TALN 2010, Montréal, 19-23 juillet 2010

Extraction de paraphrases sémantiques et lexico-syntaxiques de corpus parallèles bilingues

Jasmina Milićević

French Department, Dalhousie University
6135 University Avenue, Halifax (N.S.) CANADA B3H 4P9
jmilicev@dal.ca

Résumé

Nous présentons le travail en cours effectué dans le cadre d'un projet d'extraction de paraphrases à partir de textes parallèles bilingues. Nous identifions des paraphrases sémantiques et lexico-syntaxiques, qui mettent en jeu des opérations relativement complexes sur les structures sémantiques et syntaxiques de phrases, et les décrivons au moyen de règles de paraphrasage de type Sens-Texte, utilisables dans diverses applications de TALN.

Abstract

We present work in progress done within a project of extracting paraphrases from parallel bilingual texts. We identify semantic and lexical-syntactic paraphrases, which imply relatively complex operations on semantic and syntactic structures of sentences, and describe them by means of paraphrasing rules of Meaning-Text type, usable in various NLP applications.

Mots-clés : paraphrase lexico-syntaxique, paraphrase sémanique, règles de paraphrasage, corpus bilingues, théorie Sens-Texte

Keywords: lexical-syntactic paraphrase, semantic paraphrase, paraphrasing rules, bilingual corpora, Meaning-Text theory

1 Problématique et cadre théorique

Cet article présente le travail en cours effectué dans le cadre d'un projet visant l'extraction de règles de paraphrasage à partir de textes parallèles bilingues. Nous analysons des textes littéraires et journalistiques et leurs traductions : l'espagnol/le français vers l'anglais et l'espagnol/l'anglais/le russe vers le français. La taille totale du corpus analysé est de 100 000 mots environ. Pour le moment, l'extraction de règles se fait manuellement, mais nous prévoyons de recourir dans une phase ultérieure à des techniques automatiques d'analyse du corpus, ainsi qu'à l'élargissement de ce dernier.

Le cadre théorique que nous adoptons est la théorie linguistique Sens-Texte (Mel'čuk 1974, Kahane 2003), notamment son système de paraphrasage (Žolkovskij & Mel'čuk 1967, Mel'čuk 1974 et 1992, Milićević 2007a/b), qui a été utilisé dans diverses applications de TALN (voir, par exemple, Iordanskaja *et al.* 1991, Lavoie *et al.* 2000, Kittredge 2002 et Apresjan *et al.* 2003).

Notre démarche repose sur le postulat que la traduction est un cas particulier de paraphrasage – le paraphrasage INTERlinguistique – qui utilise essentiellement les mêmes procédés que le paraphrasage INTRAlinguistique, se laissant décrire par les mêmes règles[1]. Nous considérons des paires {$\text{Phrase}_{\text{L-SOURCE}}$, $\text{Phrase}_{\text{L-CIBLE}}$} au sein d'un corpus bilingue comme des équivalents paraphrastiques potentiels, et c'est à partir de telles paires que nous extrayons des règles de paraphrasage ; c'est la même approche que celle adoptée, par exemple, dans Dorr *et al.* (2004). Une méthode différente consiste à extraire d'un corpus bilingue des paraphrases intralinguistiques en se servant d'une des langues comme du « pivot » pour l'extraction (c'est ce que font, entre autres, Bannard & Callison-Burch 2005, Zhao *et al.* 2009)[2] ou encore à utiliser des traductions multiples d'un texte en tant que source de paraphrases intralinguistiques (comme, par exemple, Barzilay & McKeown 2001).

Nous nous concentrons sur le repérage et la description de liens paraphrastiques complexes, qui impliquent des modifications importantes soit de l'organisation lexicale et syntaxique de la phrase de départ soit du sens de cette dernière. (Ceci explique notre intérêt pour la traduction littéraire, où les interventions de ce type devraient être courantes.) De tels liens se décrivent dans notre approche au moyen de deux types majeurs de règles.

