Regards Théoriques sur le "Tagging"

Jacques Vergne et Emmanuel Giguet

Jacques.Vergne@info.unicaen.fr, Emmanuel.Giguet@info.unicaen.fr
GREYC - CNRS UPRESA 6072 - Université de Caen F-14032 Caen cedex France

Résumé

Dans cet article, nous proposons de prendre du recul par rapport à l'aspect opératoire du tagging, et nous tentons de montrer que le tagging ouvre la voie au renouveau de l'analyse syntaxique en la fondant sur l'explicitation des processus : processus de déduction locale dans les syntagmes non récursifs, et processus de mise en relation des syntagmes non récursifs, étendant ainsi à l'analyse syntaxique les propriétés calculatoires du tagging.

Introduction

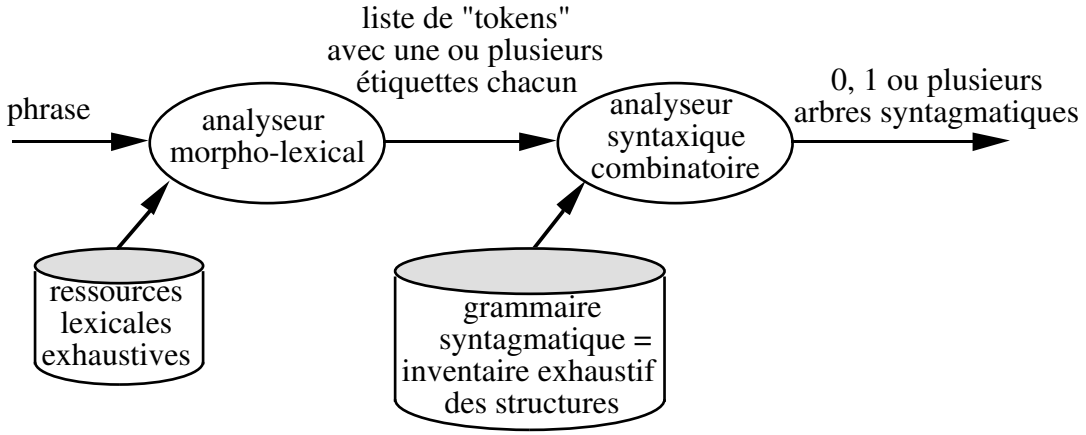
Le "tagging", ou "étiquetage", ou "marquage", consiste à affecter une "étiquette" ("tag", ou catégorie) à chaque "mot" d'un texte. Nous proposons dans cet article de prendre du recul par rapport à l'aspect opératoire de cet outil devenu d'usage courant, et de poser des questions plus fondamentales sur ses rapports avec l'analyse syntaxique automatique traditionnelle (section 1), sur le champ d'action théorique des déductions contextuelles (section 2), sur le rôle des ressources lexicales (section 3), sur la définition du jeu d'étiquettes (section 4) et sur le renouveau qu'il pourrait apporter en analyse syntaxique (section 5).

1. Tagging et analyse syntaxique traditionnelle

1.1. Deux opérations complémentaires dans une chaîne de traitement

Par "analyse syntaxique traditionnelle", nous entendons l'analyse syntaxique dans son état canonique telle que la décrit Gérard Sabah dans le chapitre 2 de (Sabah 1989), pages 37 à 71 : "Les principes de l'analyse syntaxique des phrases".

