Analyse automatique d'un discours spécialisé au moyen de grammaires locales

Takuya Nakamura

IGM – Université de Marne-la-Vallée
77454 Marne-la-Vallée Cedex 2 – France
nakamura@univ-mlv.fr

Abstract

A collection of specialized texts such as stock exchange market reports is formally analyzable into a set of sentence schemata. In this study, I take one example of these schemata, based on a specialized lexicon-grammar to describe formally a corpus constituted of an accumulation of stock exchange market reports found in a French daily newspaper. This linguistic description is also represented in the form of local grammars, in order to analyze automatically the corpus. The local grammars which represent one schema of sentence give rise to a non ambiguous analysis of approximately 22 % of the sentences in the corpus, which correspond to the totality of the sentences characterized by a sentence schema.

Résumé

Une collection de textes spécialisés, comme les bulletins boursiers, est susceptible d'être analysée formellement en une suite de schémas de phrase. Dans cette étude, nous avons pris un exemple de ces schémas de phrase, pour décrire notre corpus constitué de l'accumulation de rubriques boursières d'un quotidien français. Cette description linguistique est représentée également sous formes de grammaires locales, pour analyser automatiquement le corpus. Les grammaires locales, représentant un seul schéma de phrase ainsi décrit, ont reconnu, sans ambiguïté, 22 % des phrases du corpus, ce qui correspond à la totalité des phrases correspondant au schéma.

Mots-clés : analyse automatique d'un discours spécialisé, corpus économique, lexique-grammaire, grammaire locales.

1. Introduction

Les brefs textes journalistiques publiés régulièrement pour rapporter des événements qui se produisent dans un domaine particulier, comme les bulletins météorologiques ou les résultats sportifs, sont des discours spécialisés.

Les rapports boursiers, comme on en trouve dans un quotidien le lendemain d'un jour ouvré, en constituent un exemple. Nous allons montrer que l'impression de stéréotype que donnent ces rapports peut être caractérisée linguistiquement, en réduisant une partie significative des phrases du discours à un schéma de phrase, constitué de plusieurs composants formels. La méthode d'analyse linguistique du corpus est un lexique-grammaire spécialisé pour ce type de discours. La réduction des phrases du discours à un schéma de phrase est inspirée de Harris (1952 et 1963).

Les résultats de cette description linguistique sont manuellement transposés dans des grammaires locales (Gross, 1996 et 1997), pour pouvoir être appliquées au corpus. Grâce à une description linguistique minutieuse des phrases, ces grammaires sont lexicalisées en détail, à tel point que le recours à des noms de catégorie comme éléments de la grammaire locale, qui produit souvent des résultats erronés, peut être évité. Les composants des grammaires locales

sont clairement organisés après analyse syntactico-sémantique des phrases, afin que leur application donne des résultats sans ambiguïté. Cette double méthode - description linguistique et construction de grammaires locales - est libre des contraintes techniques externes qui sont toujours imposées lorsqu'on fait appel à un formalisme particulier de représentation informatique. Les résultats de la reconnaissance automatique sont extrêmement lisibles et explicites et ils peuvent être utilisés à des fins diverses, comme la traduction automatique ou la recherche d'informations dans les corpus.

Cette étude montre l'utilité d'une description linguistique des phrases du corpus préalable à la construction d'un système d'analyse automatique. Si l'on formalise un certain nombre de phrases réelles qui apparaissent dans le texte au moyen du lexique-grammaire, la construction des grammaires locales pour la reconnaissance de ces objets en est largement facilitée. D'autre part, le schéma de phrase étant constitué en tenant compte de tous les composants syntactico-sémantique nécessaires pour caractériser des informations spécifiques au discours y compris les compléments, considérés, en dehors de ce discours particulier, comme circonstantiels, donc aléatoires, la part d'ombre (cf. des insertions imprévisibles laissées dans les grammaires locales) est considérablement diminuée.

Dans ce qui suit, après la présentation du corpus étudié dans la section 2, nous montrons les méthodes linguistiques de l'analyse du corpus dans la section 3, puis la façon de confectionner les grammaires locales basée sur les descriptions linguistiques dans la section 4.

2. Corpus

Le corpus est constitué de rubriques *Valeurs France* parues dans le quotidien *Le Monde*. Dans cette rubrique quotidienne, on trouve de 3 à 6 brefs commentaires sur les mouvements de certaines actions échangées à la Bourse de Paris et les activités économiques ou les phénomènes sociaux liés à ces mouvements. Chaque commentaire comporte environ deux ou trois phrases.