• Les règles de paraphrasage lexico-syntaxiques spécifient des (quasi-)équivalences entre lexies considérées au sein de constructions syntaxiques. Trois règles de ce type, sous-jacentes aux paraphrases (1a-b), sont données en (1c). (Les règles opèrent entre fragments de structures syntaxiques, qui sont des arbres de dépendance. $\texttt{Oper}_1$, $\texttt{A}_1$ et $\texttt{A}_2$ sont des noms de fonctions lexicales : le verbe support d'un type particulier, l'adjectif déverbatif actif et l'adjectif déverbatif passif ; **II** et **ATTR** sont des noms de relations syntaxiques particulières.)

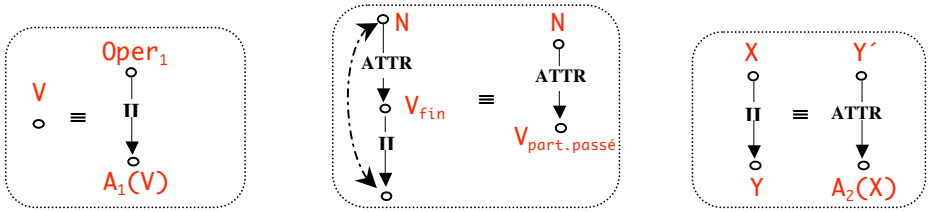
(1) espagnol ~ anglais
a. *Igualmente **asombran los Coranes que se exhiben**, …*
'Également étonnent les Corans qui s'exposent, …'
b. *Equally **astonishing is a display of Korans**, …*
'Également étonnant est l'exposition de Corans, …'

---

[1] Bien entendu, l'équivalence paraphrastique n'est pas la même chose que l'équivalence de traduction : certaines équivalences de traduction ne sont pas de paraphrases mutuelles et, dans un contexte particulier, certaines paraphrases peuvent ne pas être des équivalents de traduction. Cependant, dans la plupart des cas, les deux types d'équivalence coïncident.

[2] *Grosso modo*, on cherche dans la $\text{langue}_{\text{CIBLE}}$ tous les alignements avec une expression E (= le pivot) de la $\text{langue}_{\text{SOURCE}}$, ses alignements étant considérés comme des équivalents paraphrastiques intra-linguistiques potentiels.

c. Règle[LEX-SYNT-1] Règle[LEX-SYNT-2] Règle[LEX-SYNT-3]



La *Règle*[LEX-SYNT-1] (une substitution synonymique avec introduction d'un verbe support) rend compte du lien paraphrastique X *étonne*$_{V}$ [Y par Z(X)] ≈ X *est*$_{Oper1}$ *étonnant*$_{A1(V)}$ [pour Y à cause de X de Z] ; la *Règle*[LEX-SYN-2] (la translation « proposition relative ~ participe passé ») permet de passer de *Corans*$_{N}$ *qui*$_{N}$ *s'exposent*$_{Vfin}$ à *Corans*$_{N}$ *exposés*$_{Vpar.passé}$; finalement, la *Règle*[LEX-SYN-3] (une inversion de subordination) fait le passage entre *Corans*$_{Y}$ *exposés*$_{A2(X)}$ et *exposition*$_{X}$ *de Corans*$_{Y}$.

• Les règles de paraphrasage sémantiques décrivent des (quasi-)équivalences entre configurations de sémantèmes ; elles peuvent impliquer la décomposition de sens lexicaux (nécessaire pour « mettre à nu » l'équivalence entre expressions linguistiques), la neutralisation sémantique (« extinction » contextuelle de différences sémantiques entre expressions quasi-synonymes ou même non-synonymes, voir, par exemple, Milićević 2007a : 119*sqq*), certaines connaissances pragmatiques ou des inférences logiques. Deux règles sémantiques de paraphrasage et les paraphrases qu'elles relient suivent. (Les règles sont présentées de façon informelle ; en réalité, elles opèrent entre fragments de réseaux sémantiques, comme on le verra à la section suivante).

