Du texte vers le sens en analyse par contraintes

Francis Brunet-Manquat

GETA-CLIPS-IMAG (UJF & CNRS)
BP 53 - 38041 Grenoble Cedex 9, France
Courriel : Francis.Brunet-Manquat@imag.fr

Résumé - *Abstract*

Les progrès réalisés ces dernières années dans le domaine du traitement automatique des langues naturelles (TALN) ouvrent la voie à des traitements encore plus sophistiqués dans lesquels la sémantique devrait tenir une place centrale. Notre objectif, à long terme, est de réaliser un analyseur texte vers sens s'appuyant sur la théorie Sens-Texte d'Igor Mel'cuk. Cette analyse viserait une *compréhension* plus approfondie du texte, permettant donc d'atteindre une représentation de niveau sémantique, et une grande *robustesse* face à des entrées plus ou moins bien formées telles que celles issues de dialogues oraux. Mais renverser la théorie Sens-Texte passe par la définition et la mise en œuvre de structures de données et d'algorithmes spécifiques pour la représentation et la manipulation automatique des informations linguistiques, notamment des entrées lexicales. Pour cela, nous proposons l'utilisation du paradigme de programmation par contraintes qui offre un moyen efficace d'atteindre nos objectifs.

Recent advances in the field of natural language processing allow more sophisticated treatments in which semantics should play a key role. Our long-term goal is to produce a text towards meaning analyzer base on Igor Mel'cuk's Meaning-Text Theory (MTT). Such an analyzer should aim at a deep understanding of the input, producing a semantic representation. It should, also, be able to handled ill-formed inputs such as the ones produced by a speech recognizer. To reverse the Meaning-Text Theory we have to define and use specific data structures and algorithms to represent and handle the linguistic information, in particular the lexical entries. We investigate the use of the constraint-programming paradigm, which provides an efficient mean to reach our goals.

Mots clefs - *Keywords*

Analyse par contraintes, théorie Sens-Texte – *constraint parsing, Meaning-Text theory*

1 Introduction

Le domaine du traitement automatique des langues naturelles (TALN) a connu ces dernières années un développement important donnant lieu à la réalisation d'applications industrielles et commerciales solides. Cependant, l'état de l'art en matière de TALN, et notamment en matière de description linguistique, ouvre la voie à des traitements encore plus sophistiqués dans lesquels la sémantique devrait tenir une place centrale. L'équipe GETA est notamment impliquée dans deux projets internationaux importants : CSTAR, avec le projet européen associé NESPOLE! (http://nespole.itc.it) pour la traduction simultanée de l'oral et UNL (http://www.unl.ias.unu.edu)

pour la traduction de l'écrit. Ces deux projets se caractérisent par la présence d'une représentation *pivot* des énoncés orientée vers la sémantique. L'équipe GETA a toujours fait de la traduction par transfert multiniveaux, et ces représentations pivots nécessitent une analyse plus *fine*, notamment au niveau sémantique. Cette analyse viserait une *compréhension* approfondie de l'entrée, permettant donc d'atteindre une représentation de niveau sémantique, et une grande *robustesse* face à des entrées plus ou moins bien formées telles que celles issues de dialogues oraux. Notre but est dans un premier temps l'intégration de cet analyseur dans des systèmes de dialogue oraux ou écrits, intégrant ou non une traduction (CSTAR), puis dans des systèmes de plus large couverture, comme UNL. Pour réaliser cet analyseur, nous devons nous appuyer sur une théorie linguistique solide et universelle (pour pouvoir traiter sinon toutes les langues en tout cas le plus de langues possibles). La théorie *Sens-Texte* d'Igor Mel'cuk [Mel'cuk 1988] semble adéquate pour jouer ce rôle ; elle permet de décrire tous les niveaux linguistiques, de la morphologie/phonétique à la sémantique avec une même structure de représentation : les graphes de dépendances. De plus, elle accorde une forte importance au lexique et les descriptions lexicales sont très précises et très structurées (Dictionnaire Explicatif et Combinatoire : DEC [Mel'cuk & al. 1981]) et semblent donc se prêter à une utilisation informatique. À long terme, l'objectif est de réaliser un analyseur *Texte vers Sens* utilisant cette théorie avec l'idée d'avoir un seul algorithme pour tous les niveaux d'analyse évoqués ci-dessus.