La période de collecte va du 18 janvier 2001 au 24 août 2001, soit au total 146 rubriques comportant à peu près 560 commentaires. La taille du corpus est de 175 kilo octets.

3. L'analyse du corpus

3.1. *L'analyse du discours*

Selon Harris (1952), un discours est réductible à une succession de séquences de plusieurs catégories formelles. Ces catégories sont des classes d'équivalence, établies d'un point de vue distributionnel et transformationnel, en fonction du discours étudié. Nous considérons que l'analyse harrissienne du discours n'est pas applicable telle quelle à un commentaire trouvé dans le corpus qu'on étudie ici, puisque le nombre de phrases trouvées est petit et que les phrases contenues dans le discours sont disparates.

Si on prend pour un seul discours une accumulation de commentaires boursiers, il en émerge cependant une séquence de plusieurs catégories syntactico-sémantique définies pour le corpus boursier, désormais appelée schéma de phrase, qui représente des occurrences concrètes de phrases. D'après les résultats obtenus au terme de notre étude, un quart du corpus correspond à un seul schéma de phrase, ce qui n'est pas loin de la tentative de Harris, qui a essayé de réduire un texte entier à une succession de plusieurs séquences de catégories.

La direction donnée par Harris dans le domaine de l'analyse du discours a été poursuivie au fur et à mesure que se développe l'analyse automatique de textes d'un domaine spécifique, au moyen du formalisme de la grammaire en chaîne. Les travaux sur les textes médicaux (Hirschman et Sager, 1980) et les recherches sur les rapports boursiers ou les bulletins météo

(Kittredge, 1980) ont été effectués dans ce cadre. Ces dernières traitent du même type du phrase qu'on étudie dans cet article mais l'accent a été mis sur la façon de caractériser, de point de vue des sous-langages (*sublanguage* en anglais), les phrases dans le discours d'un domaine spécifique.

3.2. *Deux types d'informations dans le corpus*

Dans les rapports boursiers, on trouve souvent deux types d'informations, liées par une relation de cause à effet. Le premier type se trouve dans des phrases qui décrivent les mouvements en valeur d'actions particulières, observés à la Bourse. Cette information est donc spécifique à la Bourse. L'autre type relate un phénomène économique ou social qui pourrait être la cause de ces mouvements. Cette information peut être extra-boursière. En voici deux exemples types :

(1) *A 34 euros, l'action X a augmenté de 0,25 %, mercredi 26 mai, à l'ouverture de la Bourse de Paris. X a annoncé la veille le licenciement massif de trois quarts de ses employés.*

(2) *L'action X a débuté la séance du mercredi 26 mai en baisse de 1,1 %. La production du café en Amérique du Sud a subi de gros dégâts à cause d'une chaleur anormale.*

Le discours imaginé (1) est constitué de deux phrases. La première décrit le mouvement en valeur des actions d'une entreprise nommée *X* et la deuxième une activité de cette entreprise, qui est relatée ici parce qu'elle est censée avoir une influence sur la chute de valeur de l'action décrite dans la première phrase. Dans le discours également imaginé (2), la première phrase présente une similitude de forme avec celle du discours (1) et la deuxième, qui décrit un phénomène naturel, ne peut pas être comprise dans ce contexte si le lecteur n'a pas de connaissance préalable de la société *X*, qui est, par exemple, un exportateur du café.

Comme nous allons le voir, le premier type de phrase dans les deux discours peut être décrit formellement au moyen du lexique-grammaire. Mais le second type de phrase est plus difficile à schématiser formellement, parce que la phrase décrit non pas des activités boursières mais divers phénomènes dans le monde réel, impliquant divers prédicats et divers arguments.

Il existe des phrases qui appartiennent au deuxième type de phrase donné ci-dessus et qui décrivent les activités financières et commerciales d'une société comme suit :

(3) *X a (annoncé + publié + prévu) le chiffre d'affaires pour le second trimestre 2003 en (hausse + baisse) de 0,3 %, à 23 milliards d'euros.*

Elles sont susceptibles d'être décrites à partir du schéma de phrase qui sera expliqué plus loin pour analyser le premier type de phrase. Mais nous n'avons pas tenu compte de ces phrases dans cette étude.