(2) russe ~ français
a. ***Takix romanov*** *nynče* ***net***.
'De.tels romans aujourd'hui il.n'y.a.pas.'
b. ***Où trouver*** *à présent* ***un roman de telle espèce*** *?*
c. Règle[SÉM-1]: 'les X n'existent plus' ≈ 'trouver une instance d'X est impossible'

La *Règle*[SÉM-1] se base sur l'inférence triviale (faisant partie de nos connaissances pragmatiques de tous les jours) 'inexistant' ⇒ 'introuvable'. Cette règle est intéressante en ce qu'elle met en relation deux types d'actes de parole différents : la déclaration et la question rhétorique (en tant qu'une des réalisations possibles de la partie droite de la règle).

(3) français ~ anglais
a. *Tout à coup, une* ***voix*** *lointaine* [qu'il **entendit**] [...] ***fit pâlir*** *le Tétrarque.*
b. *Suddenly he* ***heard*** *the sound of a distant* ***voice*** [...]. *His cheek* ***paled***.
'Soudainement, il entendit le son d'une voix lointaine [...]. Sa joue pâlit.
c. Règle[SÉM-2]: 'P cause#1 Q' ≈ 'P ; (par conséquent,) Q[3]'

La *Règle*[SÉM-2] établit la quasi-équivalence entre la causation exprimée explicitement, au sein d'une proposition simple, au moyen du verbe FAIRE, et la conséquence exprimée implicitement – par une coordination asyndétique (= sans conjonction) de deux propositions simples. (Il s'agit en fait d'un combiné de deux règles sémantiques de paraphrasage de nature différente ; voir la section 2.)

[3] Le sémantème 'causer#1' représente la causation non-agentive ; son actant sémantique 1 est la Cause. Il s'oppose au sémantème 'causer#2', qui représente la causation agentive et a le Causateur comme son actant sémantique 1; voir Kahane & Mel'čuk (2006).

Nous voulons surtout repérer et décrire des règles sémantiques de paraphrasage, qui sont beaucoup moins connues, mais nous sommes également à la recherche de nouvelles règles lexico-syntaxiques.

Les paraphrases que peuvent prendre en charge les règles de type Sens-Texte sont, de façon générale, plus complexes que la plupart des paraphrases traitées à date dans le cadre du TALN[4]. Sur les perspectives d'automatisation de nos règles et leur utilisabilité dans le cadre du TALN, voir la Conclusion.

2 Quelques règles de paraphrasage extraites manuellement de notre corpus

Dans l'ensemble de notre corpus, nous avons constaté l'application assez fréquente de règles de paraphrasage lexico-syntaxiques. Pour la plupart, il s'agit des règles déjà connues, ce qui n'est pas étonnant étant donné la large couverture du système de paraphrasage lexico-syntaxique existant (le nombre de nouvelles règles de ce type était autour d'une dizaine). Pour ce qui est des règles sémantiques, dont la plupart sont nouvelles, leur application semble plus fréquente dans les textes parallèles impliquant les langues typologiquement plus éloignées — comme on pouvait s'y attendre. Cependant, le nombre total de nouvelles règles sémantiques de paraphrasage (une vingtaine) est inférieur à nos attentes. Souvent, une paire de paraphrases met en jeu plusieurs règles de paraphrasage, qui peuvent être de types mixtes.

Nous présentons ci-dessous cinq nouvelles règles de paraphrasage dégagées de notre corpus : quatre règles sémantiques et une règle lexico-syntaxique.

2.1 Règles sémantiques de paraphrasage

La plupart des règles sémantiques de paraphrasage trouvées jusqu'ici mettent en jeu ce qu'on peut appeler les « concepts logiques reflétés en langue » – la causation, la conséquence, la condition, l'hypothèse, etc. Bien des règles concernent l'organisation discursive – « découpage » du texte en propositions, liens communicatifs et rhétoriques entre propositions – et peuvent s'appliquer au-delà des frontières d'une seule phrase (pour des règles traitant de tels cas, voir, notamment, Danlos 2006) ; certaines mettent en jeu des connaissances pragmatiques ou des inférences (comme la Règle$^{\text{SÉM-1}}$, présentée à la section 1).

