Comparaison de trois analyseurs symboliques pour une tâche d'annotation syntaxique

Jean-Marie Balfourier, Philippe Blache, Marie-Laure Guénot, Tristan Vanrullen
Laboratoire Parole et Langage – CNRS / Université de Provence
{prenom.nom}@lpl.univ-aix.fr

Mots-clefs : Analyse superficielle, Analyse profonde, Analyseur symbolique, Campagne d'Evaluation des Analyseurs SYntaxiques (EASY), Grammaires de Propriétés (GP).

Keywords: *Shallow parsing, Deep parsing, Symbolic parser, Property Grammars (PG).*

Résumé Nous présentons quelques réflexions concernant les différentes capacités des trois parseurs symboliques engagés dans la campagne d'évaluation EASY, face à une tâche d'annotation syntaxique.

Abstract *We present some reflections about the different abilities of three symbolic parsers, evaluated in the EASY campaign, in a specific syntactic annotation task.*

Introduction

Nous présentons dans cet article trois analyseurs évalués dans le cadre de la campagne EASY reposant sur le formalisme des *Grammaires de Propriétés* (ci-après GP). Les approches utilisées sont toutes les trois symboliques, mais utilisent des techniques différentes. Le premier est un analyseur superficiel, utilisant une grammaire simplifiée adaptée à la campagne d'évaluation dans laquelle un sous ensemble des propriétés de GP est exploité. Les second et troisième analyseurs utilisent des techniques d'analyse profonde. Ils emploient tous deux une représentation de l'information syntaxique sous la forme d'une grammaire complète et contrôlent le processus d'analyse et de déterminisation des résultats grâce à des algorithmes différents qu'il nous a intéressé de comparer.

Ces trois approches présentent bien entendu des résultats différents en termes d'efficacité et de couverture. Cette expérience permet de comparer, au sein d'une même approche symbolique et à partir de ressources identiques, les approches superficielles et profondes, ainsi que différents algorithmes d'analyse.

Nous présentons dans un premier temps les ressources utilisées comme base pour l'analyse, puis nous exposons les trois analyseurs syntaxiques. Enfin, nous abordons la question de l'évaluation des résultats en tentant de différencier les origines des erreurs dans les processus d'analyse.

1 Les ressources utilisées

Le lexique et l'étiquetage. — Les trois analyseurs prennent une entrée identique se présentant sous la forme d'un texte tokenisé et étiqueté. Chaque token est imposé dans le corpus à analyser, de sorte que tous les participants soient évalués sur les mêmes données. Chacun de ces tokens est accompagné d'un ensemble d'étiquettes possibles fournies de façon automatisée grâce à l'étiqueteur WinBrill. Nous avons réétiqueté ces tokens à l'aide de notre propre étiqueteur (celui du LPL), afin de faire correspondre le jeu de traits morphosyntaxiques de ces étiquettes avec celui qu'emploie notre grammaire, ainsi que dans le but d'affiner la qualité de l'étiquetage. De fait, l'étiqueteur du LPL présente une bonne efficacité et se base sur un ensemble de traits plus fins que ceux proposés par WinBrill. Le lexique de 430000 formes utilisé par notre étiqueteur est une variante de celui du LPL (présenté dans Vanrullen *et al.* (2005)), adaptée aux entrées du corpus d'évaluation EASY. Il reste cependant un certain nombre d'erreurs qui se répercuteront bien entendu sur le résultat de l'analyse. Ce point est discuté plus loin.

La grammaire. — D'un point de vue linguistique, la participation à la campagne Easy a consisté à concevoir une grammaire formelle dont les résultats visent à coïncider avec les indications fournies dans le Protocole d'Evaluation des Analyseurs Syntaxiques (ci-après PEAS, Gendner & Vilnat (2004)), qui servait de référence à la fois pour les annotateurs (référence pour les résultats) et pour les développeurs de grammaire (production des résultats). Le formalisme que nous utilisons pour développer des grammaires au LPL est celui des Grammaires de Propriétés (ci-après GP, cf. par exemple Blache (2005)).

