE	ÉTÉ		-RE
SUR	UNE		-ES
DANS	DE		-IONS
FAIT	SES		-EMENT
DE	LEURS		-TION
PAR	LA		-ENTS
PLUS	SA		-UE
EN	AUX		-ENCE
ET	SON	NOYAU	-ATION
EU	LE		-IR
AUSSI	D'UNE		-ÉS
DEVANT	EN		-ON
PEU	SA		-EURS
COMME	SANS		-IER
POUR	CETTE		-ÉES
L'UN	AU		-E
LOIN	SE		-ITÉ
ENCORE	LEUR		-ION
AINSI	DU		-ENT
À	DES		-É
			-EMENTS
			-TIONS
			-IE
			-ATIONS

analyse des résultats Nous retrouvons essentiellement des structures nominales et quelques traces verbales (*se*, *eu*). La grande majorité des structures analysées par ce schéma sont (bien sûr) des structures prépositionnelles (qui sont les structures les plus fréquentes).

C.5 turc

Liste de morphèmes prototypiques

suffixes

-ınca -ında -larla -larıyla -ına -lara -makta -maya -maz -mez -ince -inde -lerle -iyle -lerine -lere -meye -lar -ler -dır -ıyor -ıyordu -usu -ındaki -lerini -leri -mesi -ıp -us -arak -acak -mak -elim -eyim -acaım -ayım -ıyorum -larından -us'un -unun -inden -inin -larının -ların -masın -mayın -mı -ünü -ladı -madı -ıldı -malarını -ları -mayı

préfixes

ara- ta- ya- süre- ge- gü- ha- ma- me- büyü- ne-

Tous les préfixes sont incorrects. Ils ne génèrent aucun autre élément.

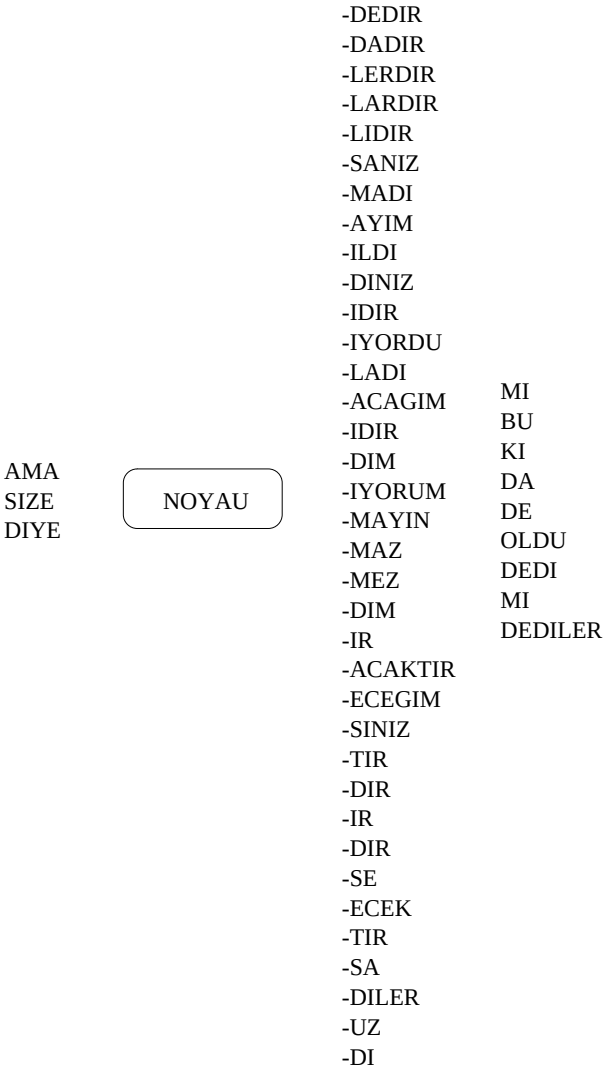
Liste complète des morphèmes

-’ -a -acak -acaktır -acağım -an -arak -ayım -ca -da -dadır -daki -dan -de -dedir -deki -den -di -diler -dim -dir -dı -dım -dınız -dır -e -ecek -eceğim -ek -en -eniz -erek -i -idir -im -imi -imiz -imize -imizi -in -ince -inde -indeki -inden -ine -ini -inin -iniz -inize -inizi -ip -ir -iyle -iz -izde -izden -ize -izi -izin -izle -ken -la -ladı -lar -lara -larda -lardan -lardır -larla -ları -ların -larına -larında -larından -larını -larının -larınız -larıyla -le -ler -lerden -lerdir -lere -leri -lerin -lerinden -lerine -lerini -lerinin -leriyle -lerle -li -lidir -lik -lı -lık -madı -mak -makta -maya -mayı -mayın -maz -mek -mesi -meye -mez -miş -mış -nin -nın -sa -sanız -se -si -sin -sine -sini -sı -sın -sına -sında -sını -sınız -sız -ta -taki -tan -te -ten -ti -tir -tı -tır -u -un -us -usu -uz -ya -ye -yi -yla -yle -yı -ü -üm -ün -ünü -ı -ıdır -ıldı -ım -ıma -ımı -ımın -ımız -ımıza -ımızı -ımızın -ın -ına -ınca -ında -ındaki -ından -ını -ının -ınız -ınıza -ınızda -ınızı -ınızın -ıp -ır -ıyla -ıyor -ıyordu -ıyorum -ız -ıza -ızda -ızdan -ızı ara- ta- ya- süre- ge- gü- ha- ma- me- büyü- ne-

La liste est composée de 195 éléments. Tous les suffixes peuvent être considérés comme corrects. La plupart des éléments correspondent à des compositions de morphèmes.