CAUSE ~ CONSÉQUENCE

(4) français ~ russe

a. *Mais la conversation* ***s'anima***$_{\text{Q}}$ ***grâce au*** *champagne* [= grâce à l'**apparition**$_{\text{P}}$ du champagne] *et bientôt tout le monde y prit part.*

b. *No šampanskoe* ***javilos***$_{\text{P}}$ ***´,*** *razgovor* ***oživilsja***$_{\text{Q}}$ *i vse prinjali v nem učastie.*
'Mais le champagne **apparut**, la conversation **s'anima** et tous y prirent part.'

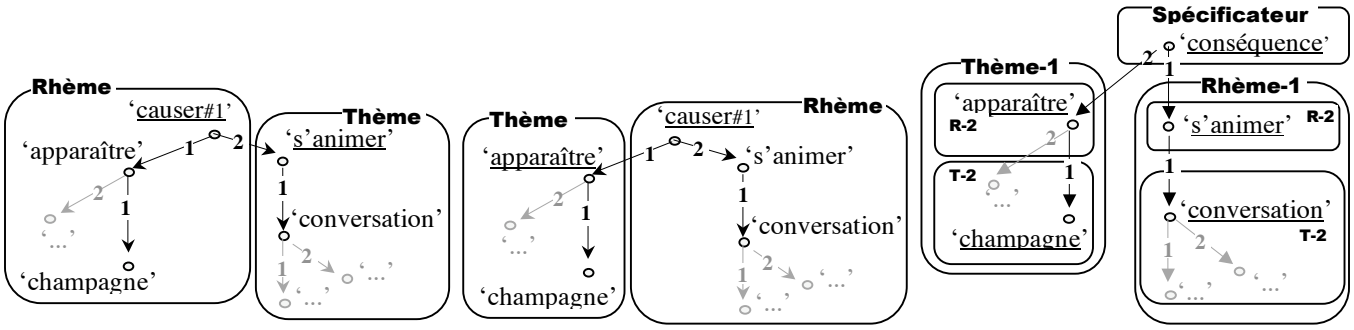
[4] Dorr *et al.* (2004) décrit un projet d'extraction automatique de paraphrases mettant en jeu des liens paraphrastiques assez complexes, comparables à ceux que peuvent traiter nos règles de paraphrasage sémantiques ; par exemple: « clause subordination vs. anaphorically linked sentences » (*This is Joe's new car, which he bought in NY. ~ This is Joe's new car. He bought it in NY.*), « different sentence types » (*Who did this ? ~ Tell me who did this.*), « inverse relationship » (*Only 20% of the participants arrived on time. ~ Most of the participants arrived late.*) et « inference » (*The tight end caugth the ball in the end zone. ~ The tight end scored a touchdown.*). À titre de comparaison, les liens les plus complexes mentionnés dans Zhao *et al.* (2009), qui est représentatif de ce que peuvent faire les approches statistiques, sont de l'ordre lexico-syntaxique ; par exemple, « phrase replacement » (*to take steps ~ to adopt measures*) et « structural paraphrase » (*the N's residents ~ the inhabitants of N*).

Les phrases en (4) affichent le lien paraphrastique « causation (explicite) ~ conséquence (implicite) », c'est donc le même lien que celui unissant les paraphrases en (3) ci-dessus. Ces deux paires de paraphrases ne se distinguent que par la réalisation du sémantème de causation : en (4a) il est exprimé au moyen de la préposition causale GRÂCE À et en (3a) au moyen du verbe causatif FAIRE. Cela veut dire que la partie gauche de règle citée en (1c) a deux variantes (conditionnées par des facteurs communicatifs, voir ci-dessous) – 'P cause#1 Q' et 'Q à cause de P' – dont les réalisations sont des paraphrases mutuelles [5]. Comparer la phrase (4a), qui réalise la variante 'Q à cause de P', avec la phrase (4c), qui, elle, est une réalisation de la variante 'P cause#1 Q' :

c. *Mais le champagne* [qui **apparut**$_P$] ***anima*** [= 'causa#1 s'animer$_Q$'] *la conversation et bientôt tout le monde y prit part*[6].

Donnons, pour fixer les idées, les représentations sémantiques des phrases en (4) ; seuls les fragments pertinents pour le paraphrasage sont représentés.