Figure 1 : Chaîne de traitement de l'analyse syntaxique traditionnelle

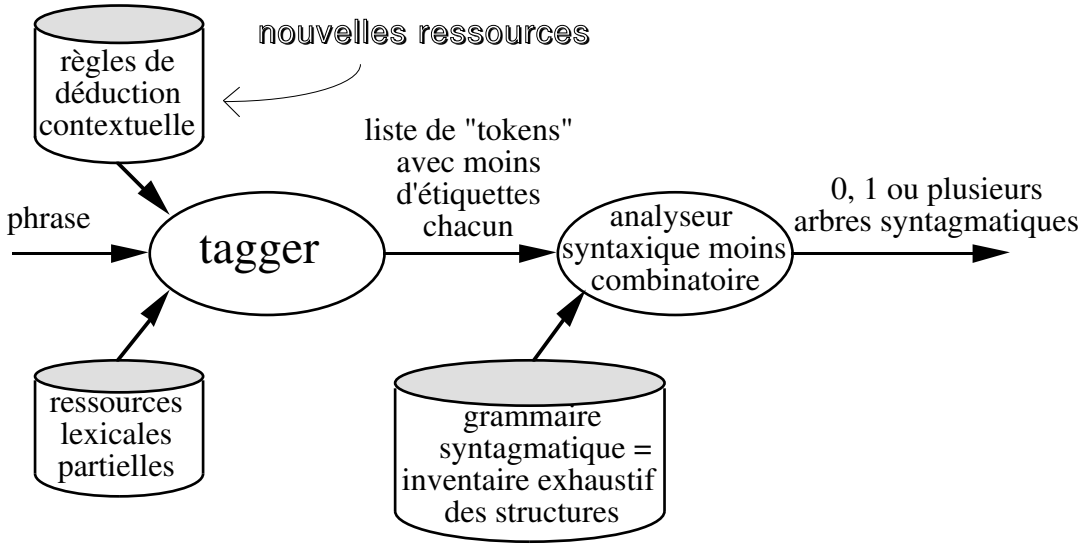


La chaîne de traitement de l'analyse syntaxique traditionnelle (voir figure 1 ci-dessus) comprend au minimum deux modules : l'analyseur morpho-lexical, qui, à partir de ressources lexicales considérées comme exhaustives, produit une liste de "tokens" ("mots" arbitraires) munis chacun d'une ou plusieurs étiquettes, suivi de l'analyseur syntaxique proprement dit, qui, pour chaque phrase, produit zéro, un ou plusieurs arbres syntagmatiques.

Le problème central de l'analyse syntaxique traditionnelle est son aspect combinatoire : l'analyseur doit choisir une étiquette pour chaque "token", et l'ensemble des étiquettes choisies (par les substitutions lexicales) doit permettre d'attribuer à la phrase une structure syntaxique attendue, reconnue dans l'inventaire exhaustif des structures attendues, codé sous la forme d'une grammaire syntagmatique. La cause profonde de cet aspect combinatoire est que l'analyse syntaxique traditionnelle est fondée sur les principes de la compilation, analyse d'un langage formel dont le lexique est clos (les mots ayant une seule catégorie) et dont la syntaxe est exhaustivement définie par une grammaire formelle, alors qu'une langue a un lexique ouvert (les mots ayant plusieurs catégories) et une syntaxe partiellement définie.

Dans la nouvelle chaîne de traitement, le tagger vient se substituer à l'analyseur morpho-lexical, pour fournir à l'analyseur syntaxique une liste de "tokens" munis chacun d'une seule étiquette (ou le moins possible), et ainsi annuler, ou réduire le plus possible, la combinatoire des catégories possibles, par l'apport de nouvelles ressources : les déductions contextuelles (voir figure 2 ci-dessous).

Figure 2 : Nouvelle chaîne de traitement de l'analyse syntaxique traditionnelle



1.2. Deux démarches opposées vis-à-vis des structures et des processus

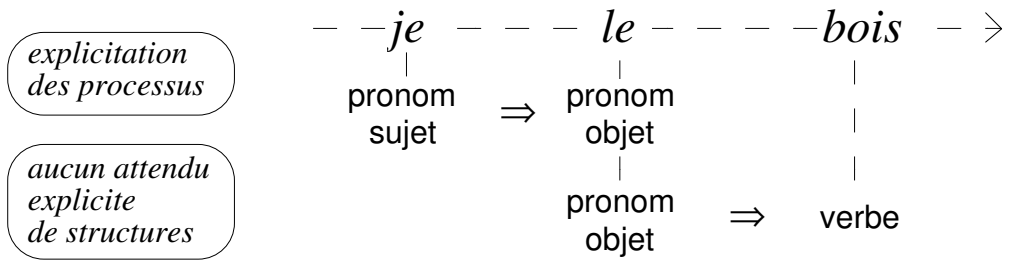
Résumons ces oppositions dans le tableau suivant :

	analyse syntaxique traditionnelle	*tagging*
structures	explicitées exhaustivement	non explicitées
processus	analogue à la compilation	explicité par les déductions contextuelles

Dans l'analyse syntaxique traditionnelle, l'ensemble des structures syntaxiques attendues est exhaustivement explicité sous la forme d'une grammaire syntagmatique.

À l'opposé, le tagging privilégie le processus par rapport aux structures. Le processus du tagging est une propagation de déductions contextuelles sur les tokens (voir figure 3 ci-dessous). Le processus est conduit par les fréquences des contiguïtés des catégories dans un tagger statistique, ou explicité par des règles contextuelles symboliques dans un tagger symbolique, mais aucune structure n'est attendue explicitement. L'algorithme d'un tagger consiste à tester sur chaque token l'applicabilité d'un nombre constant de règles, et ceci lui confère une complexité linéaire en temps par rapport au nombre de tokens de la phrase. Dans un renversement complet par rapport à l'analyse syntaxique traditionnelle, le tagging explicite le processus mais n'explicite pas les structures.