Le guide PEAS a été conçu dans un souci de consensus entre les différentes théories linguistiques prises en considération lors de son élaboration. De fait, les choix effectués à la base du guide étant plus ou moins éloignés (suivant les cas) des thèses couramment soutenues dans le paradigme des GP, la conséquence de cela est que nous avons développé une grammaire spécifique à la tâche demandée pour Easy. Cette campagne nous a donc permis de tester, dans des circonstances concrètes, la flexibilité du modèle et de vérifier sa capacité d'expression pour une théorie significativement différente de celles habituellement adoptées.

Concrètement, le développement de la GP a consisté en les étapes suivantes :

1. suite à une première lecture du guide, un choix des propriétés utilisées, et la définition de leur sémantique (leur mode de fonctionnement, variable à volonté, cf. par exemple Vanrullen *et al.* (2003)),
2. interprétation approfondie des indications données dans PEAS et transcription de celles-ci en propriétés,
3. vérification de l'adéquation des résultats fournis par les trois parseurs sur un corpus-test,
4. estimation des causes des erreurs produites : méthodes de parsing, descriptions grammaticales, interférences entre propriétés, etc.,
5. ajustements de la grammaire pour les cas qui la concernent (modifications de la sémantique des propriétés, et/ou des propriétés elles-mêmes et/ou des ensembles de propriétés),
6. reprise des étapes 3, 4 et 5 jusqu'à ce que les résultats soient satisfaisants.

Ce va-et-vient entre ajustements de grammaire et tests sur corpus représentatif nous a permis d'adapter la grammaire, à la fois aux indications de PEAS en amont, et aussi aux différents traitements qui peuvent en être faits par chacun parseurs, en aval.

2 Les parseurs

Analyseur superficiel. — L'analyseur superficiel prend en entrée un texte étiqueté et désambiguisé. Il construit dans une première passe l'ensemble des groupes, puis les relations. La construction des groupes repose sur des informations syntaxiques partielles. Plus précisément, seules les propriétés de constituance et de linéarité ont été utilisées.

La stratégie repose sur une technique d'analyse coin-gauche. Il s'agit de repérer pour chaque token, grâce aux deux propriétés citées, sa faculté d'être coin gauche d'un groupe. Dans de nombreux cas, les informations citées sont suffisantes et l'initialisation du groupe correspondant est systématique. En revanche, certaines situations nécessitent la vérification du contexte immédiat. C'est par exemple le cas des groupes PV, qui débutent par une préposition. Cette dernière initialise habituellement un GP, mais dans un contexte verbal à droite, la préposition devient coin gauche du PV. Chaque initialisation de groupe entraîne la fermeture du groupe précédent.

Le mécanisme est donc en une passe unique et consiste à analyser successivement toutes les suites de trois tokens (le token candidat coin gauche, son contexte gauche et son contexte droit). La connaissance du type du groupe en cours permet de compléter la décision d'initialisation.

Les relations sont calculées dans une seconde passe sur la base des groupes construits. Chaque relation correspond à un traitement spécifique. Un premier traitement consiste à construire différentes tables regroupant les groupes et formes susceptibles d'être source ou cible d'une relation. Chaque relation consiste ensuite à parcourir ces tables et vérifier, en fonction des positions des candidats, leur appartenance à une relation. Dans certains cas (par exemple la relation complément-verbe) les candidats sont des ressources uniques. En d'autres termes, un candidat ne peut être complément d'un seul verbe. Cette information, correspondant à une consommation de ressource, est ajoutée dans les tables pour les items concernés. La détection des relations repose donc globalement sur des critères topologiques. Il s'agit d'une approximation qui ne permet pas d'assurer un bon contrôle ce qui entraîne une surgénération.

Les techniques utilisées par l'analyseur superficiel sont donc très simples, ce qui a bien entendu des conséquences sur le résultat obtenu (en particulier pour ce qui concerne les relations). L'avantage majeur de cette technique est sa robustesse et son efficacité : le corpus total est analysé en 4 minutes environ.