La représentation sémantique [= RSém] d'une phrase est (*grosso modo*) un appariement de deux structures. La structure sémantique [= SSém], formellement un réseau sémantique, spécifie les sémantèmes sous-jacents à la phrase et les relations « prédicat ~ argument » entre eux. La structure sémantico-communicative [= SSémCom] « tourne » la SSém en un message spécifique, en marquant des sous-réseaux de cette dernière par des valeurs d'oppositions communicatives (Mel'čuk 2001), telles que thème ~ rhème ~ spécificateur ; donné ~ nouveau, focalisé ~ neutre, etc.



RSém 1, phrase (4a) RSém 2, phrase (4c) RSém 3, phrase (4b)

Les sémantèmes 'P causer#1 Q (par W)' et 'Q (être la) conséquence (de) P' ont la distribution converse d'actants ; ils sont des conversifs approximatifs puisque 'causer#1' a un actant de plus : l'actant W, qui est optionnel et représente l'élaboration de la cause.

Le soulignement identifie le nœud communicativement dominant d'un sous-réseau (Mel'čuk 2001 : 31), c'est-à-dire le nœud qui porte l'essentiel de l'information contenue dans ce dernier. Un sous-réseau articulé en thème et en rhème secondaires (**T$_2$**, **R$_2$**) correspond à une proposition syntaxique (= une expression ayant comme tête un verbe fini).

---

[5] Il est également possible d'avoir des variantes communicativement conditionnées pour la partie droite de la règle en (1c), par exemple, 'Comme conséquence (de P), Q' ; cependant, nous ne les considérerons pas, faute d'espace.

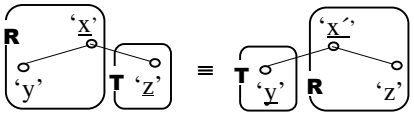
[6] En (4a) et (4c), on voit l'ellipse d'une partie de la proposition P, qui, en (4c), est accompagnée d'une inversion de subordination : *grâce à l'apparition* [P] *du champagne* ⇒ *grâce au champagne* ; *le champagne qui apparut* [P] ⇒ *le champagne*. On observe le même phénomène en (3a). De tels effacements sont effectués par des règles spéciales, que nous ne présentons pas ici.

Le passage de la RSém 1/2 à la RSém 3 implique les « transformations » de deux types différents, modélisées dans notre approche au moyen de règles de deux types :

• Un remplacement de sémantèmes (‘cause’ ~ ‘conséquence’), qui permet de passer des SSém 1/2 à la SSém 3 (et vice-versa) ; il est effectué par une règle de quasi-équivalence propositionnelle (cette règle a été présentée pour la première fois dans Milićević 2007a : 195).

• Deux restructurations communicatives, qui permettent de passer de la SSém-Comm 1 à la SSém-Comm 2, puis de celle-ci à la SSém-Comm 3 ; elles sont prises en charge par deux règles de quasi-équivalence communicative.

Règle SÉM (Communicative)-3

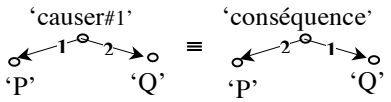


La situation changea#I.1a$_{z}$ *grâce à*$_{x}$ *son intervention*$_{y}$. ≈ *Son intervention*$_{y}$ *changea*#I.1b$_{x+z}$ *la situation* [= causa#I que la situation change#I.1a].

J’ai raté$_{z}$ *la correspondance à cause du*$_{x}$ *retard*$_{y}$ *du <qu’a eu le> premier train. ≈ Le retard*$_{y}$ *du <qu’a eu le> premier train m’a fait*$_{x}$ *rater*$_{z}$ *la correspondance.*

La règle ci-dessus établit la quasi-équivalence entre les SSémComm des RSém 1 et 2, qui ont des distributions assez similaires du matériau propositionnel entre les sous réseaux thématique et rhématique ; notamment, le nœud communicativement dominant du rhème est le même dans les deux cas. Les deux règles suivantes assurent ensemble le passage vers la RSém 3.