Un des problèmes essentiels à résoudre est la méthode à employer pour réaliser ce type d'analyseur. Notre choix s'est porté sur une méthode d'analyse par contraintes. En fait, tous les formalismes linguistiques font usage de la notion de contrainte qui indique une propriété devant être satisfaite. Les contraintes sont extrêmement utiles à la fois pour représenter l'information linguistique, mais également pour en contrôler le processus d'analyse. C'est la raison qui motive notre orientation vers un analyseur basé essentiellement sur le mécanisme de satisfaction de contraintes.

À court terme, notre objectif est la conception et la mise en œuvre d'une structure de données adaptée au traitement automatique pour les entrées lexicales inspirée de la théorie Sens-Texte et donc du dictionnaire explicatif et combinatoire (DEC). Cette structure de données doit permettre à la fois d'être utilisée comme représentation interne à l'analyse mais également devra être suffisamment intuitive sur le plan linguistique pour que le linguiste chargé du lexique puisse la manipuler sans difficulté. Bien entendu, une telle structure a besoin d'outils pour la manipuler, des outils permettant par exemple la conversion des entrées du DEC et l'ajout ou la modification des entrées lexicales par des linguistes.

Cet article n'a pas d'autre ambition que de présenter la problématique de nos futurs travaux de thèse. Nous proposons dans un premier temps de présenter les formalismes sur lesquels nous nous appuyons pour concevoir un modèle basé sur les grammaires de dépendances. Dans la seconde partie, nous décrirons l'analyse par contraintes : ses avantages, le mécanisme sur lequel elle est fondée et la représentation des données. Nous terminerons en faisant le point sur les perspectives et la méthodologie à employer pour mener ce projet à terme.

2 Aspects linguistiques

L'originalité du travail réside dans l'utilisation d'une théorie linguistique unique, le modèle Sens-Texte sur lequel nous proposons de nous appuyer d'un bout à l'autre de l'analyse. Le problème principal de ce modèle est qu'il a surtout été exploité jusqu'ici par des programmes de génération, car elle est délibérément orientée du sens vers le texte.

2.1 Fondements théoriques

L'approche Sens->Texte (Meaning-text Model), fut mise en évidence dans les années 1960 par Alexander K. Zholkovsky et Igor Mel'cuk [Zholkovsky & al. 1965]. Cette approche sert de

cadre linguistique aux grammaires de dépendances. Ce modèle linguistique met essentiellement l'accent sur la sémantique, et plus précisément sur le lexème ; la syntaxe est déterminée et contrôlée par la représentation sémantique et les entrées du dictionnaire. Elle repose sur des principes généraux applicables à toutes les langues, ce qui la rend *universelle*, ainsi les techniques descriptives et les formalismes proposés s'appliquent à toutes les langues de la même façon [Mel'cuk 1988].

Ce modèle semble être un représentant important d'un mouvement qui considère comme de plus en plus importante l'influence du lexique sur la syntaxe et la sémantique. Outre la clarté des représentations qui résulte des relations de dépendances, certains phénomènes linguistiques, comme l'unicité du sujet d'un verbe, sont exprimés beaucoup plus simplement. La théorie Sens-Texte considère le dictionnaire d'une langue comme le point central du modèle de cette langue, le dictionnaire étant vu comme un répertoire des significations de la langue [Mel'cuk & al. 1981]. Le dictionnaire utilisé par la théorie Sens-Texte se nomme le *Dictionnaire Explicatif et Combinatoire* (DEC). Une version simplifiée du DEC est en cours de réalisation, le DiCo d'Alain Polguère en collaboration avec I. Mel'cuk [Polguère 2000], ce dictionnaire électronique est composé des dérivations sémantiques et des collocations du français. Mais ce projet a pour objectif la réalisation d'un dictionnaire général public entièrement généré à partir du DiCo alors que notre souci premier est la manipulation interne pour l'analyse. Il nous faut trouver une structure de données qui nous permette à la fois d'obtenir une forme externe facilement mais également une forme interne suffisamment formelle pour autoriser l'utilisation des entrées lexicales par des algorithmes. Les arbres de dépendance à bulles proposées par Sylvain Kahane constituent un bon point de départ à cette étude. Ces arbres sont issus des arbres de dépendance classiques, ils permettent une meilleure modélisation des représentations en remplaçant les nœuds par des bulles qui peuvent contenir des sous-bulles liées à d'autres bulles et ainsi de suite. Cette structure permet de dépasser certains inconvénients des grammaires de dépendance comme la projectivité. [Kahane 2000] montre notamment comment ce formalisme peut-être exploité pour traiter l'extraction de manière élégante.