Schéma contextuel des SAD Aucune structure SAD n’a été générée.

Schéma contextuel des SAF



analyse des résultats La génération de la structure SAF turque est réalisée à partir de la structure *N-F*, puisque cette structure est déjà caractérisée par une position absolue (les structures des autres langues étaient réalisées à partir du modèle [*D N-F*]). Les marqueurs de débuts correspondent soit à des conjonctions (*ama* (mais), *çünkü* (car)),soit à des pronoms (*o* (il), *ben* (je)). Les marqueurs de fin correspondent à des marqueurs interrogatifs (*mi*, *mı*). Nous retrouvons bien les différents éléments rencontrés dans les structures SA des langues déjà étudiées. On trouve aussi des nom propres (*ısa* (Jésus), *rab* (maître)).

Schéma contextuel des SR Non traité

C.6 vietnamien

Liste de morphèmes prototypiques

suffixes

-ng -nh -eâ -oâ

préfixes

gia- ra- sa- nha- qua- ch- no- ma- na

Liste complète des morphèmes Aucun autre morphème n'est trouvé à partir de la liste des morphème prototypiques. Nous en concluons que la langue n'est pas morphologique : la segmentation n'est pas retenue.

Schéma contextuel des SAD les classes lexicales n'ont pas été implémentées pour le moment.

Schéma contextuel des SAF les classes lexicales n'ont pas été implémentées pour le moment.

Schéma contextuel des SR les classes lexicales n'ont pas été implémentées pour le moment.

C.7 swahili

Liste de morphèmes prototypiques

suffixes

-isha -alia -ara -olewa -ishwa -uliwa -aje -ane -aye -ishi -ali -eni -asi -ano -avyo

préfixes

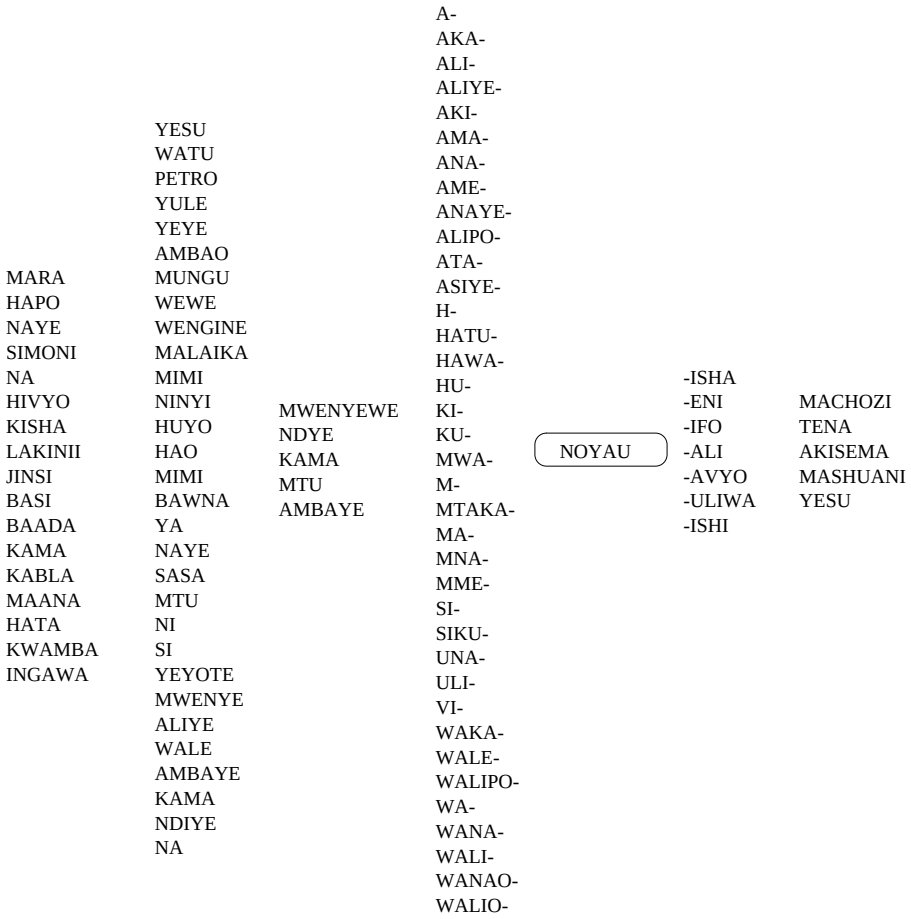
aka- ali- ame- ana- asi- ata- zi- tu- ya- uka- uli- ume- una- usi- utaka- ika- ili- ime- ina- ita- pa- si- ha- ji- ku- li- mka- mli- mme- mna- msi- mta- mwa- wa- vi- ba- ni-

