La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus

WIDLÖCHER Antoine, BILHAUT Frédérik
GREYC - CNRS UMR 6072 - Université de Caen
Campus II, Sciences 3, B.P. 5186, 14032 Caen Cedex, France
{awidloch,fbilhaut}@info.unicaen.fr

Mots-clefs : Linguistique de corpus, TAL, plate-forme logicielle

Keywords: Corpus linguistics, NLP, software framework

Résumé À travers la présentation de la plate-forme LinguaStream, nous présentons certains principes méthodologiques et différents modèles d'analyse pouvant permettre l'articulation de traitements sur corpus. Nous envisageons en particulier les besoins nés de perspectives émergentes en TAL telles que l'analyse du discours.

Abstract By presenting the LinguaStream platform, we introduce different methodological principles and analysis models, which make it possible to articulate corpus processing tasks. More especially, we consider emerging approaches in NLP, such as discourse analysis.

1 Introduction

Par delà la diversité des domaines d'investigation et des objets d'étude, un certain nombre de tendances communes se confirment peu à peu au sein de la communauté TAL. Se manifeste tout d'abord, désormais distinctement, la généralisation du travail sur corpus, mouvement qui constitue d'ailleurs un point de convergence fécond entre les travaux spécifiquement dédiés au TAL et les démarches plus immédiatemment linguistiques. Les modèles théoriques proposés doivent désormais trouver leur justification et prouver leur validité « en corpus », et leur pertinence sera jugée, tantôt sur leur capacité à rendre compte de la diversité dudit corpus (dans une perspective descriptive), tantôt à la lumière de leur capacité à l'explorer « efficacement » (dans une perspective d'ingénierie documentaire). Se pose alors inévitablement la question de la méthode et des outils à adopter pour travailler ainsi « sur corpus ».

Il devient en effet difficilement envisageable de considérer celui-ci comme une matière brute à laquelle devraient se référer *immédiatement* les différents modèles et traitements. Au contraire, la multiplication des *points de vue* sur le corpus, qu'ils soient morphologiques, syntaxiques, sémantiques, rhétoriques ou pragmatiques, qu'ils ne visent que l'une de ces dimensions ou qu'ils les croisent, rend pressante la question des interdépendances entre ces vues possibles, interdépendances qui seront d'autant plus nombreuses que des résultats satisfaisants seront obtenus par chacune des approches. Une récente Journée d'Étude de l'ATALA a d'ailleurs permis de poser très frontalement la question désormais centrale de l'articulation des traitements sur corpus. Or, si l'articulation des traitements rend indispensable une réflexion sur leur modularité, elle conduit également à réinterroger l'ensemble de leur processus d'élaboration, de la prise en

charge du corpus, jusqu'à l'*évaluation* des résulats, en imposant que soient repensées les notions même d'*observation* et d'*expérimentation*, à travers, en particulier, une réflexion sur les cycles d'*évaluation/validation* puis d'ajustement des méthodes d'analyse.

Enfin, de nouvelles perspectives en TAL confirment ces nouveaux besoins. Si nous considérons l'intérêt récent accordé à l'analyse automatique de l'organisation discursive, par exemple en termes *thématiques* (Bilhaut, 2004) ou *rhétoriques* (Widlöcher, 2004), il apparaît clairement que ces investigations sont rendues possibles par la préexistence de résultats satisfaisants aux niveaux de granularité inférieurs, en matière d'analyse morpho-syntaxique et sémantique, aux niveaux lexicaux et syntagmatiques. Ces travaux se trouvent consubstantiellement liés à des stratégies d'*empilement* de traitements successifs permettant l'*enrichissement incrémental* des vues sur le corpus et l'abstraction progressive des formes de surface par l'utilisation des analyses préalables. À travers la présentation de LinguaStream[1], nous envisageons ici différents éléments méthodologiques et techniques pouvant permettre d'assumer ces nouvelles orientations.

2 La plate-forme LinguaStream

LinguaStream (Bilhaut & Widlöcher, 2005) a pour principale ambition de faciliter la réalisation d'expériences sur corpus non triviales en TAL, ainsi que le cycle d'évaluation/ajustement qui en découle. Sans outil adapté, le coût de développement induit par chaque nouvelle expérience devient en effet un frein considérable à l'approche expérimentale. Pour répondre à ce problème, LinguaStream facilite la mise en œuvre de procédés complexes tout en requérant des compétences informatiques minimales. Plate-forme générique fondée sur le principe d'**enrichissement incrémental** des documents électroniques, elle facilite la conception et l'évaluation de chaînes de traitements complexes, par assemblage visuel de modules d'analyse de types et de niveaux variés : morphologique, syntaxique, sémantique, discursif... Chaque palier de la chaîne de traitement se traduit par la découverte et le marquage de nouvelles informations, sur lesquelles pourront s'appuyer les analyseurs subséquents.