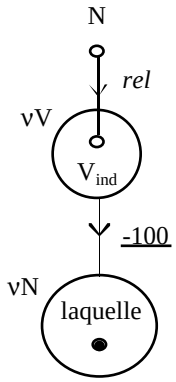
2.2 Vers une structure de données unique

Le premier objectif est donc de définir un langage de description lexicale s'inspirant de la théorie Sens-Texte et du formalisme à bulles proposé par Sylvain Kahane. Ce langage devra comporter :

- Une forme externe dans un langage standard (type XML) exploitable par des outils plus ou moins standard pour produire par exemple une forme papier ; cette forme externe devra être suffisamment intuitive sur le plan linguistique pour que le linguiste chargé de la création du lexique puisse la manipuler sans difficulté.

- Une forme interne (structure de données) manipulable par programme ; elle devra être suffisamment formelle pour autoriser l'utilisation des entrées lexicales par un algorithme d'analyse capable de faire un traitement du Texte vers le Sens.

Cette structure de données devra pouvoir décrire une langue sans être contraint par un aspect procédural quelconque. L'idée est de représenter les informations linguistiques d'une langue par un ensemble de contraintes. Tous les formalismes linguistiques font usage de la notion de contrainte qui indique une propriété devant être satisfaite. Les grammaires de dépendance à bulles se prêtent bien à une expression sous forme de contraintes (figure 1). En effet, les théories non génératives, et en particulier les grammaires de dépendances basées sur des règles de correspondance, peuvent s'appuyer sur des théories basées sur les contraintes en ce qui concerne la représentation des données linguistiques, ce qui permet de considérer l'information sous la forme d'un ensemble d'équations [Blache 2000].



Déclaration des constituants :
N_1 := N(0);
V_1 := V[indicatif](0);
$laquelle_1$:= laquelle(1);
B_1 := νV; // *nucléus verbal*
B_2 := νN; // *nucléus nominal*

Déclaration des relations :
$V_1 \in B_1$;
$laquelle_1 \in B_2$;
$N_1 \rightarrow$[relation] V_1;
$B_1 \rightarrow$[-100] B_2

Figure 1 : Exemple d'une représentation par contraintes[1]
de la structure élémentaire du pronom relatif *laquelle*

Les contraintes sont extrêmement utiles à la fois pour représenter l'information linguistique, mais également pour contrôler le processus d'analyse d'un énoncé. Dans la suite, nous décrirons les avantages et les principes d'une telle analyse.

3 Analyse par contraintes

L'utilisation des contraintes dans les systèmes d'analyse permet de modéliser les différents aspects du langage : syntaxe, sémantique, discours, etc. En plus, elles autorisent l'utilisation d'une méthode d'analyse basée sur un mécanisme de *satisfaction de contraintes*. Nous appellerons cette méthode l'*analyse par contraintes*.

3.1 Caractéristiques

L'analyse par contraintes s'appuie sur une description de l'information linguistique uniquement sous forme d'un ensemble de contraintes et sur des mécanismes permettant de propager et vérifier la cohérence d'un tel ensemble [Maruyama 1990] [Blache 2000] [Schröder & al. 2000]. L'idée de voir l'analyse de langage naturel comme un mécanisme de satisfaction de contraintes amène de nombreux avantages :

- L'approche par contraintes permet une ***description complète*** des différents niveaux d'analyse (morphologique, syntaxique, sémantique, etc.) car la totalité de l'information linguistique de l'énoncé analysé sera représentée sous la forme d'un ensemble de contraintes.

- Le mécanisme de satisfaction de contraintes permet la manipulation de données incomplètes, ce qui favorise l'***analyse robuste***. Il sera ainsi possible d'analyser partiellement une phrase mal formée. Un tel mécanisme semble très efficace pour le traitement de la parole spontanée (projet CSTAR).