Nous nous sommes limité à la description du premier type de phrases mentionné plus haut, (les premières phrases des discours (1) et (2)) et leurs variantes formelles.

La description syntaxique de ces phrases est basée sur le lexique-grammaire[1], qui recense pour un prédicat donné la nature et le nombre de ses compléments possibles et les propriétés syntaxiques et transformationnelles qu'elles présentent. Cette description se traduira sous forme d'une table syntaxique dont les entrées sont constituées par les prédicats et où les colonnes indiquent les propriétés distributionnelles et syntaxiques.

[1] Pour connaître le détail, voir Gross (1975 et 1981) et Leclère (2002).

3.3. *Les phrases exprimant la variation d'une valeur*

Le type de phrase traité décrit *la variation en valeur d'une action particulière cotée à la Bourse de Paris*. En voici quelques exemples types :

(4) *ABC gagnait 2,84 %, vendredi 24 août dans les premières transactions, à 18,10 euros.*
(5) *L'action ABC s'appréciait de 0,94 %, vendredi dans les premiers échanges, à 37,48 euros.*
(6) *Le titre d'ABC (a fait un bond + a bondi) de 3,29 %, à 16,78 euros, mercredi matin*
(7) *L'action du spécialiste de la carte à puce ABC restait stable dans les premières transactions, lundi 20 août, à 3,09 euros.*

L'information essentielle, dans les rapports boursiers, est syntaxiquement structurée autour de plusieurs dizaines de prédicats (verbaux, nominaux ou adjectivaux, i.e. les séquences soulignées dans les exemples ci-dessus) et leurs compléments. Diverses formes de phrases ainsi construites partagent toutes une unité de sens stable, qui est la variation de quantité. Le sens de variation est rendu syntaxiquement explicite par la présence de deux compléments numériques, l'un relatif et l'autre absolu.

Les sujets de ces prédicats sont réalisés par un ensemble de groupes nominaux qui sont principalement juxtaposés d'une manière appositive (cf. (5)-(7)) autour de noms têtes comme *action, titre* ou *valeur*. Les noms propres des sociétés dont les transferts de l'action sont rapportés dans les textes ont été répertoriés et jouent un rôle syntaxique de sujet par métonymie (cf.(4)).

Les deux compléments numériques ont un lien fort avec les verbes, malgré leur nature adverbiale. Surtout, le complément numérique relatif est indispensable pour définir un certain nombre de verbes qu'on appelle ici « verbes de variation ». Il entretient donc un lien plus étroit avec ces verbes que le complément numérique absolu.

D'autres compléments, cette fois clairement adverbiaux, de temps et de lieu, sont considérés comme des compléments spécifiques de phrase. Le complément de lieu est plus facilement omis, étant implicitement constant[2], que celui de temps. L'information du temps est évidemment cruciale dans un rapport boursier en tant qu'élément d'observation.

Par ailleurs, un même prédicat peut apparaître sous forme de variantes formelles. Plusieurs prédicats verbaux utilisés dans ce contexte peuvent subir une transformation de nominalisation au moyen de verbes supports (Gross, 1981). Ils se rangent dans le même paradigme syntaxique. L'exemple (6) montre un cas de nominalisation possible.

L'organisation linéaire des prédicats et de leurs compléments est assez régulière. La nature du corpus constitué de rapports factuels fait qu'il n'y a de phrases ni interrogatives, ni impératives. Toutes les phrases observées sont déclaratives.

Les prédicats peuvent être classés, sémantiquement, dans trois groupes : ceux qui expriment le mouvement en hausse d'une valeur, ceux qui exprime le mouvement contraire, c'est-à-dire la baisse d'une valeur, et le dernier type qui exprime la stabilité.

Le type de phrase étudié se schématise comme une suite de composants formels :

A V D% DN AT AL

Chaque composant doit être considéré comme une classe d'équivalence, à l'intérieur de laquelle toutes les variantes formelles de la classe doivent être énumérées. ***A*** regroupe des

[2] Dans notre corpus, le lieu d'observation est toujours à la Bourse de Paris.

groupes nominaux fonctionnant comme sujet des prédicats de variation, ***V*** les prédicats de variation, ***D%*** les compléments numériques relatifs et ***DN*** les compléments numériques absolus. ***AT*** représente les adverbes de temps et ***AL*** les adverbes de lieu.