Un environnement de développement intégré permet de construire visuellement ces chaînes de traitement, à partir d'une « palette » de composants (une cinquantaine est intégrée en standard) facilement extensible grâce une API Java, un système de macro-composants, et des *templates*. Certains sont spécifiquement dédiés au TAL, et d'autres permettent de résoudre différents problèmes liés à la gestion des documents électroniques (traitements XML en particulier). D'autres peuvent être utilisés pour effectuer des calculs sur les annotations produites par les analyseurs, ou encore générer des diagrammes. Chacun dispose d'un ou plusieurs points d'entrée et/ou de sortie que l'on relie pour obtenir la chaîne voulue, celle-ci étant représentée par un graphe où les divers composants apparaissent sous forme de « boîtes » reliées entre elles. Chaque composant propose un nombre variable de paramètres permettant d'adapter leur comportement. Les marquages produits sur un même document sont organisés en couches indépendantes, supportant enchâssements et chevauchements. La plate-forme se base systématiquement sur les **standards et outils** XML, et peut traiter tout fichier de ce type en préservant sa structure orginelle. À l'exécution, elle se charge de l'ordonnancement des sous-tâches, et différents outils permettent *in fine* de visualiser les documents analysés et leurs annotations.

Principes fondamentaux

En premier lieu, la plate-forme recourt systématiquement à des **représentations déclaratives** pour spécifier les différents traitements, ainsi que leur enchaînement sous forme de graphe. Les différents formalismes disponibles permettent ainsi de transcrire directement l'expertise

[1] On trouvera une présentation complète de la plate-forme à l'adresse `http://www.linguastream.org`.

linguistique à mettre en œuvre, l'appareil procédural qui en résulte étant pris en charge par la plate-forme. Les règles données ont donc une valeur tant **descriptive**, en tant que représentations formelles d'un phénomène linguistique, que **prescriptive**, en tant qu'instructions de traitement fournies à un processus informatique.

De plus, la plate-forme exploite la **complémentarité des modèles d'analyse**, plutôt que de privilégier un hypothétique modèle « omnipotent » capable d'exprimer efficacement tout type de contrainte. Nous faisons en effet l'hypothèse qu'un analyseur complexe doit adopter successivement plusieurs regards sur le même matériau linguistique, auxquels répondront des formalismes distincts. On pourra combiner, au sein d'un même traitement, des expressions régulières au niveau morphologique, une grammaire locale au niveau syntagmatique, un transducteur au niveau phrastique et une grammaire DSDL (cf. *infra*) au niveau discursif. L'interopérabilité des différents modèles d'analyse proposés est garantie par l'usage d'une **représentation unifiée** des marquages et des annotations. Ces dernières sont uniformément représentées par des **structures de traits**, modèle communément utilisé en TAL et en linguistique, et permettant de représenter des annotations riches et structurées. Tout composant d'analyse pourra produire son propre marquage en s'appuyant sur les analyses précédentes : les formalismes proposés permettent de spécifier des contraintes sur les annotations existantes par **unification**. La plate-forme favorise ainsi l'**abstraction progressive des formes de surface**. Chaque palier d'analyse pouvant accéder simultanément aux annotations produites par tous les paliers antérieurs, les analyseurs de plus haut niveau sont généralement conduits à s'abstraire progressivement du matériau textuel pour ne plus reposer que sur des représentations symboliques antérieurement calculées.