Liste complète des morphèmes

-isha -alia -ara -ishwa -uliwa -aje -ane -aye -ifu -ishi -ali -eni -asi -iko -ano -yavyo uta- hamku- walio- kilicho- nita- wali- yali- sita- tuna- a- haya- h- haku- i- k- walipo- m- alivyo- u- tuki- ham- asiye- hau- nina- hatu- tuli- aliyo- iliyo- aki- mki- waka- uki- wame- niki- mana- ange- alio- mlio- nitakapo- nitaka- ulio- hata- nili- wata- yata- alipo- hawaku- ataka- walivyo- mtaka- tuka- tume- siku- hu- hai- wanao- atakaye- wana- yana- aliye- tuta- hawa- nika- nime- waki- anaye- ki- -wasi-

Les préfixes sont donc beaucoup plus dévelopés que les suffixes.

Schéma contextuel des SAD



Parmi les marqueurs de débuts, on trouve en grande partie des conjonctions (*lakini, kama, na*) et des pronoms (*mimi, wewe, yeye, watu, wote, sisi, ninyi*). On trouve aussi des noms propres et des noms fréquents (*bwana (monsieur), mtu (homme)*), ce qui arrive fréquemment lorsqu'un syntagme nominal est composé d'un seul mot. La fin du SAD est composé d'adverbes et du nom propre *yesu*. Les affixes correspondent bien à des affixes verbaux.

Schéma contextuel des SAF Non traité

Schéma contextuel des SR Non traité

Annexe D

Quelques résultats d'algorithmes de clustering

Voici quelques exemples de classification des mots. Nous avons pris les vingt mots les plus fréquents de notre corpus *français01*, et les avons classés selon différents contextes. Le contexte est construit avec les cent mots les plus fréquents. Le commentaire est donné à la section 3.3.3. La classification a été effectuée à partir de l'outil développé dans [Guilpin and Caron, 1997], qui crée une interface utilisateur pour les opérations de clustering du logiciel *Splus* [Baumgarten, 1994]. La distance utilisée est la distance *binaire*.