4. La construction de grammaires locales

4.1. *Le système UNITEX et les grammaires locales*

Le système UNITEX (Paumier, 2003) est un système de traitement automatique de la langue naturelle qui permet d'étiqueter les textes avec les dictionnaires électroniques existants et/ou créés selon les besoins. Les grammaires locales sont des graphes d'automates à états finis[3] qui, partant du point de départ, reconnaissent les unités linguistiques qui se trouvent sur un chemin. Lorsque un chemin a été exploré et qu'on arrive au point final, la séquence a été reconnue (Gross, 1997).

Le système UNITEX permet d'appliquer les grammaires locales aux textes étiquetés lexicalement.

Nous analysons notre corpus au moyen d'un ensemble de grammaires locales (266 graphes) dont l'unité maximale correspond au schéma de phrase donné plus haut et d'un dictionnaire des noms propres créé spécialement pour analyser le corpus (416 entrées).

Certains des composants du schéma de phrase sont exposés ci-dessous, soit sous forme d'une grammaire locale, soit sous forme d'une expression rationnelle.

4.2. *Le composant* **A**

Le constituant *A* du schéma de phrase ci-dessus représente les groupes nominaux suivant :

(L'action + le titre + la valeur) ((de + E) Npr + de Ns Npr + E)[4]

Les opérations d'effacement opérées sur les séquences (8) permettent d'obtenir la classe d'équivalence de groupes nominaux suivante. Ils sont co-référents dans le contexte en question :

(9) *(l'action + le titre + la valeur) de [la société de construction automobile]*$_{Ns}$ *[ABC]*$_{Npr}$
= (l'action + le titre + la valeur) de [la société de construction automobile]$_{Ns}$
= (l'action + le titre + la valeur) (de + E) [ABC]$_{Npr}$

Les groupes nominaux (9) sont ultimement réductibles aux groupes nominaux suivants :

(10) *((l'action + le titre + la valeur) + ABC)* (= (9))

Ces séquences sont donc considérées comme produites à partir de la séquence la plus longue en y opérant les effacements. Ce genre de faits linguistiques est facilement exprimé au moyen d'expressions régulières comme (8) ou (9), ou de façon encore plus lisible, de grammaires locales (Figure 1). Les noms propres de sociétés sont recensés dans un dictionnaire électronique (Figure 2).

[3] Cette équivalence peut être vraie dans le domaine du traitement automatique de la langue naturelle.

[4] *Npr* désigne ici les noms propres. *Ns* désigne un ensemble de groupes nominaux comportant une sous-classification des activités. Ex : *société de production audiovisuelle, entreprise de fournitures de bureau*, etc. *E* représente l'élément vide. ''+'' représente un choix dans un paradigme.

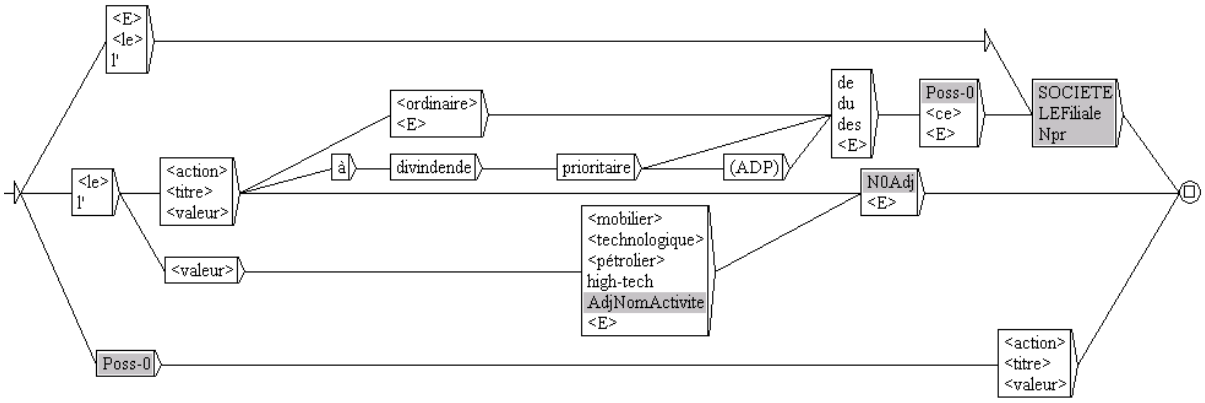

*Figure 1. Grammaire locale **A***
La flèche qui se trouve à gauche montre le point de départ et le carré à l'intérieur d'un cercle à droite est un point final. Les boîtes grises sont un appel à sous-graphe.

```
138 Eridiana Béghin-Say,.N+propre
139 Essilor,.N+propre
140 Etam Développement,.N+propre
141 Eurofins Scientific,.N+propre
142 Euronext,.N+propre
143 Euronext Paris,.N+propre
144 Europ@web,.N+propre
145 Eurosport France,.N+propre
146 Eurosport International,.N+propre
147 Eurotunnel,.N+propre
148 Executive Life,.N+propre
```