Un autre aspect important concerne la **variabilité du grain d'analyse** au cours du traitement. De nombreux modèles d'analyse imposent la définition d'un grain d'analyse minimal, dit « jeton » ou *token*. C'est par exemple le cas de toute grammaire ou transducteur : ces formalismes supposent l'existence d'une unité textuelle (comme le caractère ou le mot) à laquelle s'appliquent les patrons. Quand la définition de ce grain minimal est nécessaire au fonctionnement d'un composant, la plate-forme permet de spécifier **localement** le ou les types d'unités à considérer comme jetons. Toute unité préalablement délimitée peut jouer ce rôle : il pourra s'agir du découpage habituel en mots, mais aussi de toute autre unité ayant été préalablement marquée : syntagmes, phrases, cadres du discours, etc. Le grain minimal peut donc être différent pour chaque palier de l'analyse, ce qui augmente considérablement la portée des différents modèles d'analyses utilisables dans la plate-forme. D'autre part, chaque module d'analyse spécifie les marquages antérieurs auxquels il souhaite faire référence, pouvant ainsi ne tenir compte que des marquages qu'il estime pertinents, et donc s'affranchir partiellement de la linéarité du texte. La combinaison de ces fonctionnalités permet d'adopter un **point de vue** sur le document spécifique à chaque étape d'une chaîne de traitement.

La **modularité** des chaînes de traitements favorise quant à elle la **réutilisabilité** des composants dans des contextes différents : un module d'analyse développé au sein d'une première chaîne pourra être réutilisé dans d'autres chaînes. De façon similaire, toute chaîne pourra être réutilisée en tant que constituant d'une chaîne de plus haut niveau, sous forme de « macro-composant ». Réciproquement, pour une chaîne donnée, on pourra **substituer** à un composant tout autre composant **fonctionnellement équivalent**. Pour une sous-tâche donnée, un prototype rudimentaire pourra être remplacé *in fine* par un équivalent pleinement opérationnel. Ceci rend possible la mise en comparaison des traitements, en soumettant ces derniers à des contextes rigoureusement identiques, condition *sine qua non* d'une comparaison pertinente.

Modèles d'analyse

Nous avons évoqué plus haut quelques-uns des composants susceptibles de prendre part à une chaîne de traitement. Parmi ceux spécifiquement dédiés au TAL, on peut distinguer deux familles. La première regroupe les analyseurs « prêts à l'emploi », dédiés à une tâche précise.

Il s'agira par exemple de l'étiquetage morpho-syntaxique, une interface avec TreeTagger étant intégrée par défaut. Ces composants sont paramétrables, mais il n'est pas possible de modifier fondamentalement leur fonctionnement. D'autres au contraire proposent un *modèle d'analyse*, c'est-à-dire un formalisme de représentation de contraintes linguistiques, éventuellement associé à un modèle opératoire, par lequel l'utilisateur peut spécifier intégralement le traitement à opérer. Ils permettent d'exprimer des contraintes tant sur les formes de surface que sur les annotations insérées par les analyseurs précédents. Toutes les annotations sont représentées sous forme de structures de traits, et les contraintes sont systématiquement spécifiées par unification sur ces structures. Quelques-uns des systèmes proposés sont :

– Un système appelé EDCG (pour *Extended*-DCG), permettant de décrire des **grammaires locales d'unification** en se basant sur la syntaxe DCG (*Definite Clause Grammars*) de Prolog. Une telle grammaire peut être décrite dans le plus pur style déclaratif, bien que les spécificités du langage logique restent accessibles aux utilisateurs expérimentés.
– Un système, nommé MRE (pour *Macro-Regular-Expressions*), permettant de décrire des patrons par **transducteurs**, s'appliquant aussi bien aux formes de surface qu'aux annotations préalablement calculées. Sa syntaxe est similaire à celle des expressions régulières communément utilisées en TAL et en linguistique sur corpus. Mais à la différence de ces dernières, ce formalisme ne s'applique pas spécifiquement aux caractères ni aux mots, et peut porter sur toute unité textuelle préalablement analysée.
– Un formalisme d'expression de **contraintes au niveau discursif**. En cours d'élaboration, DSDL (*Discourse Structure Description Language*), que nous décrirons plus loin, permet l'exploration des organisations discursives par l'expression et la satisfaction de contraintes, pouvant être non séquentielles exprimées à l'aide d'un ensemble de fonctions discursives primitives (présence/absence, cohérence sémantique...), et pouvant porter en particulier sur les annotations produites en amont et sur des relations entre ces dernières.
– Un système d'annotation à partir de **lexiques sémantiques**, un système de **tokenisation** basé sur des expressions régulières (au niveau caractère), un système permettant de délimiter des objets linguistiques en se basant sur le balisage XML du document, etc.