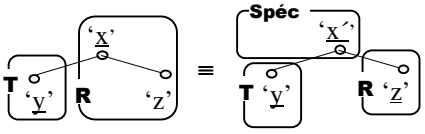
Règle SÉM (Propositionnelle)-4



Son intervention$_{y}$ [= le fait qu’il intervint] *changea*#I.1b$_{X=Z}$ *la situation*. ≈
(a) *Il intervint*$_{y}$ *et, par conséquent*$_{x}$, *la situation changea*#I.1a$_{z}$.
(b) *Il intervint*$_{y}$. *La situation changea*#I.1a$_{z}$.

Le retard$_{y}$ *du <qu’a eu> le premier train m’a fait*$_{x}$ *rater*$_{z}$ *la correspondance*. ≈
(a) *Le premier train a eu du retard*$_{y}$ *si bien que*$_{x}$ *j’ai raté*$_{z}$ *la correspondance.*
(b) *Le premier train a eu du retard*$_{y}$, *j’ai raté*$_{z}$ *la correspondance.*

Règle SÉM (Propositionnelle)-5



Son comportement$_{y}$ *bizarre m’a irrité*$_{x+z}$ ≈
(a) *Il s’est comporté*$_{y}$ *de façon bizare alors*$_{x}$, *j’ai été irrité*$_{z}$.
(b) *Il s’est comporté*$_{y}$ *de façon bizarre ; j’ai été irrité*$_{y}$.

Comme on peut le constater, nos règles permettent à la fois l’expression explicite et implicite de la conséquence : nous croyons que le choix entre les deux variantes doit se faire par une règle stylistique assez générale, qui, en outre, prendrait en charge d’autres relations discursives admettant les deux types d’expression[7].

Nos règles de paraphrasage ne spécifient pas non plus les réalisations de sémantèmes ; celles-ci sont prises en charge par les règles de lexicalisation, qui exploitent l’information consignée dans les articles de dictionnaires de lexies correspondantes. Par exemple, la règle de lexicalisation du sémantème ‘Q est une conséquence de P’ doit spécifier, pour le cas de réalisation implicite de celui-ci, les informations suivantes. 1) L’ordre linéaire : les propositions syntaxiques réalisant ‘P’ et ‘Q’ doivent toujours apparaître exactement dans cet ordre, car, en absence d’un marqueur explicite, c’est l’ordre linéaire (ensemble avec la prosodie particulière) qui signale la relation de consécution entre ‘P’ et ‘Q’[8] ; 2) Prosodie/ponctuation : à

[7] Par exemple, l’explication : *Il ne **tolérait** pas le désobéissance **; <, car**> c’**était** un monarque absolu* ; la causation : *Le déficit **s’aggrave : <parce que**> le Canada **importe** plus qu’il n’exporte* ; etc.

[8] Comme on le sait, la conséquence/la causation entretient un lien étroit avec la succession temporelle : la succession de deux événements dans le temps est souvent interprétée comme la présence d’un lien causal entre eux, même si, dans les faits, il

l'écrit, les propositions réalisant 'P' et 'Q' peuvent être séparés par une virgule, un point-virgule ou un point ; 3) Niveau de langue : la construction appartient au style narratif.

Par contre, une règle de paraphrasage doit spécifier, le cas échéant, les conditions sous lesquelles la substitution qu'elle décrit est possible. (Nous ne le faisons pas ici pour nos règles.)

CONDITION SUIFFISANTE ~ SUCCESSION TEMPORELLE

(5) anglais ~ français

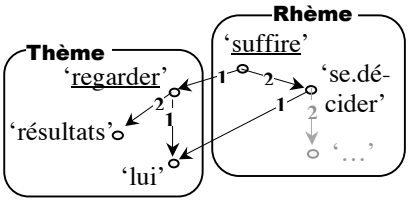
a. *He **made up his mind**$_{Q}$ immediately **after looking**$_{P}$ at the results.*
'Il **prit une décision** immédiatement **après avoir jeté un coup d'oeil** sur les résultats.'

b. *Il lui **suffit** d'un simple **coup**$_{P}$ **d'œil** sur les résultats **pour prendre** une **décision**$_{Q}$.*