Figure 3 : Processus de propagation de déductions contextuelles dans le tagging



Du point de vue de son processus, un analyseur traditionnel intègre toutes les caractéristiques du modèle de la compilation, à la différence non négligeable qu'un compilateur analyse un langage de programmation, exhaustivement connu et modélisé, et qu'un analyseur syntaxique de langue analyse une langue, connue partiellement, et donc partiellement modélisée. En conséquence, alors que le processus d'un compilateur est déterministe, celui d'un analyseur syntaxique traditionnel de langue est non déterministe, car l'absence de critères locaux pour prendre des décisions locales implique des retours en arrière aux tokens précédents (voir les "points d'embarras" de Sabah 1989, page 58). Les algorithmes sont de complexité polynomiale avec des "formalismes simples", ou bien le problème devient NP-complet "pour les formalismes plus évolués, comme les grammaires d'unification" (voir Abeillé et Blache 1997, pages 79 et 80).

2. Les déductions contextuelles dans le tagging : leur champ d'action théorique

Tentons de circonscrire a priori le champ d'action des déductions contextuelles.

Définissons un segment intermédiaire entre les mots et les phrases : le syntagme non récursif (SNR dans la suite de cet article), ou syntagme simple, ou syntagme noyau, ou syntagme sans ses syntagmes compléments, ou "core phrase", ou "chunk" dans la littérature en anglais (voir Abney 1996). Ce segment est stable entre des langues différentes, et trouve approximativement son équivalent à l'oral sous la forme du groupe accentuel.

On a donc une hiérarchie de segments à trois niveaux : mots, SNR, phrases, hiérarchie où le segment d'un niveau est constitué de segments du niveau inférieur : un tout est d'un type différent du type de ses parties, contrairement au syntagme récursif, qui est constitué de mots et de syntagmes récursifs (certaines parties sont de même type que le tout).

Les mots dans un SNR, et les SNR dans une phrase ont des comportements très différents : les mots d'un SNR forment un agrégat très contraint autour d'un nom ou d'un verbe, une structure stable et explicitable exhaustivement, mais les SNR dans une phrase sont soumis à des contraintes plus relâchées, et forment des structures instables, dont l'explicitation exhaustive nous semble actuellement un objectif hors de portée.

À titre d'illustration, dans ses vers bien connus du Bourgeois Gentilhomme, Molière fait permuter les SNR dans la phrase, mais laisse inchangé l'ordre des mots dans les SNR :

Belle marquise , vos beaux yeux me font mourir d'amour .
D'amour mourir me font , belle marquise , vos beaux yeux .
Vos beaux yeux d'amour me font , belle marquise , mourir .

On ne peut alors considérer une phrase comme une suite continue, indifférenciée de mots, dans laquelle toutes les contiguïtés seraient équivalentes. On doit considérer de manière différente une contiguïté de deux tokens à l'intérieur d'un SNR, et une contiguïté de deux tokens à la frontière de deux SNR contigus. Une déduction contextuelle sûre devra donc s'appuyer sur la structure interne stable du SNR et être interne au SNR, et non entre le dernier token d'un SNR et le premier du SNR suivant.

Comme un SNR est constitué d'un élément central lexical (nom ou verbe), entouré d'éléments périphériques grammaticaux (prépositions, déterminants, clitiques, adverbes, ...), la déduction contextuelle canonique consiste en une déduction de la forme :

étiquette d'un mot grammatical ⇒ étiquette du mot lexical contigu

Plus précisément, un mot grammatical marque le plus souvent (en français) le début d'un SNR ainsi que le type de ce SNR, ce qui signifie que la déduction est indirecte car elle passe par le type du SNR :

étiquette d'un mot grammatical ⇒ type du SNR ⇒ étiquette du mot lexical contigu
ou : étiquette d'un mot grammatical ⇒ type du SNR ⇒ étiquette du mot gramm. contigu

Exemples :

une ferme	*une* > déterminant	⇒ *une ferme* SNR nominal	⇒ *ferme* > nom
ne ferme	*ne* > négation	⇒ *ne ferme* SNR verbal	⇒ *ferme* > verbe
je le	*je* > pronom sujet	⇒ *je le* ... SNR verbal	⇒ *le* > pronom objet
le bois	*le* > pronom objet	⇒ ... *le bois* SNR verbal	⇒ *bois* > verbe

On peut alors définir un cas où il est impossible de faire une déduction locale et interne au SNR : c'est le cas d'un SNR constitué d'un seul mot lexical. Cette déduction locale impossible ne porte pas à conséquence si ce mot n'a qu'une seule catégorie possible.

Dans l'exemple suivant, ***remporte*** constitue à lui seul un SNR, et ne peut être que verbe :

*Comme prévu, M. Museveni **remporte** la quasi-totalité des votes dans l'Ouest, ...*

Mais, si ce mot a plusieurs catégories possibles, alors la décision ne peut pas être prise par déduction contextuelle locale : la décision ne pourra être prise que par la mise en relation du SNR de type multiple avec un SNR de type connu. Dans le corpus qui a servi à l'évaluation des essais de l'action GRACE (action internationale d'évaluation comparative des "taggers" du français [1]) en septembre 1996 (environ 10 000 mots du journal Le Monde, fichier "lemon06"), environ 1% des mots sont dans ce dernier cas (1/6 sont des noms, 1/6 sont des verbes, et 2/3 sont des adjectifs ou des participes passés en position adjectivale).