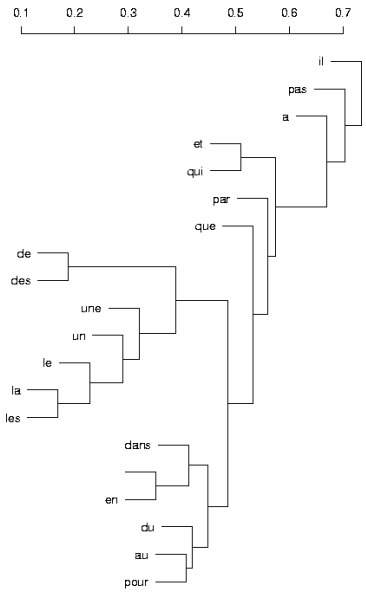


FIG. D.1 – Catégorisation de mots : contexte : un mot avant

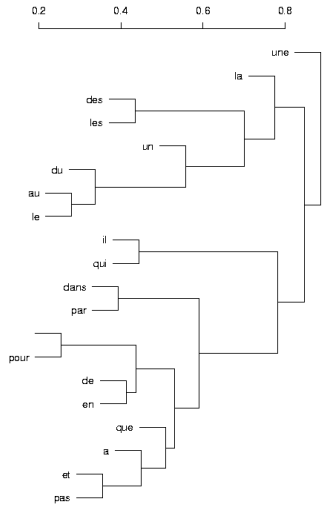


FIG. D.2 – Catégorisation de mots : contexte : un mot après

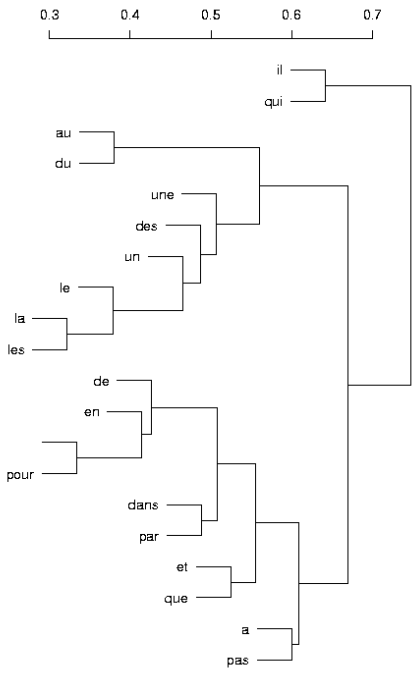


FIG. D.3 – Catégorisation de mots : contexte : un mot avant et après

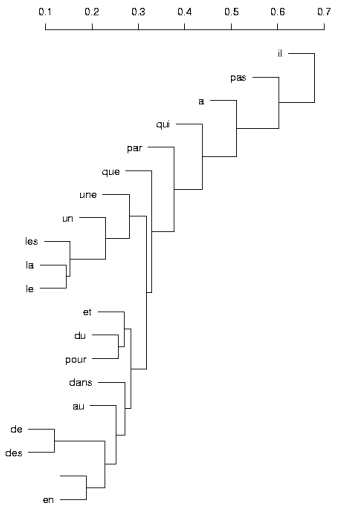


FIG. D.4 – Catégorisation de mots : contexte : deux mots avant

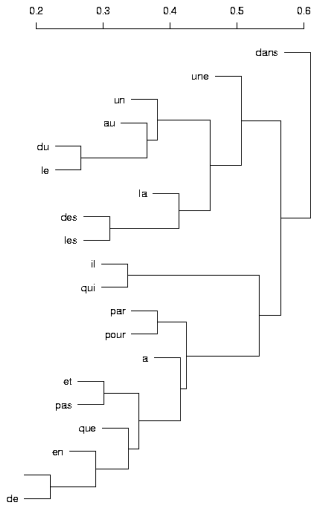


FIG. D.5 – Catégorisation de mots : contexte : deux mots après

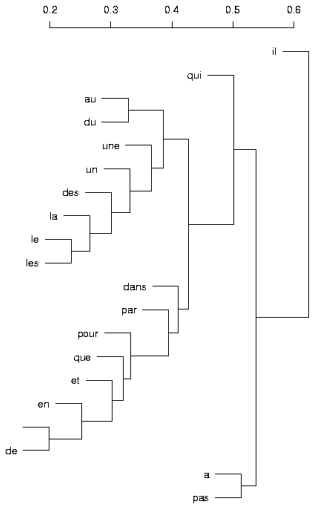


FIG. D.6 – Catégorisation de mots : contexte : deux mots avant et après

Index

Index

Bibliographie

[Abney, 1992] Abney, S. (1992). Prosodic structure, performance structure and phrase structure. In *Speech and Natural Language Workshop*, pages 425–428. Morgan Kaufmann.

[Abney, 1995] Abney, S. (1995). Chunks and dependencies : Bringing processing evidence to bear on syntax. In *Computational Linguistics and the Foundations of Linguistic Theory*.

[Andreewsky, 1973] Andreewsky, A. (1973). *Apprentissage, analyse automatique du langage, application à la documentation*. Paris : Dunod.

[Antworth, 1990] Antworth, E. L. (1990). Pc-kimmo : a two-level processor for morphological analysis. *Academic Computing*, 16.

[Argamon et al., 1998] Argamon, S., Dagan, I., and Krymolowski, Y. (1998). A memory-based approach to learning shallow natural language patterns. In *COLING'98*, Montréal.

[Aristote, 1990] Aristote (1990). *Poétiques*. Livre de Poche.

[Arnauld and Lancelot, 1660] Arnauld, A. and Lancelot, C. (1660). *la grammaire générale et raisonnée (réédition (1969))*. Foucault, Paris.

[Baumgarten, 1994] Baumgarten, M. (1994). *Une introduction à S-plus*. École polytechnique fédérale de Lausanne.

[Benveniste, 1966] Benveniste, E. (1966). *Problèmes de linguistique générale*. Éditions Gallimard, Paris.

[Bloomfield, 1933] Bloomfield, L. (1933). *Language*. Holt and Winston.

[Bouaud et al., 1997] Bouaud, J., Habert, B., Nazarenko, A., and Zweigenbaum, P. (1997). Regroupement issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. In *Ingénierie des connaissances*, pages 207–223, Roscoff.

[Bourigault, 1993] Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *ATALA revue t.a.l.*, 34(2).

[Brendel et al., 1986] Brendel, V., Beckmann, J., and Trifonov, E. (1986). Linguistics of nucleotide sequences : Morphology and comparison of vocabulaires. *Journal Biomol Struture Dyn*, 4 :11–21.

[Brent, 1996] Brent, M. (1996). Advances in the computational study of language acquisition. *Cognition*, 61 :1–18.

[Brent and Cartwright, 1996] Brent, M. and Cartwright, T. A. (1996). Distributional regularity and phonetic constraint are useful for segmentation. *Cognition*, 61 :93–125.

[Brent et al., 1995] Brent, M., Murthy, S. K., and Lunsberg, A. (1995). Discovering morphemic suffixes : A case study in mdl induction. In *Fifth International Workshop on AI and Statistics*.

[Brill, 1993] Brill, E. (1993). *A Corpus-Based Approach to Language Learning*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

[Brill, 1995] Brill, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. In *Workshop on Very Large Corpora, ACL'95*.

[Brown et al., 1992] Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-grams models of natural language. *Computational Linguistics*, 18(4) :467–479.

[Carroll, 1994] Carroll, L. (1994). *Through the looking glass*. Penguin Popular Classics.