Premier analyseur profond. — Le premier analyseur profond utilise une version XML de la grammaire comme ressource permettant l'analyse. Deux algorithmes spécifiques sont mis en oeuvre : celui lié à l'analyse et celui destiné à déterminiser la sortie. La technique d'analyse repose sur une transformation de la grammaire en un graphe de contraintes. Ce graphe, ainsi que la sémantique des contraintes sont confrontés aux corpus à analyser. L'analyse consiste à produire l'ensemble des contraintes satisfaites et enfreintes par chaque succession de tokens constituant un énoncé phrastique. En cours d'analyse, les catégories de la grammaire EASY sont construites en fonction du nombre de contraintes qui les satisfont. Cette technique utilise la mesure de *densité de satisfaction* présentée dans Vanrullen (2004). Si une construction est suffisamment satisfaite (ceci en fonction d'un seuil préétabli pour chaque corpus), elle sera conservée dans le résultat. Une fois l'analyse achevée, une déterminisation des résultats est réalisée en utilisant à nouveau la mesure de densité de satisfaction : les constructions définitivement choisies pour le résultat final sont celles qui maximisent la somme des densités de satisfaction pour

chaque énoncé. Cette maximisation est réalisée grâce à un calcul sur des cliques (au sens de la théorie des graphes), où chaque clique présente un conflit entre plusieurs hypothèses d'analyse. Cet analyseur présente l'avantage de séparer les algorithmes et les données (le programme est indépendant de la grammaire et de sa sémantique, contrairement au troisième analyseur qui doit les inclure au sein même du programme). Son inconvénient réside par contre dans sa lenteur. La complexité moyenne reste polynomiale, mais la durée du traitement est de plusieurs jours sur plusieurs machines pour le million de mots que constitue le corpus. Pour cette raison, bien qu'il était possible de calculer aussi bien les constituants que les relations à l'aide de ce parseur, nous n'avons pu effectuer que la première tâche dans les délais impartis. Ceci confronte les contingences de la campagne d'évaluation aux possibilités réelles des analyseurs, lorsque ceux-ci sont programmés dans le cadre de la recherche expérimentale en laboratoire.

Second analyseur profond. — Ce second analyseur profond, dans son principe, permet la production de tous les arcs possibles conformes à une grammaire de propriétés donnée. Cette analyse se fait en différentes passes, chaque passe engendrant un niveau hiérarchique supplémentaire dans les syntagmes produits. Les constituants EASY n'ayant qu'un niveau de hiérarchie, une seule passe est nécessaire pour les produire. Mais l'analyse a été complétée par des passes supplémentaires afin de produire les relations.

Pour cela, la grammaire EASY a été étendue par l'ajout de catégories syntagmatiques décrivant l'intégralité d'une phrase. Le choix a été fait, au sein de cette grammaire étendue, de présenter chaque relation comme une contrainte de dépendance particulière entre 2 (ou 3 pour certaines relations) constituants d'un syntagme supérieur. Ainsi, la relation <sujet-verbe> est une contrainte de dépendance entre le groupe nominal et le groupe verbal au sein de la catégorie phrase.

Une fois produit par l'analyseur l'ensemble des constituants possibles d'une phrase, on recherche sa meilleure couverture possible (le syntagme le plus englobant, si possible du genre phrase) ainsi que l'ensemble de ses constituants jusqu'aux catégories lexicales. Les constituants EASY correspondent alors au premier niveau de cet assemblage et les relations, à toutes les dépendances nécessaires à sa construction.

3 Interprétation des résultats

Les résultats de la campagne ne sont pas encore disponibles ; il serait donc prématuré de parler d'une évaluation complète de nos résultats. Cependant à la lumière des travaux de développement effectués au cours de la participation à la campagne et des comparaisons entre les différentes sorties obtenues en fonction des techniques employées, on a pu mettre en évidence un certain nombre de caractéristiques (avantages et limites) propres à chacune des approches adoptées, pour le traitement d'un même corpus à l'aide d'une même grammaire.

Il est bien évident que, nos parseurs se basant sur des entrées étiquetées et cette étape préalable n'étant pas fiable à 100%[1], les erreurs provenant de la phase d'étiquetage sont autant de causes d'erreurs systématiques de parsing (bien que dans ce cas ce ne soit pas le parsing lui-même qui soit à mettre en cause). Cela étant dit, même à partir d'entrées correctement étiquetées,

[1] Pas plus que la transcription elle-même : on peut trouver autant de coquilles dans les textes de sources écrites que dans les transcriptions de corpus oraux, coquilles qui augmentent d'autant la probabilité d'erreur d'étiquetage.

nous avons pu lors de nos étapes de test successives mettre en évidence un certain nombre de différences caractéristiques de traitements entre les parseurs. Compte-tenu du fait que les indications d'annotation ne sont pas censées laisser place à l'ambiguïté, cela signifie que dans ces cas seule l'une des réponses est à considérer comme étant celle attendue. Nous allons donc maintenant présenter quelques-unes de ces différences de traitements, qui peuvent provenir soit des techniques de parsing (méthodes d'introduction des "groupes" ou des "relations"), soit de la grammaire elle-même.