Dans l'exemple suivant, ***montre*** constitue à lui seul un SNR, et peut être verbe ou nom ; la décision ne peut pas être prise localement, mais par la mise en relation avec son sujet ***présence***:

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

Cette déduction est du type :

présence sujet potentiel de *montre* si *montre* est un SNR verbal
⇒ *montre* SNR verbal ⇒ *montre* verbe

En conclusion de cette section, l'étiquetage des mots qui forment à eux seuls un SNR de type multiple est impossible par déduction contextuelle (il devra donc être confié à une étape ultérieure de l'analyse). Notons en outre que tout étiquetage fondé sur une relation entre deux SNR est aussi impossible par déduction contextuelle ; citons par exemple : la résolution de *de d' du des* préposition ou partitif-article, la résolution de *que qu'* conjonction ou pronom relatif, les genre et nombre des pronoms relatifs selon leur antécédent, les genre, nombre, cas, type des clitiques réfléchis et des *nous* et *vous*, toute propagation de genre, nombre, personne par accord entre SNR (sujet - verbe, antécédent - pronom relatif, ...).

3. Les ressources lexicales dans le tagging

Dans les sections précédentes, nous avons focalisé notre attention sur les déductions contextuelles. Mais ces déductions s'appuient sur des ressources lexicales, et doivent s'articuler avec elles. Étudions de quelle manière.

3.1. Trois articulations possibles entre ressources lexicales et déductions contextuelles

• Voyons comment le problème du tagging est habituellement posé : pour chaque token, toutes ses catégories sont exhaustivement énumérées à partir de sources d'informations lexicales

[1] http://www.ciril.fr/~pap/grace.html

(lexique de lemmes ou de formes, reconnaisseur de formes verbales, règles sur les finales - ou "guesser" - pour les mots absents du lexique), et le tagger doit choisir parmi les catégories possibles (comme l'homographie polycatégorielle est couramment appelée "ambiguïté", ce processus de choix est souvent appelé "désambiguïsation", vue comme une annulation ou une diminution de l'"ambiguïté").

Une première solution pour effectuer un choix est d'attribuer au contexte le rôle de supprimer des catégories possibles de cette liste (c'est le cas des grammaires de contraintes - voir Voutilainen 1994). Nous appellerons une telle déduction : "déduction négative", qui correspond à une règle de la forme :

dans tel contexte, tel token **ne peut pas** avoir telles catégories

Le principal défaut d'une déduction négative est que la catégorie attendue du token doit appartenir à la liste des catégories possibles, ce qui revient à faire l'hypothèse que tout token appartient au lexique et est muni de la liste exhaustive de ses catégories possibles, hypothèse manifestement fausse pour les mots lexicaux, même avec un "guesser".

• Une autre solution pour choisir est d'utiliser des "déductions affirmatives", de la forme :

dans tel contexte, tel token **a** telle catégorie

Cette solution a l'avantage de pallier cette double incomplétude des ressources lexicales : certains tokens n'y sont pas, et certains tokens présents n'ont pas toutes leurs catégories possibles.

Par exemple, dans *je positive*, *positive* est pré-étiqueté adjectif dans le lexique, et la déduction est la suivante :

je positive *je* > pronom sujet ⇒ *je positive* SNR verbal ⇒ *positive* > verbe

• Mais il y a encore une autre manière de poser le problème : on remarque que les différentes catégories possibles d'un token sont loin d'être équiprobables : en général une catégorie est de beaucoup la plus fréquente. Prenons l'exemple caricatural de *le l' la les* dans ce même corpus de 10 687 mots [2] du journal Le Monde : 1054 occurrences qui se répartissent en 1029 déterminants (97,6%), et 25 pronoms clitiques objet (2,4%).

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

pronom clitique sujet suivi de *le l' la les* ⇒ *le l' la les* pronoms clitiques objet
négation *ne* suivie de *le l' la les* ⇒ *le l' la les* pronoms clitiques objet
le l' la les suivi d'un verbe sûr ⇒ *le l' la les* pronoms clitiques objet

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

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

3.2. Types d'homographies polycatégorielles, et étude statistique

Étudions les principaux types d'homographies, et faisons-en une étude statistique sur ce même corpus du journal Le Monde (de 10 687 mots), corpus pour lequel nous disposons d'un étiquetage manuel réalisé par des linguistes du comité de coordination de l'action GRACE,

---

[2] Ce comptage a été effectué automatiquement dans un éditeur (BBEdit sur Mac) : l'apostrophe et le tiret sont des séparateurs, la ponctuation et les guillemets ne sont pas comptés comme mots.

selon des spécifications proposées par le comité de coordination, puis discutées durant l'adjudication des essais. En consultant automatiquement des ressources lexicales incluant les homographies, nous obtenons les résultats suivants, en mettant en évidence l'alternative la plus rare :