Figure 2. Dictionnaire de noms propres de sociétés

4.3. *Le composant V*

Le constituant *V* regroupe ici toutes les variantes formelles des prédicats de variation. La catégorie se subdivise en plusieurs classes. Si le prédicat est verbal, la transitivité par rapport au complément de mesure relative est un critère de subdivision. Si le prédicat est nominal, plusieurs sous-classes sont créés en fonction des verbes supports choisis.

Les verbes transitifs qui prennent comme objet direct le complément de mesure relative sont donnés ci-dessous. Cette propriété est marquée dans la colonne *NVDnum%* de la table (cf. Annexe). Ils forment la classe *Vvt* :

(11) *gagner, perdre, prendre, s'adjuger, céder, abandonner, regagner*

Avec les verbes intransitifs, le complément de mesure relative apparaît précédé de la préposition *de* :

(12) *progresser, chuter, baisser, dégringoler, bondir…*

Ces verbes forment la classe *Vvi* et la propriété figure sous la colonne *NVDeDnum%* dans la table.

Les prédicats peuvent être nominalisés. Selon les verbes supports, nous pouvons avoir dans la classe *V* les combinaisons de *Vsup* et *Npred* suivants :

(13) *[(accuser + enregistrer + connaître +…)]$_{Vsup}$ UN (progression + baisse) de D%*

(14) *[(pointer + être + s'échanger +...) en]* $_{Vsup}$ *(augmentation + baisse) de D%*

Les verbes supports transitifs (*Vsupt*) sont exemplifiés en (13) et les verbes supports avec la préposition *en* (*VsupEn*) sont donnés en (14). Les propriétés sont représentées dans les colonnes *N Vsupt Dét V-n De Dnum %* et *VsupEn en Vvn De Dnum %* respectivement.

4.4. *Le composant D%*

D% regroupe les expressions de pourcentage, c'est-à-dire le complément de mesure relative, direct ou indirect.

(15) *(de + E) Dnum (% + pour cent) = (de + E) 15 (% + pour cent)*

4.5. *Le composant DN*

Le composant *DN* est un complément de mesure absolue. Il comporte un déterminant numérique et un nom d'unité monétaire. Il spécifie la valeur des actions échangées à la Bourse.

(16) *Dnum Nunité = 45 euros*

4.6. *Le composant AT*

Les compléments de temps ponctuels qu'on peut trouver dans le corpus sont construits autour des noms d'activités boursières, avec des noms temporels comme *début, milieu*. En voici des exemples :

(17) *(dans les premièr(e)(s) + au début de LE + au cours de LE) (transactions + échanges + cotation)*

Les groupes nominaux prépositionnels comme (17) se comportent comme des adverbes. Mais il n'est pas souhaitable de les étiqueter comme adverbe dans le dictionnaire, parce qu'ils sont composés de locutions prépositives figées donnant le sens de temps et de noms d'activité, qui peuvent varier selon les contextes. Dans ce genre de situation, la construction de la grammaire locale énumérant les noms qui peuvent apparaître dans un corpus particulier est utile.

Des compléments de temps généraux peuvent aussi apparaître dans le corpus. Par exemple :

(18) *dans la matinée du mardi 10 octobre = le mardi 10 dans la matinée*
le mardi 10 = le mardi

Ces différents compléments de temps sont symbolisés par *AT*.