Constituants. — Comme on l'a vu précédemment, le fichier que les parseurs prennent en input est un texte étiqueté. A chaque token correspond une liste d'étiquettes possibles, et parmi elles une (sous-)liste des propositions retenues par le désambiguïseur. Pour des raisons de simplicité (et de probabilité), les deux premiers parseurs ont retenu comme technique de ne considérer que le premier élément de cette liste pour leurs analyses, alors que le troisième prend en compte toutes les possibilités proposées. Il s'est avéré que dans certains cas cette dernière technique ait permis de "rattrapper" une imprécision récurrente du désambiguïseur, et ait ainsi permis de produire une analyse en constituants juste là où les deux autres étaient nécessairement erronées, par exemple dans les cas fréquents d'ambiguïté entre un déterminant et un amalgame préposition + déterminant qui ont la même forme (*des, du*,...), et pour lesquels le désambiguïseur ne choisissait pas toujours la meilleure possibilité. Cela avait pour conséquence que les deux parseurs se contentant de la première possibilité retenue par le désambiguïseur introduisaient systématiquement, en cas d'erreur, des GN à la place de GP et *vice versa*, alors que le troisième pouvait vérifier la cohérence de son choix et opter pour l'étiquette qu'il évaluait comme fournissant l'analyse la plus satisfaisante.

Relations. — Seuls deux des trois parseurs ont produit des relations, le premier et le troisième. Les deux techniques d'introduction des relations relèvent de deux approches totalement différentes :

- Comme il le fait pour les groupes, le premier parseur établit des relations en fonction de leur *constituance* (construction des tables regroupant les candidats possibles) et de leur *linéarité* (introduction des relations pour tous les cas où l'ordre des éléments est celui recherché).
- Le troisième parseur a intégré les relations comme étant des contraintes de *dépendance* caractéristiques, mettant en relation les deux ou trois éléments concernés au sein de syntagmes de niveaux supérieurs aux groupes Easy.

Il en résulte que là où le premier parseur génère non seulement toutes les relations attendues mais aussi un nombre conséquent de relations superflues, le troisième introduit souvent moins de relations, cependant chacune de celles-ci dépend directement de l'exactitude des groupes qui les contiennent et a par conséquent une plus forte probabilité d'être juste. Prenons l'exemple du traitement de l'énoncé suivant :

(1) Tout en adoptant le principe de l'adhésion de ces pays, le Conseil européen a précisé que ceux-ci devraient répondre à certains critères et que la capacité de l'Union à accueillir de nouveaux membres devrait également être prise en compte.

Pour cet énoncé, notre premier parseur a introduit 20 relations, dont 7 justes. Notre troisième parseur a introduit 22 relations, dont 11 justes. Le détail est donné en figure 1[2].

[2]Bien évidemment ces pourcentages ne sont calculés que sur l'énoncé donné en exemple et ne sauraient en aucun cas être représentatifs des résultats généraux des parseurs ; il s'agit juste ici d'illustrer les différences de traitements et non d'évaluer les résultats.

	Parseur 1		Parseur 2	
Relation	Nombre	dont justes (précision)	Nombre	dont justes (précision)
Mod-N	9	4 (44 %)	4	3 (75 %)
Suj-V	3	1 (33 %)	1	1 (100 %)
Cod-V	2	0 (0 %)	2	2 (100 %)
Cpl-V	1	0 (0 %)	4	2 (50 %)
Mod-V	1	0 (0 %)	1	1 (100 %)
Aux-V	1	1 (100 %)	2	2 (100 %)
Comp	2	0 (0 %)	8	0 (0 %)
Coord	1	1 (100 %)	0	0
Total	20	7 (35 %)	22	11 (50 %)

FIG. 1 – Détail des relations pour les parseurs 1 et 3 sur l'énoncé de l'exemple (1).