- homographies sur les catégories de mots grammaticaux :

2,4% de pronoms	sur	1054	*le l' la les* homographes déterminant / pronom
8,5% de partitifs-articles	sur	1088	*de d' du des* homographes préposition / partitif-article
16,2% de pronoms relatifs	sur	130	*que qu'* homographes conjonction / pronom relatif
4,7% de noms	sur	171	*est être avoir* homographes verbe auxiliaire / nom
17,1% de noms	sur	111	*bien mal moins plus* ... homographes adverbe / nom

- homographies sur les catégories de mots lexicaux :

16,4% de verbes	sur	487	homographes nom / verbe non auxiliaire
33,8% de noms	sur	714	homographes adjectif / nom

(ces derniers concernent les absents du lexique, normalement étiquetés par notre analyseur adjectif ou nom, par déduction locale dans le SNR nominal ou verbal)

- homographies sur les attributs de noms, adjectifs, pronoms, verbes :

24,6% de féminins	sur	2475	homographes masculin / féminin
25,7% de pluriels	sur	501	homographes singulier / pluriel
3,8% de subjonctifs	sur	185	homographes indicatif / subjonctif

Cette étude confirme qu'une des alternatives est toujours beaucoup plus fréquente.

3.3. Une expérience d'évaluation de l'étiquetage par défaut

Pour évaluer l'intérêt et l'importance des étiquettes par défaut, codées dans les ressources lexicales, par rapport aux déductions contextuelles, faisons l'expérience d'étiqueter uniquement avec les étiquettes par défaut des ressources lexicales, sans **aucune** déduction contextuelle locale. Comme valeurs par défaut, prenons systématiquement l'alternative la plus fréquente :

- homographies sur les catégories de mots grammaticaux :

déterminant	pour *le l' la les* homographes déterminant / pronom
préposition	pour *de d' du des* homographes préposition / partitif-article
conjonction	pour *que qu'* homographes conjonction / pronom relatif
verbe	pour *est être avoir* homographes verbe auxiliaire / nom
adverbe	pour les homographes adverbe / nom

- homographies sur les catégories de mots lexicaux :

nom	pour les homographes nom / verbe non auxiliaire
adjectif	pour les homographes adjectif / nom

- homographies sur les attributs de noms, adjectifs, pronoms, verbes :

masculin	pour les homographes masculin / féminin
singulier	pour les homographes singulier / pluriel
indicatif	pour les homographes indicatif / subjonctif

L'expérience porte sur ce même corpus de 10 687 mots du journal Le Monde des essais de l'action GRACE, étiqueté à la main par des linguistes du comité GRACE. Nous pouvons alors comparer un étiquetage automatique avec l'étiquetage manuel [3].

Le jeu d'étiquettes GRACE est dérivé du jeu MULTEXT [4] ; il comprend 11 catégories de base, avec 0 à 6 attributs ayant de 2 à 8 valeurs, ce qui donne 311 étiquettes différentes. La tokenisation GRACE est très fine : tout mot contenant une apostrophe ou un tiret est subdivisé, et apostrophes et tirets sont des tokens (respectivement 4,8% et 1,0% des tokens), comme le reste de la ponctuation (10,5% des tokens), ce qui donne 12741 "tokens" pour 10 687 "mots".

---

[3] Pour les mettre en conformité avec les spécifications de l'étiquetage GRACE résultant de l'adjudication des essais, nous avons modifié environ 2% des tokens du corpus de référence étiqueté à la main.

[4] http://www.lpl.univ-aix.fr/projects/multext

Dans ces conditions expérimentales, et en reproduisant aussi fidèlement que possible le protocole d'évaluation GRACE, on obtient alors les résultats suivants :

aucune étiquette multiple (décision = 1,0000)

92,1% des tokens ont la même catégorie de base que dans le corpus de référence
(précision sur les catégories de base = 0,9208)

82,5% des tokens ont exactement la même étiquette que dans le corpus de référence
(précision sur les étiquettes complètes = 0,8254)

En conclusion de notre expérience, nous constatons que les ressources lexicales avec des **valeurs par défaut**, sans aucune déduction contextuelle, permettent d'obtenir 9 catégories de base sur 10 égales à celles du corpus de référence, sur le jeu de 11 catégories de base, et 8 étiquettes complètes sur 10 égales à celles du corpus de référence, sur le jeu de 311 étiquettes complètes (notons qu'il est important de préciser avec quelle tokenisation et avec quel jeu d'étiquettes on obtient telle précision). Avec les seules ressources lexicales, l'étiquetage automatique a ainsi parcouru la majeure partie du chemin vers l'étiquetage manuel.

Nous faisons l'hypothèse que la valeur par défaut (venant du lexique), jusqu'à preuve du contraire (apportée par le contexte), est une propriété générale des langues, permettant une économie dans les processus de communication langagiers.