4.7. *Le composant AL*

Les compléments de lieu, symbolisés par *AL,* montrent plusieurs variantes de formes, à cause de la troncation. Dans le corpus, toutes les séquences suivantes sont considérées comme équivalentes :

(19) *à la Bourse de Paris, à Paris, en Bourse, à la Bourse*

4.8. *Le graphe principal*

Pour mener à bien la reconnaissance des phrases dans le corpus, il faut linéaliser les composants décrits plus haut. On appelle graphe principal la grammaire locale qui contient tous les sous-graphes. Les composants *AT, AL, D%* et *DN*, de nature adverbiale, peuvent subir une permutation au sein d'une même phrase, ce qui multiplie le nombre de chemins dans la grammaire locale. Voici le graphe principal appliqué lors du parsing du corpus :

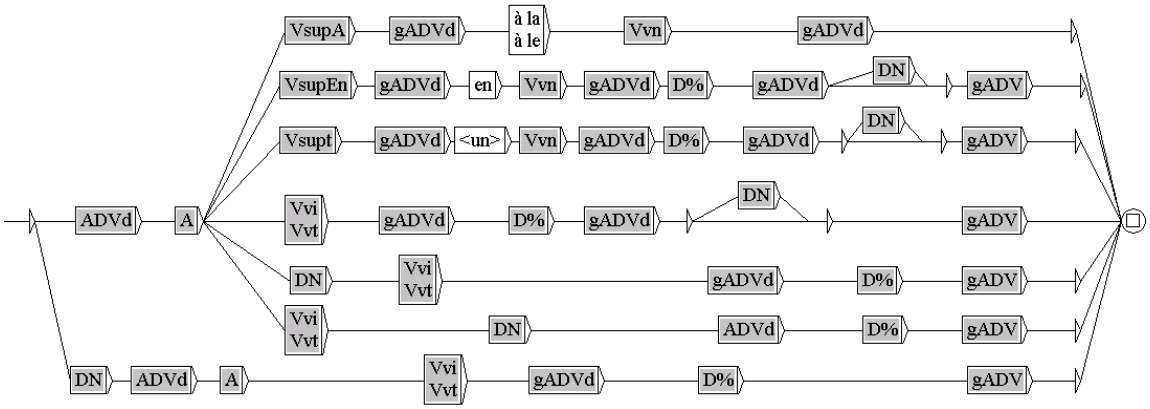

Figure 3. Le graphe principal[5]

Le nombre des sous-graphes dans ce graphe principal est d'à peu près 270[6]. Avec ces grammaires locales, toutes les phrases, y compris celles qui n'ont jamais été rencontrées au cours de la construction de ces grammaires apparaissant dans le corpus étudié qui sont linguistiquement analysables selon le schéma de phrase donné en haut ont été entièrement reconnues sans ambiguïté. Les phrases concernées représentent environ 22 % du corpus total.

Il faut souligner, cependant, le fait que ce résultat a été obtenu en fonction d'un corpus choisi d'avance et que les grammaires locales ont été également conçues spécifiquement pour être appliqué à ce corpus. Il n'est donc pas étonnant que leur application à d'autres corpus, même s'il est de la même nature que celui qu'on a utilisé, ne donne pas, d'emblée, un résultat aussi fin. Or nous pensons que des ajustements mineurs (cf. la différence du style) remettent assez rapidement en marche les grammaires locales conçues ici, puisque les grammaires locales sont modulaires, et qu'il est possible d'atteindre un taux de reconnaissance semblable, du fait que l'information communiquée est de même nature et que nous pensons que la variation de formes linguistiques qui correspond à cette information doit être assez limitée.

Le résultat de l'application de ce graphe au corpus donne une concordance, dont un échantillon est donné ci-dessous (Figure 4) :