- L'analyse par contraintes offre également une très grande ***souplesse*** de traitement, car elle permet de régler la granularité de l'analyse en ne prenant en compte que certaines contraintes. Il est alors possible de réaliser des analyses superficielles utilisées par exemple par des applications d'extraction d'informations.

---

[1] Les informations entre crochets servent à spécifier la variable ou le type de dépendance, la valeur binaire entre parenthèses détermine si le nœud est de type blanc (0) ou noir (1),

- La méthode par contraintes consiste à construire une ***représentation unique*** pour toutes les interprétations d'une phrase à analyser. Aucun arbre d'analyse explicite n'est généré tant que ce n'est pas nécessaire. Le processus de satisfaction de contraintes réduit les ambiguïtés structurales sans générer les arbres d'analyse intermédiaires, limitant ainsi la combinatoire de l'analyse.

3.2 Structure de données

La méthode d'analyse par contraintes s'appuie sur une représentation des informations linguistiques sous la forme d'un système de contraintes que nous nommerons *réseau de contraintes*. Chaque solution qui satisfait toutes les contraintes simultanément correspond a un unique réseau de contraintes. L'analyse et les désambiguïsations successives s'opèrent en ajoutant de nouvelles contraintes au réseau. Notre structure s'inspirera de la théorie CDG, Constraint Dependency Grammar, mise au point par Hiroshi Maruyama [Maruyama 1990].

Le réseau de contraintes doit contenir toutes les structures syntaxiques et sémantiques implicites. Il peut être représenté comme un graphe de dépendances (théorie Sens-Texte). Ce graphe sera constitué de plusieurs niveaux de dépendance en parallèle (syntaxique, sémantique, auxiliaire, etc.) [Schröder & al. 2000]. Chaque niveau sera construit dans un même temps, et sera relié aux autres pour permettre une analyse plus pertinente. Donc la totalité de l'information extraite de l'énoncé à analyser sera contenue dans une unique structure de donnée. Les nœuds du graphe sont appelés des *rôles*. Un mot de l'énoncé correspond à un rôle pour chaque niveau. Chacun des rôles sera associé à un ensemble d'équations, appelé *domaine*, correspondant aux contraintes pouvant lui être appliquées. La structure ainsi obtenue correspondra à un énorme ensemble de contraintes (figure 2). Il faudra ensuite propager et vérifier la cohérence d'un tel ensemble à l'aide d'un mécanisme de satisfaction de contraintes.

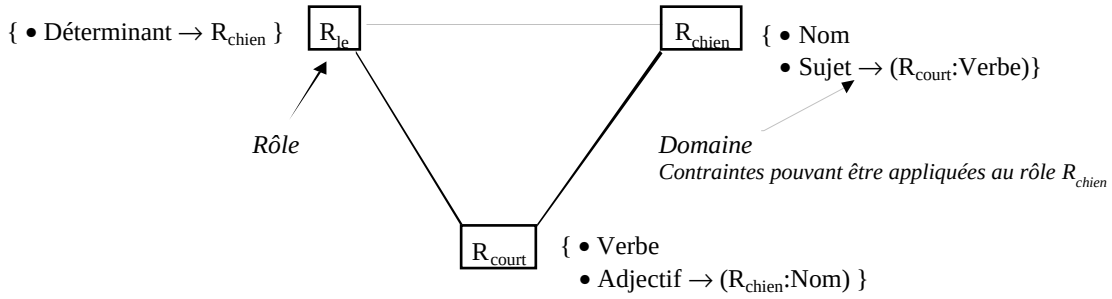


Figure 2 : Réseau de contraintes de la phrase "le chien court" au niveau syntaxique

3.3 Satisfaction de contraintes

Le mécanisme de satisfaction de contraintes consiste à vérifier la satisfaisabilité (la consistance) d'un ensemble de contraintes. Pour cela, aucun processus de dérivation n'est nécessaire pour contrôler la structure d'un énoncé. Ce mécanisme est constitué de trois étapes : Former le réseau de contraintes initial, propager les contraintes, c'est-à-dire lier certains rôles entre eux à l'aide des contraintes et vérifier la consistance. Si la consistance n'est pas vérifiée, enlever les inconsistances locales, c'est-à-dire les contraintes ne satisfaisant pas à tous les rôles, et recommencer la propagation de contraintes jusqu'à obtenir un réseau consistant. Au final, la structure obtenue contiendra la description de l'énoncé avec toutes les interprétations possibles.