3.4. Contributions des déductions contextuelles et des mises en relation

En introduisant successivement les déductions contextuelles, puis les mises en relations des SNR, voici un tableau résumant les valeurs obtenues en précision (la décision reste égale à 1,0000) :

ressources	précision sur les 11 catégories de base		précision sur les 311 étiquettes complètes	
lexique avec valeurs par défaut	92,1%	+92,1%	82,5%	**+82,5%**
+ déductions contextuelles : absolu	98,0%		94,8%	
relatif		+5,9%		**+12,3%**
+ mises en relation : absolu	99,3%		97,1%	
relatif		+1,3%		**+2,3%**

En comparant les contributions des trois types de ressources, on observe la répartition suivante (en précision sur les étiquettes complètes) :

- le lexique avec des valeurs par défaut résout 82,5% des tokens,
- la déduction contextuelle améliore la précision pour 12,3% des tokens, ce qui fait plafonner la précision à 94,8%,
- et enfin la mise en relation des SNR permet de résoudre 2,3% des tokens, qui sont environ pour moitié les mots qui forment à eux seuls un SNR de type multiple (ceux évoqués ci-dessus en section 2), et pour moitié certains mots grammaticaux étiquetés par la mise en relation : résolution de *de d' du des* préposition ou partitif-article, résolution de *que qu'* conjonction ou pronom relatif, genre et nombre des pronoms relatifs selon leur antécédent, genre, nombre, cas, type des clitiques réfléchis et des *nous* et *vous*.

L'écart de 3% en précision sur les étiquettes complètes qui sépare les 97,1% des 100% est dû pour moitié à des erreurs de notre analyseur, et pour moitié à un écart incompressible dû aux faits que l'étiquetage manuel ne peut être d'une régularité parfaite et que les spécifications aussi précises soient-elles ne peuvent pas prévoir tous les cas réels.

Remarque sur les valeurs des précisions données ci-dessus : ces valeurs ne peuvent pas être comparées avec les valeurs obtenues par d'autres taggers dans d'autres conditions expérimentales : tokenisation, jeu d'étiquettes (nombre d'étiquettes, jeu de type traditionnel ou distributionnel), corpus, spécifications d'étiquetage du corpus de référence, protocole et métrique de calcul des écarts entre résultats calculés et corpus de référence. L'intérêt de ces valeurs est surtout de permettre ici la comparaison relative entre les apports des trois types de

ressources dans la précision du tagging. La comparaison des performances de taggers différents ne peut se faire avec rigueur que s'ils sont comparés dans le cadre de conditions expérimentales strictement identiques, comme dans l'action GRACE par exemple. Remarquons l'importance du nombre d'étiquettes dans la comparaison des précisions de différents taggers : un petit nombre d'étiquettes rend la décision plus facile (la probabilité est plus grande de tomber par hasard sur la bonne étiquette), mais jusqu'à un certain point, car un trop petit nombre d'étiquettes rendrait les déductions contextuelles locales plus difficiles par manque de finesse et de régularité dans la description des contextes.

4. Définition du "token", définition du "tagset"

Le mot, qui exprime une segmentation conventionnelle et instable entre des langues différentes, et évolutive à l'intérieur d'une langue, n'est pas le bon candidat pour être systématiquement le "token", segment de phrase, unité traitée dans un traitement de langue : un mot peut être divisé en plusieurs tokens (cas des amalgames, des mots composés), ou plusieurs mots peuvent être regroupés en un seul token : locutions diverses, noms propres composés, numéraux composés, mots composés. Le token est défini en rapport avec l'objectif du traitement.

Le "tagset", ou jeu d'étiquettes des tokens, doit-il, par fidélité aux traditions, reprendre les parties du discours? Rappelons-nous que les parties du discours constituent une taxinomie traditionnelle et empirique des "mots", consacrée surtout à l'enseignement, une théorie parmi de nombreuses autres possibles (voir un exemple de théorie originale dans Tesnière 1959, pages 51 à 94), et que nous avons toute liberté d'en créer une autre, adéquate au tagging. Posons-nous alors la question : quelles seraient les caractéristiques d'un jeu d'étiquettes adéquat aux déductions contextuelles, déductions sur les régularités des contiguïtés des catégories, sur les **régularités distributionnelles**, captées à travers le filtre du jeu d'étiquettes.

Par exemple, ces trois étiquettes traditionnelles caractérisent un SNR nominal :

"une" : article "cette" : adjectif démonstratif "sa" : adjectif possessif

Elles sont maintenant plus souvent regroupées sous une étiquette unique : le déterminant, étiquette fondée sur la régularité de la classe distributionnelle des déterminants, différenciée de celle des adjectifs.

En revanche, "il", "elle", "nous", "y", "ceux", "qui", "dont" sont tous des pronoms, mais ils ont des distributions bien différentes !