{S} L'action Sopra gagnait 5,70 %, à 67,90 euros, jeudi matin.{S} De son côté, le titre Cap Gemini Ernst & Young pro
{S} L'action ST Microelectronics a cédé 3,13 %, jeudi matin, à 37,09 euros.{S} Le titre Thomson Multimédia a perdu 3
{S} L'action STMicroelectronics a débuté la séance de jeudi en forte baisse, perdant 5 % à l'ouverture, à 47,50 euro
{S} L'action Stmicroelectronics gagnait 2,77 %, jeudi 12 juillet dans les premiers échanges, à 33,40 euros.{S} Le gr
{S} L'action Suez Lyonnaise des Eaux était en hausse de 3,45 %, à 158,8 euros.{S} Le groupe de services collectifs,
{S} L'action Suez Lyonnaise gagnait 0,17 %, à 179 euros, mardi matin.{S} Le Financial Times de mardi révèle que ST 3
{S} L'action Suez perdait 0,24 %, à 37,38 euros, mardi 7 août à l'ouverture de la Bourse de Paris.{S} Le groupe a dé
{S} L'action Suez s'appréciait de 0,94 %, vendredi dans les premiers échanges, à 37,48 euros.{S} Jeudi, des rumeurs
{S} L'action Technip chutait de 10,33 %, à 148,5 euros, jeudi 28 juin dans les premiers échanges, et le titre Coflex
{S} L'action TF1 gagnait 1,73 %, à 55,95 euros, mercredi 31 janvier.{S} Le groupe a enregistré un résultat net provi
{S} L'action TF1 grimpait de 5,31 %, jeudi, dans les premières transactions, à 34,11 euros.{S} Les opérateurs ont ét
{S} L'action TF1 perdait 0,62 %, à 45,79 euros, vendredi matin.{S} La société a annoncé une hausse de 1,7 % de ses r
{S} L'action TF1 s'appréciait de 1,72 % mardi matin, à 41,3 euros.{S} Le titre figure dans une liste, publiée par le
{S} L'action Thales gagnait lundi matin 1,05 %, à 43,45 euros. {S} Le groupe de cosmétiques Clarins doit publier son
{S} L'action Thales s'inscrivait en hausse de 1,43 %, mardi matin, à 44,82 euros.{S} Le groupe de électronique de dé
{S} L'action Thomson Multimédia grimpait de 4,2 %, lundi matin, à 51,6 euros.{S} Le groupe a enregistré en 2000 un b
{S} L'action Thomson Multimédia s'échangeait en baisse de 5,43 % à 40,38 euros.{S} Le groupe a publié de bons résult
{S} L'action TotalFinaElf gagnait 1,14 %, à 177 euros.{S} Le groupe pétrolier a obtenu 30 % de un contrat pour la va
{S} L'action Transgène cédait 0,86 %, lundi matin, à 11,58 euros.{S} Les 63 millions de euros levés dans le cadre de
{S} L'action Tredi Environnement reculait de 2,47 %, vendredi dans les premières transactions, à 39 euros.{S} Le spé

Figure 4. Concordance de la reconnaissance

[5] Une boîte grisée appelle un sous-graphe.
[6] Un sous-graphe peut appeler à son tour un autre sous-graphe.

La partie soulignée correspond à une séquence reconnue par les grammaires locales. La lettre *S* majuscule entre parenthèses signifiant le commencement et la fin d'une phrase, les séquences analysées de la Figure 4 correspondent à des séquences reconnues comme des phrases par *UNITEX*.

5. Conclusion et perspectives

22 % des phrases du corpus sont analysées par le schéma de phrase donné. Le résultat de la reconnaissance automatique n'a pas d'ambiguïté. Cela montre l'utilité d'une analyse linguistique, syntaxico-sémantique et lexicale, préalable à la construction de grammaires locales. Cela est nécessaire pour mener à bien l'analyse automatique de ce type de texte. La technique linguistique de réduction des variantes formelles de phrase à un schéma de phrase par rapport à un discours particulier est une aide pour la création de grammaires locales optimales.

La continuation de ce travail doit tenir compte de plusieurs tâches :

1. pouvoir traiter d'autres types de schémas de phrase, e.g. formaliser le deuxième type de phrase apparaissant dans les rapports boursiers[7] (cf. 3.2.)
2. pouvoir traiter les phrases complexes, e.g. le schéma de phrase étudié ici peut être réutilisé pour décrire par exemple les compléments phrastiques de verbes de parole : *X a annoncé une baisse de 2,1 % de son chiffre d'affaires, à 34 milliards d'euros :* la partie soulignée dans cette phrase peut être formalisée avec les composants que nous avons fixés, complétés par une nouvelle unité, *chiffre d'affaires*.

Cette méthode peut être appliquée à des discours d'autres natures : bulletin météo, résultats sportifs, comptes rendus techniques, etc. On peut cependant prévoir que l'application subirait des contraintes pareilles à celles rencontrées ici, c'est-à-dire que certaines phrases sont faciles à décrire dans des schémas et d'autres sont plus problématiques. Pour le moment, l'objet de description et les résultats sont limités mais nous pensons que l'accumulation de descriptions linguistiques et de grammaires locales, tournées vers des corpus concrets, diminuera ces contraintes.

Références

Gross M. (1975). *Méthodes en syntaxe.* Hermann.

Gross M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages*, vol. (63) : 7-52.

Gross M. (1996). Construction de grammaires locales et automates finis. *Working Papers*, vol. (5). Centro linguistico, Universita' commerciale ''L. Bocconi''.