3 Analyse du discours

Voyons quels avantages l'analyse automatique du discours peut tirer des principes proposés. Un premier apport significatif résulte de l'approche par enrichissement incrémental et par abstraction progressive des formes de surface. S'il est naturel d'opérer au niveau de ces dernières pour une analyse par exemple morphologique ou syntaxique, il va sans dire que l'analyse discursive ne peut s'accomoder de la diversité combinatoire apparaissant à ce niveau, et qu'un filtrage s'impose. En plus de la possibilité d'opérer la pure et simple *occultation* d'éléments peu pertinents pour tel ou tel besoin interprétatif, la plate-forme permet d'opérer ladite abstraction de deux manières complémentaires. En premier lieu, l'unicité du modèle de marquage et d'annotation donne à chaque étape d'analyse l'accès aux représentations symboliques produites en amont, et permet ainsi de ramener la diversité combinatoire de surface à celle des valeurs interprétées, généralement moins nombreuses. En second lieu, le principe de variabilité du grain d'analyse déjà évoqué permet d'exploiter au niveau discursif des modèles d'analyse habituellement dédiés à des niveaux de granularité inférieurs. Par exemple, des règles EDCG pourront aussi bien décrire des patrons syntagmatiques qu'une grammaire textuelle, selon le grain choisi.

Par ailleurs, la plate-forme propose des modèles d'analyse spécifiquement adaptés au niveau discursif. Le langage DSDL en particulier, s'écarte des paradigmes généralement adoptés par les autres formalismes (y compris MRE ou EDCG), qui reposent fondamentalement sur des principes de *linéarité* (on tient compte de tous les éléments successifs) et de *séquentialité* (un ordre est imposé), principes souvent inadaptés au niveau discursif. En permettant l'expression de

contraintes *non séquentielles* et *non linéaires*, le formalisme DSDL autorise l'expression et la détection de motifs pouvant porter sur des éléments distants dans le texte, sans faire d'hypothèse sur leur ordre, ce qui s'avère particulièrement adapté à l'analyse du discours.

Afin de donner une idée plus concrète des principes méthodologiques présentés, envisageons à présent une configuration linguistique particulière, assez représentative des problèmes posés par l'analyse discursive, en abordant le problème de l'encadrement du discours (Charolles, 1997), et plus particulièrement de la détection automatique des *cadres temporels*. Rappelons que cette théorie qualifie ainsi des segments textuels homogènes du point de vue d'un critère d'interprétation fixé dans une expression en position détachée en début de phrase, dite *introducteur de cadre*. L'opérationnalisation en TAL de ce modèle psycho-linguistique impose la résolution de deux problèmes principaux : détection des introducteurs, puis évaluation de leur *portée*, c'est-à-dire détermination de la borne droite du cadre introduit. Bien que cette dernière tâche soit très problématique dans la mesure où les critères formels de clôture des cadres sont difficiles à établir, un certain nombre d'indices ont toutefois pu être dégagés dans le cas précis des cadres temporels (Bilhaut *et al.*, 2003), que nous évoquerons ci-après.

Le problème de la détection des introducteurs temporels se décline lui-même en deux sous-problèmes : l'analyse des expressions temporelles, et celle des introducteurs s'appuyant sur elles. Les principes de modularité évoqués trouvent ici leur justification, puisque nous souhaiterons généralement traiter ces problèmes indépendamment. L'analyse sémantique des expressions temporelles fait l'objet d'une grammaire EDCG, exprimant des contraintes sur les résultats d'une analyse morpho-syntaxique préliminaire, et associant aux expressions reconnues une représentation de leur « sens » sous forme de structures de traits. Sur cette base, la détection des introducteurs peut être mise en place à l'aide de critères essentiellement positionnels. Les contraintes exprimées sont fondamentalement séquentielles : nous recherchons des zones de texte vérifiant des motifs imposant la présence, dans un ordre fixé, d'éléments immédiatement successifs. Ces règles sont donc simplement exprimables à l'aide du formalisme MRE (outre les expressions temporelles, nous exploitons ici le marquage des phrases et des connecteurs de discours) :

```
{type : phrase, anchor : start}
<introducteur>
{type : connecteur}? {tag : pre} {type : temporel} /as $t
</introducteur> /sem {axe : temps, valeur : $t} ","
```

Les contraintes sur les structures de traits produites en amont (ici en gras), ainsi que sur les formes de surface (ici, la virgule en fin de motif) permettent de délimiter l'introducteur. Nous recherchons les éléments précédés d'un début de phrase et composés, d'un éventuel connecteur de discours, d'une préposition et d'une expression temporelle. Le reste de l'expression correspond au marquage et à l'annotation produits en sortie. L'élément reconnu aura le type « introducteur » et sera associé à l'annotation sémantique qui lui fait suite. Précisons que la variable `$t` permet de faire « remonter » l'information contenue dans la structure de traits associée à l'expression temporelle, pour un usage ultérieur.