Donc, pour capter (puis utiliser) des régularités distributionnelles, chaque étiquette du jeu d'étiquettes doit définir une classe distributionnelle de tokens. Par exemple, parmi les adjectifs, on devra différencier les épithètes antéposées, les épithètes postposées, et les attributs, car leurs distributions sont différentes.

Un jeu d'étiquettes distributionnel, associé au concept de SNR nominal ou verbal, implique que ce jeu est partitionné en deux sous-ensembles (leur intersection est vide) : le jeu des étiquettes dans le SNR nominal et le jeu des étiquettes dans le SNR verbal. Par exemple, l'étiquette de l'adjectif épithète, dans le SNR nominal, est différente de celle de l'adjectif attribut, dans le SNR verbal.

5. Les concepts du tagging : un renouveau pour l'analyse syntaxique

Comme nous l'avons vu en section 1.2., l'apport conceptuel principal et original du tagging est l'explicitation des processus, associé à l'abandon de l'explicitation des structures. Les processus explicités sont des déductions contextuelles.

Nous y ajoutons le concept central de syntagme non récursif (défini en section 2), des ressources lexicales avec des valeurs par défaut (étudiées en section 3) généralisées pour l'ensemble des mots grammaticaux, et un jeu d'étiquettes distributionnel (présenté en section 4). Le tagger, devenu moteur de déduction contextuelle, produit en sortie, non seulement une liste de tokens étiquetés, mais aussi une liste de syntagmes non récursifs délimités et typés. Environ 1% des tokens ont à la fois les deux propriétés suivantes : ils ont plusieurs catégories possibles

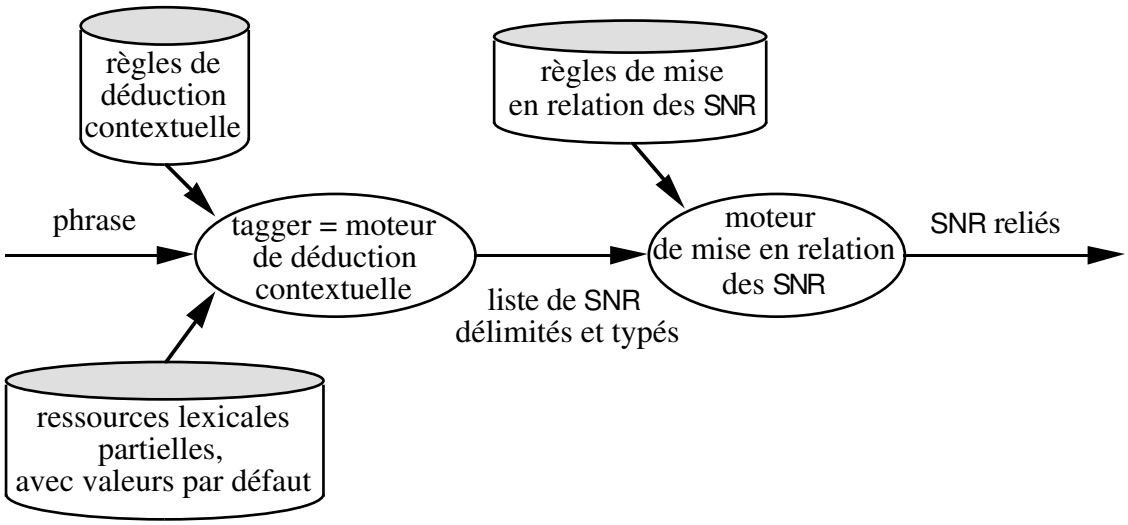
parmi verbe, nom, adjectif et ils constituent seuls un SNR de type multiple (voir ci-dessus en section 2). Les autres tokens et les autres SNR ont respectivement une catégorie unique et un type unique. Après les déductions contextuelles, les catégories par défaut de *de d' du des* et *que qu'* sont encore respectivement préposition et conjonction.

Que manque-t-il alors pour terminer l'analyse syntaxique? Il faut relier les SNR (relations de dépendances, d'apposition, de coordination, d'antécédence, ...).

Dans l'esprit du tagging (l'explicitation des processus), nous avons conçu un processus de mise en relation des SNR, qui ne fait aucune hypothèse explicite sur les structures syntaxiques situées entre les SNR reliés, ni sur la distance qui les sépare. Ce processus est implémenté comme dans le tagger, sous la forme d'un moteur qui interprète des règles [5] qui explicitent le processus. Ses principes et son implémentation sont décrits dans (Giguet et Vergne 1997).

L'ensemble, les ressources lexicales, les deux moteurs et les deux bases de règles, constituent un analyseur de complexité linéaire, dans lequel les processus sont explicités et qui produit en sortie les structures syntaxiques sous la forme de syntagmes non récursifs reliés (voir figure 4 ci-dessous).