[Cartwright and Brent, 1997] Cartwright, T. A. and Brent, M. R. (1997). Syntactic categorization in early language acquisition : formalizing the role of distributional analysis. *cognition*, 63(2) :121–170.

[Champollion, 1997] Champollion, J. F. (1997). *Grammaire égyptienne*. Solin Acte sud (Réédition).

[Chanod and Tapanainen, 1995] Chanod, J. P. and Tapanainen, P. (1995). Create a tagset, lexicon and guesser for a french tagger. In *ACL SIGDAT workshop : From Texts To Tags : Issues In Multilingual Language Analysis*, University College Dublin, Ireland.

[Charniak, 1993] Charniak, E. (1993). *Statistical Language Learning*. A bradford Book, The MIT Press.

[Chatman, 1955] Chatman, S. (1955). Immediate constituents and expansion analysis. *Word*, 11 :377–385.

[Chomsky, 1965] Chomsky, N. (1965). *Aspect of the Theory of Syntax*. MIT Press, Cambridge.

[Chomsky, 1969a] Chomsky, N. (1969a). *La linguistique cartésienne*. Éditions du Seuil, Paris.

[Chomsky, 1969b] Chomsky, N. (1969b). *Structures syntaxiques*. Éditions du Seuil.

[Chomsky, 1970] Chomsky, N. (1970). *Principles on goverment and binding*. Dordrecht, Netherlands.

[Church and Hanks, 1990] Church, K. and Hanks, P. (1990). Word assosiation norms, mutual information and lexicography. *Computational Linguistic*, 16.

[Collins and Brooks, 1995] Collins, M. and Brooks, J. (1995). Prepositional phrase attachment through a backed-off model. In *Third Workshop on Very Large Corpora*.

[Coulmas, 1989] Coulmas, F. (1989). *The writing systems of the world.* Blackwell.

[Daelemans and Powers, 1992] Daelemans, W. and Powers, D., editors (1992). *Background and experiments in Machine Learning of Natural Language (Proc. 1st Int. SHOE Workshop).* Tilburg University.

[Daille, 1994] Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* PhD thesis, Université de Paris 7.

[Daille et al., 1996] Daille, B., Habert, B., Jacquemin, C., and Royauté, J. (1996). Empirical observation of term variations and principles for their description. *Terminology*, à paraître.

[de Marcken, 1995] de Marcken, C. (1995). The unsupervised acquisition of a lexicon from continous spreech. Technical report, MIT Artificial Intelligence Lab. Memo 1558.

[de Saussure, 1972] de Saussure, F. (1972). *Cours de linguistique générale.* Payot.

[Debili, 1982] Debili, F. (1982). *Analyse syntactico-sémantique fondée sur une acquisition automatique de relations lexicales sémantiques.* PhD thesis, Université de Paris 11 Orsay.

[Decker and Focardi, 1995] Decker, K. M. and Focardi, S. (1995). Technology overview : A report on data mining. Technical report, CSCS-ETH, Swiss Scientific Computer Center.

[Dessen, 1995] Dessen, P. (1995). Les secrets de la séquence. *Biofutur*, 146 :39–43.

[Elman, 1990] Elman, J. (1990). Finding struture in time. *Cognitive Science*, 14 :179–211.

[Finch, 1993] Finch, S. (1993). *Finding structure in Language.* PhD thesis, Center for cognitive Science, University of Edinburgh.

[Finch and Chater, 1992] Finch, S. and Chater, N. (1992). Bootstrappping syntactic categories using statistical methods. In Daelemans, W. and Powers, D., editors, *Background and experiments in machine learning of Natural Language*, pages 229–236, ITK, Tilburg.

[Firth, 1957] Firth, J. C. (1957). *A synopsis of linguistic theory.* Palmer, F.R. (ed) (1968) Selected papers of J.R. Firth 1952-9. Harlow : Longman.

[Fluhr, 1977] Fluhr, C. (1977). *Algorithme à apprentissage et traitement automatique des langues.* PhD thesis, Paris Sud.

[Franova and Kooli, 1998] Franova, M. and Kooli, M. (1998). Recursion manipulation for robotics : Why and how ? In *EMCSR'98.*

[François, 1968] François, F. (1968). *La description linguistique.* Le Langage, André Martinet (éd.),Encyclopédie de la Pléiade. Gallimard.

[Fries, 1952] Fries, C. (1952). *The Structure of English.* London.

[Février, 1948] Février, J. (1948). *Histoire de l'écriture.* Grande Bibliothèque Payot.

[Giguet, 1996] Giguet, E. (1996). The stakes of multilinguality : Multilingual text tokenization in natural language diagnosis. In *Proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence (PRICAI) Workshop "Future issues for Multilingual Text Processing"*, Cairns, Australia.

[Giguet, 1998] Giguet, E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues.* PhD thesis, Université de Caen, section d'informatique.