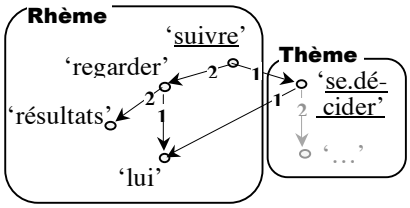
Les paraphrases en (5) affichent le lien entre la condition suffisante, que nous représentons par le sémantème 'P suffit pour que Q ait lieu' et la succession temporelle, représentée par 'Q suit P'[9]. La condition suffisante se réalise par la lexie SUFFIRE ou par l'expression *n'avoir qu'à V-er* ; la succession, quant à elle, s'exprime typiquement au moyen de la lexie APRÈS (ou ang. AFTER), comme en (5a), mais peut également s'exprimer au moyen d'une conjonction, comme en (5c) :

c. *Il **jeta**$_{P}$ un **coup d'œil** sur les résultats **puis** <**et ensuite**> il **se décida**$_{Q}$.*

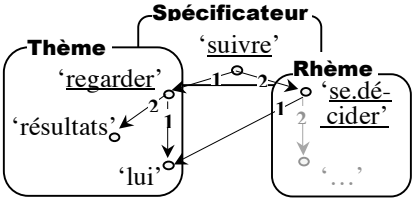
Voici les RSém (simplifiées) des phrases en (5) :



RSém 4, phrase (5b)



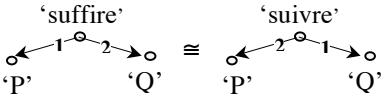
RSém 5, phrase (5a)



RSém 6, phrase (5c)

La règle de quasi-équivalence propositionnelle nécessaire pour relier ces représentations suit.

Règle $^{\text{SÉM(Propositionnelle)-6}}$



(1) 'P' de 'suffire' s'exprime comme « de V-inf » : ***Il suffit de franchir** la latitude du tropique **pour constater** immédiatement les améliorations sensibles côté confort. ≈ **Après avoir franchi** la latitude tropique, on **constate** immédiatement ... | **Il** me **suffisait de cliquer** sur « relever » **pour qu'**il **continue**. ≈ **Après que** je **cliquais** sur « relever », il **continuait**.*

(2) 'P' de 'suffire' s'exprime comme « de N » : ***Il** lui **suffit d'**un simple **examen** clinique **pour constater** l'état de déshydratation. ≈ **Après** un simple **examen** clinique, il **constata** ... | **Il** lui **suffit** de quelques heures de repos par nuit **pour reprendre** le souffle* [...] ≈ ***Après** seulement quelques heures de repos par nuit, **il reprend** le souffle* [...]

---

n'en est rien (cf. Nazarenko 2000: 45, où il est question de « propension [des humains] à l'interprétation causale »). Un lien consécutif implicite risque d'autant plus à être confondu avec la succession temporelle. Ceci est cependant un problème d'analyse et concerne peu une approche comme la nôtre, orientée vers la synthèse. Notons par ailleurs qu'une règle de paraphrasage reliant la conséquence/la causation et la succession temporelle (mais pas vice versa !) est parfaitement possible ; elle rend compte du fait que les phrases suivantes sont toutes des paraphrases acceptables de la phrase (4b) : *Le champagne apparut **et** (**ensuite**) la conversation s'anima. ~ **Après** l'apparition du champagne, la conversation s'anima.*

[9] Ce sémantème correspond à l'acception I.B2 du vocable SUIVRE du *Nouveu Petit Robert* : venir, se produire après dans le temps.

Cette règle doit s'appliquer en conjonction avec les règles de quasi-équivalence communicative ci-dessus : *Règle*$^{SÉM(Communicative)-3}$ (appliquée de droite à gauche) pour passer de la RSém-4 à la RSém-5 et la *Règle*$^{SÉM(Communicative)-5}$ pour assurer le passage vers la RSém-6. Seules des réalisations avec *après* (à partir d'appariements comme celui de la RSém-5) sont illustrées.

Notons que *simple* et *immédiatement*/*seulement* sont des ajouts stylistiques, qui renforcent, respectivement, l'idée de condition suffisante et de succession temporelle ; bien qu'ils soient très fréquemment utilisés (ce que démontre une recherche Google), ils peuvent être omis et donc ne font pas partie de la règle[10].