On voit que même si le nombre total de relations introduites n'est pas très différent (20 dans un cas, 22 dans l'autre), par contre la pertinence de ces relations diffère d'un parseur à l'autre, puisque dans la moitié des cas le parseur 3 ne fournit que des bonnes relations (pour Suj-V, Cod-V, Mod-V et Aux-V), et au moins 50 % de justes dans la moitié des cas restants (Pour Cpl-V et Mod-N). Par contre toutes ses propositions de Comp sont erronées, et il n'a pas trouvé la relation de Coord que le premier parseur a su construire.

Grammaire. — Le Protocole d'Evaluation propose une description des Groupes Nominaux indiquant, globalement, que fait partie du GN tout ce qui est compris entre le déterminant et le nom (ou l'objet qui occupe sa place). Cela ne comprend pas, donc, tous les possibles constituants du syntagme nominal (au sens classique) qui figurent après le nom. Le traitement de ce point a été facilement représentable dans notre GP. Cependant, on peut lire plus loin dans le guide PEAS que cette description a une limite : elle n'est plus valable pour les "*éléments en langue étrangère, (l)es formules, (l)es équations mathématiques ou chimiques*", ni pour les "*références bibliographiques au sein de textes (comme dans les articles)*" (Gendner & Vilnat (2004), section D.1.X). Pour ces cas précis, il est dit qu' "*ils peuvent être regroupés dans des constituants*" (*ibid.*). Les trois exemples présentés (et leurs analyses respectives) dans PEAS sont les suivants :

(2) a. La patiente présentait **<GN> un placenta prævia </GN>.**

b. L' amour est plein de quiétude et gardé de sentinelles <GP> à toutes les portes des sens </GP>, et **<GP> in cunclis sensibus custoditus </GP>**.

c. Le cas souvent étudié (**<GN> Hamburger 99 </GN>**) est revu dans cet article.

L'exemple (2a) montre un cas où un mot étranger (*prævia*) a reçu un traitement différent de la norme, du fait de sa nature de "mot étranger" : s'il avait été considéré comme un mot "normal" il n'aurait pas été intégré au GN *un placenta*, mais aurait été l'objet d'un GA unaire (puisque postposé au nom auquel il se rapporte). L'exemple (2b) montre le parallèle fait entre le GP *à toutes les portes des sens* et le groupe suivant, qui est une expression latine, *in cunclis sensibus custoditus*, qui se voit affecter l'étiquette de GP parce que coordonné au GP qui le précède[3]. Enfin, l'exemple (2c) montre le cas exceptionnel de la citation (référence) où un objet qui non

[3]Ou alors du fait de l'analyse syntaxique du groupe latin, mais le parsing du latin n'étant pas l'objet de la grammaire ni de la campagne, et le latin n'étant pas la seule langue possible dans ce cas, nous avons par principe abandonné cette possibilité.

seulement n'est pas un Nom propre (une date en l'occurrence) mais qui est également post-posé, peut faire partie d'un GN lui-même constitué d'un Nom propre (ce qui théoriquement est impossible, en vertu de la description du GN donnée en section B.2).

Les cas tels que celui de l'exemple (2c) ont pu être décrits sans problème dans la grammaire. En revanche, les exceptions illustrées par les exemples (2a) et (2b) ont été impossibles à exprimer. En effet, les groupes que la grammaire permet d'introduire ne contiennent pas d'information qui permette de savoir si un groupe donné est constitué d'éléments de langue étrangère, de formules ou d'équations mathématiques ou chimiques. Cette information, si elle pouvait figurer, proviendrait de l'étiquette des constituants et non de l'analyse, puisqu'en termes strictement syntaxiques, ces "natures" de groupes ne sont pas pertinentes en soi : leur analyse demeure la même que pour tous les autres groupes. Or ni l'étiquetage fourni pour les besoins de la campagne, ni notre couple étiqueteur-désambiguïseur, ne permet cela : soit le lexique contient le mot à étiqueter (cas des expressions mathématiques) et nos étiquettes ne donnent pas d'information de ce type, soit le lexique ne contient pas le mot à étiqueter (cas des mots de langue étrangère[4]), et dans ce cas la tâche consistera à évaluer quelle est la catégorie la plus probable du mot inconnu en fonction de son contexte, mais rien ne pourra nous permettre d'affirmer qu'il s'agit d'un mot de langue étrangère.