Figure 4 : Tagger intégré dans notre analyseur



Les résultats sur des corpus variés en français, articles de journaux (Le Monde), littérature, textes scientifiques sont visibles sur internet à l'adresse : http://www.info.unicaen.fr/~giguet. C'est cet analyseur que nous avons utilisé pour réaliser l'expérience rapportée ci-dessus (en section 3.3.). C'est aussi cet analyseur qui nous a permis de participer à l'action GRACE, en lui ajoutant en module final une fonction de transfert qui transforme notre tokenisation et notre étiquetage en la tokenisation et l'étiquetage GRACE. Les déductions ayant été faites sur notre jeu d'étiquettes, ce transfert permet de faire les évaluations sur le jeu d'étiquettes GRACE. Ce dernier, à cause de sa trop grande fidélité aux parties du discours traditionnelles et de ses propriétés distributionnelles insuffisantes, n'est pas très approprié aux déductions contextuelles. Mais il a l'intérêt majeur d'être le résultat d'un large consensus de la communauté scientifique du tagging du français, et de permettre de faire les calculs d'écart entre le corpus de référence et les résultats de différents taggers dans des conditions expérimentales unifiées et rigoureuses.

6. Conclusion

Face aux difficultés rencontrées par l'analyse syntaxique traditionnelle, principalement dues à son aspect combinatoire, et à l'obligation de disposer d'un inventaire exhaustif des structures syntaxiques d'une langue, le tagging constitue une échappatoire prometteuse, mais il s'est surtout centré sur ses aspects opératoires, et s'est peu interrogé sur ses bases théoriques.

---

[5] Dans les deux cas, les règles sont de la forme : conditions ⇒ actions.

Dans cet article, nous avons tenté de pallier cette lacune, en montrant que le tagging prend le contre-pied de l'analyse syntaxique traditionnelle en mettant l'accent sur l'explicitation des processus, et que, par là même, il ouvre la voie au renouveau de l'analyse syntaxique en la fondant sur l'explicitation des processus : processus de déduction contextuelle dans les syntagmes non récursifs, et processus de mise en relation des syntagmes non récursifs. On étend ainsi à toute l'analyse syntaxique les propriétés calculatoires du tagging, et on obtient ainsi des algorithmes de complexité linéaire en temps.

L'explicitation des processus en analyse syntaxique (au détriment de l'explicitation des structures) pourrait peut-être aussi conduire des linguistes à déplacer leur intérêt, des structures vers les deux processus de transformation entre les deux "ordres" ainsi définis par Tesnière dans (Tesnière 1959), page 16, § 1 :

1.- L'ordre structural des mots est celui selon lequel s'établissent les connexions.

et page 18, § 8 :

8.- Nous appellerons ordre linéaire celui d'après lequel les mots viennent se ranger sur la chaîne parlée. L'ordre linéaire est, comme la chaîne parlée, à une dimension.

Tesnière définit alors ainsi les deux **processus** de transformation [6] entre ces deux "ordres", page 19, § 4 :

[...] nous pouvons dire que [...] ***parler*** *une langue, c'est en transformer l'ordre structural en ordre linéaire, et inversement que* ***comprendre*** *une langue, c'est en transformer l'ordre linéaire en ordre structural.*

Références

Abeillé Anne et Blache Philippe (1997), "État de l'art : la syntaxe", in *Traitement automatique des langues,* volume 38, n°2, pp. 69-90, ATALA, Paris.

Abney Steven (1996), "Part-Of-Speech Tagging and Partial Parsing", in Ken Church and Steve Young and Gerrit Bloothooft, editors, *An Elsnet Book,* Corpus-Based Methods in Language and Speech, Kluwer Academic, Dordrecht.

Chanod Jean-Pierre et Tapanainen Pasi (1995), "Tagging french - comparing a statistical and a constraint based method", in *Proceedings of the Seventh Conference of EACL (EACL'95)*, pages 149-156, Dublin, Ireland.

Constant Patrick (1991), *Analyse Syntaxique par Couches*, thèse de doctorat présentée à l'École Nationale Supérieure des Télécommunications, Paris.

Giguet Emmanuel et Vergne Jacques (1997), "From part of speech tagging to memory-based deep syntactic analysis", in *International Workshop on Parsing Technologies 1997 Proceedings*, Boston.

Sabah Gérard (1989). *L'intelligence artificielle et le langage,* tome 2 : *processus de compréhension* , Hermès, Paris.

Tesnière Lucien (1959). *Éléments de syntaxe structurale*, Klincksieck, Paris.

Vergne Jacques (1995), "Les cadres théoriques des Traitements Automatiques des Langues syntaxiques : quelle adéquation linguistique et algorithmique? une étude et une alternative", in *Actes de "TALN'95" Conférence sur le Traitement Automatique du Langage Naturel*, Marseille, pp. 24-33.

Voutilainen Atro (1994), *Three studies of grammar-based surface parsing of unrestricted English text*, Publications of the departement of general linguistics, n°24, Université d'Helsinki.

[6] Processus appelés "génération" et "analyse" dans la communauté du TAL.