[Giguet and Vergne, 1997] Giguet, E. and Vergne, J. (1997). From part-of-speech tagging to memory-based deep syntactic analysis. In *Proceedings of the International Workshop on Parsing Technologies (IWPT'97)*, MIT, Boston, Massachussets, USA.

[Greenberg, 1963] Greenberg, J. (1963). *Universals of Language.* Cambridge, MIT.

[Grevisse, 1969] Grevisse, A. (1969). *Précis de grammaire française.* J. Duculot.

[Grevisse, 1986] Grevisse, A. (1986). *Le bon Usage.* Duclot.

[Guilpin and Caron, 1997] Guilpin, T. and Caron, N. (1997). Outil de classification distributionnelle des mots. Projet de licence, Université de Caen, section d'informatique.

[Guiraud, 1968] Guiraud, P. (1968). *Langage et théorie de la communication.* Le Langage, André Martinet (éd.),Encyclopédie de la Pléiade. Gallimard.

[Habert et al., 1997] Habert, B., Nazarenko, A., and Salem, A. (1997). *Les linguistiques de corpus.* Armand Colin.

[Hagège, 1982] Hagège, C. (1982). *La Structure de Langues.* Number 2006 in Que Sais-je ? Presses Universitaires de France.

[Halliday, 1985] Halliday, M. (1985). *An Introduction to Functional Grammar.* Arnold.

[Halliday, 1961] Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word*, 17(3) :241–292.

[Harris, 1946] Harris, Z. (1946). From morpheme to utterance. *Language*, 22 :161–173.

[Harris, 1951] Harris, Z. (1951). *Structural Linguistics.* The University of Chicago Press.

[Harris, 1954] Harris, Z. (1954). Distributional structure. *Word*, 10(2-3) :146–162. Traduction française : Language (20), 1970.

[Harris, 1955] Harris, Z. (1955). From phonemes to morphemes. *Language*, 31(2) :190–222.

[Harris, 1990] Harris, Z. (1990). *Theory of Language and Information : a mathematical approach.* Oxford University Press.

[Hejmslev, 1966] Hejmslev, L. (1966). *Le langage.* Les éditions de Minuit, Paris.

[Hindle and Rooth, 1993] Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1).

[Hockett, 1961] Hockett, C. (1961). Linguistic elements and their relations. *Language*, 37 :29–53.

[Honkela, 1997] Honkela, T. (1997). Comparisons of self-organized word category maps. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Helsinki University of Technology*, pages 298–303.

[Huckle, 1995] Huckle, C. (1995). Grouping word using statistical context. In *EACL-95, student session.*

[Hughes and Atwell, 1994] Hughes, J. and Atwell, E. (1994). The automated evaluation of inferred word. In Cohn, A., editor, *Proceedings of the 11 European Conference on Artificial Intelligence (ECAI-94)*, pages 535–539.

[Hutchens, 1994] Hutchens, J. L. (1994). *Natural Language Grammatical Inference.* PhD thesis, University of Western Australia.

[Hutchens and Alder, 1998] Hutchens, J. L. and Alder, M. D. (1998). Finding structure via compression. In Powers, D. M. W., editor, *Computational Natural Language Learning*, pages 79–82, Adelaide.

[Kazakov, 1997] Kazakov, D. (1997). Unsupervised learning of naïve morphology with genetic algorithms. In *Workshop on Empirical Learning of Natural Language Processing Tasks*, Prague.

[Kiss, 1972] Kiss, G. R. (1972). Grammatical word classes : a learning process and its simulation. *Psychology of learning and motivation*, 7 :1–41.

[Kohonen, 1978] Kohonen, T. (1978). The self-organization map. In *IEEE*, volume 78, pages 1464–1480.

[Longacre, 1960] Longacre, R. (1960). String constituent analysis. *Language*, 36(1) :63–88.

[Longacre, 1964] Longacre, R. (1964). *Grammar discovery procedures : A field manual.* The Hague, Mounton and Company.

[Lucas, 1995] Lucas, N. (1995). Le style scientifique en japonais et en français. In Beillevaire, P. and Gossot, A., editors, *Japon pluriel, Acte du premier colloque de la société française des études japonaises*, pages 393–402. Éditions Phillipe Picquier.

[Lyons, 1969] Lyons, J. (1969). *Introduction to Theoretical Linguistics.* Cambridge University Press.

[Magerman, 1991] Magerman, D. (1991). Mutual information, deducing linguistic structure. In Powers, D. and Reeker, L., editors, *Machine Learning of Natural Language and Ontology.*

[Mahmoudian, 1981] Mahmoudian, M. (1981). *La Linguistique.* Paris : Seghers.

[Mahon and Smith, 1996] Mahon, J. M. and Smith, F. (1996). Improving statistical language model performance with automaticcaly generated word hierarchies. *Computational Linguistics*, 22(2) :217–247.

[Malherbe, 1995] Malherbe, M. (1995). *Les langages de l'humanité.* Robert Lafon.

[Mandelbrot, 1968] Mandelbrot, B. (1968). *Les constantes chiffrées du discours.* Le Langage, André Martinet (éd.),Encyclopédie de la Pléiade. Gallimard.