La satisfaction de contraintes reste un processus très lourd en terme de complexité algorithmique, nous devrons donc trouver des stratégies d'analyse dans les processus de propagation de contraintes et de vérification de consistance pour rendre l'analyse efficace.

4 Perspectives

L'un des objectifs à atteindre est la *réutilisabilité*, aussi bien en terme de lexique que d'analyse. La réutilisabilité des données est un point essentiel, l'idée est de permettre l'utilisation, l'ajout d'entrée et l'augmentation du lexique à long terme. Il faut donc prévoir une structure capable d'évoluer dans le temps sans perte d'informations. Cette structure permettra la construction incrémentale de lexiques, tendant vers des bases lexicales desquelles on peut extraire des lexiques d'applications. L'analyseur pourrait être également réutilisé pour différentes langues, pour différents types d'analyse (extraction d'informations, analyse syntaxique, analyse sémantique de texte).

Pour arriver à nos fins, notre démarche de travail sera constituée de trois étapes. La première étape sera de terminer notre étude du DEC et du DiCo. Cette étude doit nous amener à définir (ou simplement choisir parmi celle qui existe) une structure de représentation informatique du DEC qui corresponde à nos besoins. Nous utiliserons cette structure pour développer une version informatique du DEC qui sera intégrée à la base Papillon, une base lexicale pour le français et le japonais basé sur des dictionnaires monolingues et des liens interlingues [Mangeot 2000]. La seconde étape sera de réaliser des outils de manipulation pour cette structure, des éditeurs et des analyseurs. Nous devrons développer des algorithmes efficaces permettant la propagation de contraintes dans des temps raisonnables sans trop restreindre le pouvoir d'expression des contraintes. La dernière étape constituera à tester ces outils dans le cadre d'application pilote tels qu'UNL et CSTAR sur un lexique du français.

Références

[Blache 2000] **Blache P. (2000)**, *Le rôle des contraintes dans la théories linguistiques et leur intérêt pour l'analyse automatique : les Grammaires de propriétés,* Proc. TALN-2000, pp. 41-50.

[Brunet-Manquat 2000] **Brunet-Manquat F. (2000)**, *Programmation par contraintes pour l'analyse de dépendances,* Rapport de DEA, UJF et INPG, 13 juin 2000.

[Kahane 2000] **Kahane S. (2000)**, *Extractions dans une grammaire de dépendance à bulles.* **41**/1, pp. 187-216.

[Mangeot 2000] **Mangeot M. (2000)**, *Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links,* Proc. WAINS'7, 7th Workshop on Advanced Information Network and System, Bangkok, Thailand.

[Maruyama 1990] **Maruyama H. (1990)**, *Structural disambiguation with constraint propagation.* Annnual Meeting of the ACL, **28**/pp. 31-38.

[Mel'cuk & al. 1981] **Mel'cuk I., Iordanskaja L. & Arbatchewsky-Jumarie N. (1981)**, *Un nouveau type de dictionnaire: le dictionnaire explicatif et combinatoire du français comtemporain (six articles de dictionnaire).* Cahiers de lexicographie, **38**/1, pp. 3-34.

[Mel'cuk 1988] **Mel'cuk I. (1988)**, *Dependency syntax: theory and practice.* State University of New-York Press.

[Polguère 2000] **Polguère A. (2000)**, *Towards a theoretically-motived general public dictionnary of semantic derivations and collocations for french,* Proc. EURALEX'2000, pp. 517-527.

[Schröder & al. 2000] **Schröder I., Menzel W., Foth K. & Schulz M. (2000)**, *Modeling Dependency Grammars with Restricted Constraints.* Traitement Automatique des Langues, **41**/1, pp. 97-126.

[Zholkovsky & al. 1965] **Zholkovsky A. & Mel'cuk I. (1965)**, *On a Possible Method and instruments for Semantic Synthesis (of Texts).* **6**/pp. 23-28.