Pour la détermination de la portée de l'introducteur, la méthode présentée dans (Bilhaut *et al.*, 2003) s'appuie sur des critères énonciatifs tels que la cohésion des temps verbaux, sur la structuration en paragraphes, et sur des calculs sémantiques de cohérence entre l'introducteur et les autres expressions temporelles. La nature de ces contraintes diffère radicalement des précédentes. D'une part, nous pouvons désormais nous abstraire de la linéarité du texte : contrairement à une approche par expressions régulières, nous pouvons ici ignorer un certain nombre d'éléments du flot textuel. D'autre part, s'il existe bien des contraintes interprétatives entre l'introducteur et certains éléments de la zone introduite, il n'est pas souhaitable de concevoir ces contraintes comme imposant un ordre strict entre ces éléments. Pour l'expression de telles contraintes à la fois *non linéaires* et *non séquentielles*, nous disposons du formalisme DSDL

et pouvons formuler la « grammaire » ci-dessous. Nous recherchons une unité textuelle composée de phrases complètes, commençant par un élément identifié comme introducteur et ne comportant pas d'autre élément de ce type, dont tous le verbes sont au même temps, et au sein de laquelle les expressions temporelles portent sur une plage comprise dans l'intervalle fixé par l'introducteur, en ne retenant que le plus long des candidats partageant un même introducteur.

```
Rule {type : "cadre"} :
        start({type : "introducteur"})
        end({type : "phrase"})
        homogeneity(comparator : portée)
        not presence(pattern : {type : "intro"}, amount : 2)
        size(mode : #LONGEST)

Comparator portée ({type : "verbe"} as $v1, {type : "verbe"} as $v2) :
        $v1/temps = $v2/temps

Comparator portée ({type : "intro"} as $i, {type : "tempo"} as $t) :
        ($t/debut >= $i/debut) and ($t/fin <= $i/fin)
```

Il est ainsi possible, à l'aide des principes méthodologiques promus par la plate-forme, et en nous appuyant sur la complémentarité des modèles d'analyse, de mettre en place un analyseur de cadres temporels, certes encore imparfait, mais ne faisant usage que de formalismes purement déclaratifs propices à la capitalisation de l'expertise linguistique mise en œuvre.

4 Conclusion

Initialement développée dans le cadre du projet GeoSem[2], la plate-forme évolue maintenant indépendamment. Elle est aujourd'hui utilisée dans le cadre d'un projet TCAN[3], de différents travaux de recherche en TAL, notamment en analyse sémantique du discours : organisation thématique (Bilhaut, 2004), ou rhétorique (Widlöcher, 2004). Le logiciel est également utilisé à des fins pédagogiques au GREYC et à l'ERSS, et a été mis à la disposition de laboratoires tels que LIUPPA ou LATTICE. La plate-forme reste en elle-même indépendante des modèles d'analyse utilisés, pour peu qu'ils partagent le même système de marquage et d'annotation, et il est donc envisageable d'intégrer des modules exploitant d'autres modèles d'analyse.

Références

BILHAUT F. (2004). Analyse automatique de la structure thématique du discours pour la navigation documentaire. In *Journée ATALA « Modéliser et décrire l'organisation discursive à l'heure du document numérique »*.

BILHAUT F., HO-DAC M., BORILLO A., CHARNOIS T., ENJALBERT P., DRAOULEC A. L., MATHET Y., MIGUET H., PÉRY-WOODLEY M.-P. & SARDA L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. In *Actes de Traitement Automatique du Langage Naturel (TALN)*, Batz-sur-Mer, France.

BILHAUT F. & WIDLÖCHER A. (2005). La plate-forme LinguaStream. In *Journée ATALA « Architectures logicielles pour articuler les traitements sur corpus »*.

CHAROLLES M. (1997). L'encadrement du discours : Univers, champs, domaines et espaces. Cahier de Recherche Linguistique no 6. Université de Nancy 2.

WIDLÖCHER A. (2004). Analyse macro-sémantique : vers une analyse rhétorique du discours. In *Actes de RECITAL 2004*, p. 183–188, Fès, Maroc.

[2] « Traitement sémantique de l'information géographique », programme CNRS « Société de l'information ».

[3] « Intervalles temporels et applications à la linguistique textuelle », projet interdisciplinaire du CNRS.