[Marcus, 1991] Marcus, M. (1991). The automatic acquisition of linguistic structure from large corpora : An overview of work at the university of pennsylvania. In *AAAI Spring Symposium*.

[Martinet, 1970] Martinet, A. (1970). *Élements de linguistique générale*. Armand colin.

[Mel'čuk, 1987] Mel'čuk, I. (1987). *Dependency syntax, theory and practice*. Albany : Suny Press.

[Miclet and de la Higuera, 1996] Miclet, L. and de la Higuera, C., editors (1996). *Grammatical Inference : Learning Syntax from sentences*, volume 1147 of *Lecture Notes in Artificial Intelligence*. Springer Verlag.

[Morel and Lacheret-Dujour, 1998] Morel, M. and Lacheret-Dujour, A. (1998). Utilisation d'une structure arborescente pour une hiérarchisation fine des règles de transcription graphème-phonème. In *Actes des XXIIèmes journées d'études sur la parole*.

[Nevin, 1993] Nevin, B. E. (1993). A minimalist program for linguistics. a perspective on the work of zellig harris. *Historiographia Linguistica*, 20(2/3) :355–398.

[Pereira et al., 1993] Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *EACL93*.

[Peyo, 1959] Peyo (1959). *La flûte à six schtroumpfs*. Dupuis.

[Pike, 1967] Pike, K. (1967). *Language in relation to a unified theory of the structure of human behavior*. Mouton & Co, The Hague - Paris.

[Pinker, 1984] Pinker, S. (1984). *Language Learniability and Language Development*. Harvard University Press, Cambridge, Massachusetts.

[Pitman, 1948] Pitman, R. S. (1948). Nuclear structures in linguistics. *Language*, 24(3) :287–292.

[Ploux and Victorri, 1998] Ploux, S. and Victorri, B. (1998). Constructions d'espaces sémantiques à l'aide de dictinnaires de synonymes. *Traitement automatique des langues*, 39(1) :161–182.

[Powers, 1998] Powers, D. M. W., editor (1998). *New Methods in Language Processing and computational Natural Language Learning*, Macquarie University.

[Powers and Daelemans, 1992] Powers, D. M. W. and Daelemans, W. (1992). Shoe : The extraction of hierarchical structure for machine learning of natural language. project summary. In Daelemans, W. and Powers, D., editors, *Background and experiments in machine learning of Natural Language*, pages 125–159, ITK, Tilburg.

[Ramat, 1985] Ramat, P. (1985). *Typologie Linguistique*. Presse Universitaire de France.

[Ramshaw and Marcus, 1995] Ramshaw, L. A. and Marcus, M. P. (1995). Text chunking using transformation-based learning. In *ACL Third Workshop on Very Large Corpora*, pages 82–94.

[Redington et al., 1996] Redington, M., Chater, N., and Finch, S. (1996). Distributional information and the acquisition of linguistics categories : A statistical approach. In *Fifteenth Annual Conference of the Cognitive Science Society*, pages 848–853, Hillsdale, NJ : Erlbaum.

[Rosmorduc, 1994] Rosmorduc, S. (1994). *Analyse morpho-syntaxique de textes non ponctués, application aux textes hiéroglyphiques.* PhD thesis, École normale supérieure de Cachan.

[Sapir, 1921] Sapir, E. (1921). *Language, an introduction to the study of speech.* New York.

[Schütze, 1993] Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the american for Computational Linguistics*, volume 31, pages 251–258.

[Schütze, 1995] Schütze, H. (1995). Distributional part-of-speech tagging. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 141–148, Dublin.

[SciencesAvenir, 1998] SciencesAvenir (1998). Le dernier méson. *Sciences et Avenir*, 616 :20–21.

[Smadja, 1993] Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistic*, 19(1) :143–177.

[Sokal and Sneath, 1963] Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* San Francisco : W. H. Freeman.

[Sproat et al., 1994] Sproat, R., Shih, C., Gale, W., and Chang, N. (1994). A stochastic finite-state word-segmentation algorithm for chinese. In *Proceedings of ACL-94.*

[Stolcke and Omohundro, 1994] Stolcke, A. and Omohundro, S. M. (1994). Best-first model merging for hidden markov model induction.

[Stolcke and Shriberg, 1996] Stolcke, A. and Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. Technical report, Speech Technology and Research Laboratory.

[Tesnière, 1959] Tesnière, L. (1959). *Éléments de syntaxe structurale.* Klincksieck,Paris.

[van den Bosch et al., 1996] van den Bosch, A., Daelemans, W., and Weijters, T. (1996). Morphological analysis as classification : an inductive approach. In *NEMLAP'96, Ankara.*

[Vendryes, 1923] Vendryes, J. (1923). *Le Langage : introduction Linguistique à l'Histoire.* Albin Michel, l'evolution de l'humanité edition.

[Vergne, 1999] Vergne, J. (1999). Entre arbre de dépendance et ordre linéaire, les deux processus de transformation. *Les cahiers de grammaires*, à paraître.