Dans ces cas donc, nous avons fait le choix de nous référer uniquement aux étiquettes fournies aux analyseurs, et aux propriétés de la grammaire. Tous les mots donc, qu'ils soient d'origine étrangère ou non, qu'ils soient des formules mathématiques ou non, on été traités selon les définitions des groupes fournies en section B, même si cela constitue une limite de l'adéquation des résultats de nos parseurs avec l'annotation de référence.

Protocole. — A l'étude détaillée du corpus test, nous avons pu mettre en évidence certains cas dont le traitement demandait un choix, lequel n'était pas spécifié dans PEAS. C'est le cas notamment du traitement des bribes et des amorces : il n'est pas spécifié dans le guide si dans un cas de disfluence, l'on doit considérer chaque occurrence de la répétition (du *reparandum* au *repair*) comme faisant partie du groupe (ce qui donne l'annotation de (3a)), ou alors si l'on ne doit faire figurer dans le groupe que l'occurrence du *repair* (exemple (3b)) :

(3) a. **<NV> il il se tachait </NV>** sa sa **<NV> il ne ne buvait </NV>** que des Blancs

b. il **<NV> il se tachait </NV>** sa sa il ne **<NV> ne buvait </NV>** que des Blancs

Dans un cas comme celui-ci, nous avons donc du faire un choix parmi les possibilités, ne sachant pas lequel de ces choix avait été fait par les annotateurs lors de l'établissement de la référence. En l'occurrence, nous avons choisi de faire figurer toutes les occurrences des répétitions dans les groupes, comme dans l'exemple (3a). Mais si la décision des annotateurs a été différente, alors dans tous ces cas notre annotation sera considérée comme erronée alors qu'il ne s'agit précisément que d'une question de convention et non de justesse d'analyse.

Conclusion

L'utilisation d'une approche symbolique dans une tâche d'annotation de corpus n'est sans doute pas la plus naturelle. Les techniques stochastiques sont en effet parfaitement adaptées à un trai-

[4]Sauf certains latinismes et anglicismes courants. Il serait d'ailleurs intéressant de définir précisément ce que l'on entend par "élément de langue étrangère" dans le cadre du Protocole, pour savoir si des entrées telles que *a priori*, *cool* ou bien *(e)mail*, *erzatz* en font partie ou non.

tement de ce type qui s'appuie sur un style d'annotation fermé. Cependant, les analyseurs symboliques offrent d'autres avantages, en particulier si le formalisme qu'elles utilisent permet une flexibilité de traitement ; c'est le cas des Grammaires de Propriétés dans lesquelles la granularité d'analyse peut être choisie. Ce réglage s'effectue en choisissant le type et le nombre de contraintes à satisfaire. Nous sommes donc en mesure, à partir d'une même grammaire et d'une même stratégie d'analyse, de proposer plusieurs types de traitement offrant des résultats plus ou moins détaillés en fonction des besoins. Là où les approches stochastiques nécessitent un réglage particulier en fonction de chaque tâche d'annotation demandée, une approche symbolique du type de celle décrite ici permet au contraire d'envisager une réutilisabilité à la fois des ressources exploitées (lexique, grammaire), mais également des moteurs utilisés.

Références

Philippe Blache. Property grammars : A fully constraint-based theory. In H Christiansen, P Skadhauge, & J Villadsen, editors, *Constraint Satisfaction and Language Processing*. Springer-Verlag, 2005.

Véronique Gendner & Anne Vilnat. Les annotations syntaxiques de référence peas, version 1.6. Révisions par : Laura Monceaux, Patrick Paroubek, Isabelle Robba, 2004.

T. Vanrullen, P. Blache, C. Portes, S. Rauzy, J.F. Maeyhieux, J.M. Balfourier, M.L. Guénot, & E. Bellengier. Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales. In *Actes de TALN 2005*, 2005.

Tristan Vanrullen. Analyse syntaxique et granularité variable. In *Actes de RECITAL 2004*, 2004.

Tristan Vanrullen, Marie-Laure Guénot, & Emmanuel Bellengier. Formal representation of property grammars. In *Proceedings of ESSLLI Student Session*, 2003.