Gross M. (1997). The Construction of Local Grammars. In Roche E. et Schabes Y. (Eds), *Finite State Language Processing.* The MIT Press.

Harris Z. (1952). Discourse Analysis. *Language*, vol. (28).

Harris Z. (1963). *Discourse Analysis Reprints.* Mouton & Co.

Hirschman L. et Sager N. (1982). Automatique Information Formatting of a Medical Sublanguage. In Chapitre 2 de Kittredge R. et Lehrberger J. (Eds), *Sublanguage : Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

[7] De ces phrases disparates sommairement catégorisées dans ce deuxième type de phrase semble apparaître une échelle de degré de « décrivabilité » de phrases, mesuré en fonction de la nature de l'information. Si la phrase rapporte des événements boursiers, financiers ou commerciaux, elle est plus facile à décrire que la phrase qui rapporte un phénomène naturel, par exemple.

Kittredge R. (1982). Variation and Homogeneity of Sublanguages. In Chapitre 4 de Kittredge R. et Lehrberger J. (Eds), *Sublanguage : Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

Leclère Chr. (2002). Organization of the lexicon-grammar of French verbs. *Lingvisticae Investigationes*, vol. (25 : 2).

Paumier S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse, Université de Marne-la-Vallée, IGM.

Annexe : Extrait de la table du lexique-grammaire spécialisé[8]

A	**Prédicats de variation**		**Vvi**		**Vvt**	**VsupEn en Vvn De Dnum %**						**Vsup à Dét V-n**						**Vsupt Dét V-n De Dnum %**				
N0 = ACTION	Vv ou Vsup Adj	Vvn	NVDeDnum%	NVDnum%	NVDnum%DePoss0Valeur	être	s'échanger	s'inscrire	pointer	se négocier	Vtps	être réservé	être orienté	s'orienter	reprendre	repartir	pointer	faire un V-n	enregistrer UN V-n	connaître UN V-n	afficher UN V-n	accuser UN V-n
+	apprécier(s')	(appréciation)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	augmenter	augmentation	+	–	–	+	?	?	?	?	+	?	?	?	?	?	?	–	?	?	?	?
+	baisser	baisse	+	–	–	+	+	+	+	+	+	+	?	?	?	?	?	–	+	?	?	?
+	bondir	bond	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	?	–	?	?
+	chuter	chute	+	–	–	?	?	?	?	?	–	–	–	–	–	–	–	?	?	?	?	?
+	décrocher	(décrochage)	+	–	–	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
+	dégringoler	(dégringolade)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	–	–	–
+	déprécier(se)	(dépréciation)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	effondrer(s')	(effondrement)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	effriter(s')	(effritement)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	envoler(s')	(envolée)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	glisser	(glissement)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	–	–	–

[8] Dans cette table du lexique-grammaire spécialisé, le signe ''+'' signifie que l'expression en question est trouvée dans le corpus. Le signe ''–'' montre le cas contraire. Le signe d'interrogation ''?'' signifie que l'expression en question n'est pas agrammaticale, mais qu'elle n'a pas apparu dans le corpus.

A	Prédicats de variation		Vvi		Vvt	VsupEn en Vvn De Dnum %						Vsup à Dét V-n						Vsupt Dét V-n De Dnum %				
N0 = ACTION	Vv ou Vsup Adj	Vvn	NVDeDnum%	NVDnum%	NVDnum%DePoss0Valeur	être	s'échanger	s'inscrire	pointer	se négocier	Vtps	être réservé	être orienté	s'orienter	reprendre	repartir	pointer	faire un V-n	enregistrer UN V-n	connaître UN V-n	afficher UN V-n	accuser UN V-n
+	grimper	(grimpée)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	progresser	progression	+	–	–	+	?	?	?	?	+	?	?	?	?	?	?	–	+	?	?	?
+	rebondir	(rebond)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?	?
+	reculer	recul	+	–	–	+	?	?	?	?	?	–	–	–	–	–	–	–	+	?	+	+
+	replier(se)	repli	+	–	–	+	?	?	?	+	+	?	?	?	?	?	?	?	?	?	+	?
+	reprendre(se)	(reprise)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?	?
+	stable(être + rester)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+		hausse	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	?	+	+	+	?
+	abandonner	(abandon)	–	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	adjuger(s')		–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	céder	(cession)	–	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	gagner	gain	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	+	?
+	perdre	(perte)	–	+	+	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?	?
+	prendre		–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	regagner	(regain)	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?