[Vergne and Giguet, 1998] Vergne, J. and Giguet, E. (1998). Regards théoriques sur le "tagging". In *proceedings of the fifth annual conference Le Traitement Automatique des Langues Naturelles (TALN 1998)*, Paris, France.

[Wanner and Gleitman, 1982] Wanner, E. and Gleitman, L. (1982). *Language Acquisition : The State of the Art.* Cambridge University Press.

[Wolff, 1977] Wolff, G. (1977). The discovery of segments in natural language. *British Journal of Psychology*, 68 :97–106.

[Wolff, 1980] Wolff, G. (1980). Language acquisition and the discovery of phrase structure. *Language and Speech*, 23(3) :255–269.

[Woodley, 1995] Woodley, M. C. P. (1995). Quels corpus pour quels traitements automatiques ? *TAL*, 36 :213–232.

[Zhang, 1996] Zhang, M. (1996). A faster structured tag word classification method. In *PRICAI-96 Workshop on Future Issues for Multi-lingual Text Processing, Cairns, Australia.*

[Zipf, 1949] Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort : An Introduction to Human Ecology.* AW.

[Zuret, 1998] Zuret, D. (1998). *Discovery of Linguistic Relations Using Lexical Attraction.* PhD thesis, MIT, Cambridge.

Concepts and Algorithms
to discover Formal Structures in Natural Languages

Abstract

This presentation describes a method which allows the uncovering of syntactic structures from untagged corpora (no lexicon, just raw text). It can be considered as a continuation of Zellig Harris distributional work developed in the 50'. Following the distributional hypothesis, only formal criteria are used (no resort to semantics).

The method is based on a simple idea of the language : it is a linear object in which the boundaries (beginning and ending) of the different structures are marked by characteristic elements. The structures so delimited are the simple phrase (non recursive) and the clause, which are both multilingually and formally defined. The phrase Boundaries Indicator (BI) corresponds to morphemes (linked or free), and the clause BI to morphemes and phrases.

From this theoretical structure, we extract the list of all the categories an element can belong to (beginning and ending BI of phrases and clauses). Once structures and categories are identified, we build specified contexts for each category in order to classify all the words of the texts. These contexts are built thanks to prototypical elements which are easily identified from formal criteria (their identification relies on their behaviour related to punctuation marks). We can thus classify a word into several categories. The categorization first deals with clause elements (such as conjunctions, verbal phrases), and then with nominal phrases.

This method allows word categorization and a segmentation of the corpus into phrases. These concepts and algorithms were partially tested on several natural languages such as French, German, Turkish, Vietnamese, Swahili.

Keywords : Machine Learning, Natural Language Processing, Distributionalism, Clustering, Multilinguism.

Concepts et algorithmes
pour la découverte des structures formelles des langues

Résumé

Que peut-on apprendre sur la structure d'une langue à partir d'un texte écrit dans cette langue, et ceci sans connaissance particulière sur celle-ci et avec l'aide (disons l'utilisation) d'un ordinateur ? Voilà la question à laquelle nous avons essayé de répondre. Cette réponse peut être vue comme une continuation des travaux en analyse distributionnelle développée dans Zellig Harris. L'objectif de ce travail est donc de découvrir les structures formelles d'une langue en étudiant ces régularités formelles contenues dans un corpus

Notre méthode de découverte se base sur une simple conception formelle de la langue : un objet linéaire dans lequel les frontières (de début et de fin) des différentes structures sont indiquées par des éléments caractéristiques. Les structures ainsi identifiées sont le syntagme simple (non récursif), et la proposition, structures à la fois multilingues et formelles. Ces indicateurs de frontières correspondent à des morphèmes (libres ou liés) pour le syntagme, et à des morphèmes ou des syntagmes pour la proposition.

À partir de ces structures théoriques, nous construisons la liste de toutes les catégories qu'un élément (morphème ou mot) peut prendre. Une fois ces structures et catégories recensées, nous construisons des contextes spécifiques à chaque catégorie afin de catégoriser les éléments du texte. Nous obtenons donc un mécanisme permettant d'assigner à un élément plusieurs catégories si cet élément apparaît dans différents contextes. Ces contextes sont construits à l'aide des éléments prototypiques de marqueurs de frontières de structures, identifiables grâce à leur position par rapport à la segmentation physique du texte (en particulier les ponctuations).

Les résultats obtenus permettent la catégorisation des mots du corpus, ainsi qu'une segmentation partielle en syntagmes. La méthode a été appliquée à une dizaine de langues comme le français, l'allemand, le turc, le vietnamien et le swahili.

Mots-clés : Apprentissage automatique, langues naturelles, distributionalisme, catégorisation (linguistique), multilinguisme.

Discipline : Informatique

GREYC CNRS UPRESA 6072
Groupe de Recherche en Informatique, Image, et Instrumentation de Caen
Université de Caen Basse-Normandie
Campus II F-14032 Caen Cedex