Des paraphrases avec DÈS QUE sont possibles dans le cas (1) ci-dessus : ***Dès que je cliquais*** *sur « relever » il* ***continuait*** ; dans le cas (2), on peut avoir des paraphrases mettant en jeu AVEC, MOYENNANT : ***Avec*** *<****Moyennant****>* ***un simple examen*** *clinique, il* ***constata*** *...*

2.2 Règles lexico-syntaxiques de paraphrasage

Toutes les nouvelles règles lexico-syntaxiques dégagées jusqu'à maintenant de notre corpus sont des règles relativement mineures, décrivant des « transformations » bien spécifiques. Ainsi, la règle ci-dessous met en jeu une restructuration syntaxique locale, déclenchée par des facteurs communicatifs ou rhétoriques, avec peu de changements lexicaux.

FAIRE N, EN V-ANT ~ FAIRE N, QUI V-E

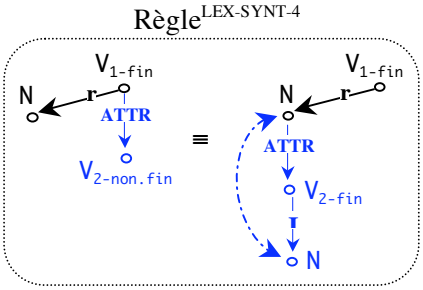
(6) anglais ~ français
a. *... it* ***maintained***$_{V1\text{-}fin}$ *an unrelenting embargo against Iraq,* ***causing***$_{V2\text{-}non.fin}$ *the deaths of civilians ...*
' ... **en causant la mort** de civiles ...'
b. *... ils ont maintenu*$_{V1\text{-}fin}$ *un* ***embargo*** *implacable contre l'Irak,* ***qui tue***$_{V2\text{-}.fin}$ *des innocents ...*

En (6a), le participe présent *causing* 'causant' dépend du verbe *maintain* (on le voit parce que le participe peut être paraphrasé par *thereby* [= by doing so] *causing* ; de plus, dans la traduction française, on a un gérondif, qui dépend nécessairement d'un verbe). Ce participe est rendu en (6b) par une relative qualifiant le nom *embargo*, l'objet direct du verbe *maintenir*. Donc, on a ici deux opérations (prises en charge par une seule règle) : 1) une translation « participe présent ~ proposition relative » et 2) un transfert du sous-arbre de la relative vers le nouveau gouverneur nominal.

La substitution illustrée en (6) n'est possible que si $V_{1\text{-}fin}$ est un verbe support : *L'armée* ***a lancé*** [IncepOper$_1$(ATTAQUE)] *contre la ville une* ***attaque*** *d'envergure,* ***en la réduisant*** *<****qui l'a réduite****> en ruine* vs. *L'armée a regardé les bras croisés l'attaque contre la ville,* ***en la sacrifiant*** *<*qui l'a sacrifiée> à l'ennemi*. En (6), *maintenir* = ContOper$_1$ (EMBARGO).

L'antéposition de la relative par rapport au verbe principal (le premier exemple ci-dessous) est possible mais contrainte par la structure communicative. La répétition du nom est souhaitable, voire obligatoire, pour des raisons stylistiques, si la relative en est séparée par un constituant relativement lourd : *L'armée yéménite a lancé, jeudi 10 septembre 2009 à l'aube, une* ***attaque*** *massive contre les rebelles chiites, dans le nord du Yémen,* ***attaque qui*** *à repoussé l'ennemi vers l'intérieur du pays.*

[10] Il y a ici l'interférence avec une autre règle de paraphrasage, sous-jacente aux paraphrases de type *L'examen* ***était très simple*** *~ J'****ai fait*** *l'examen* ***en un clin d'œil***, qui se base sur l'implication pragmatique « X est simple à faire ⇒ X se fait rapidement ».



r = **I** (V_1 est de type $Func_0$)
Une cyberattaque a frappé l'Internet mardi, ***en provoquant*** *la panne de plusieurs serveurs* <*Une cyberattaque,* ***qui a provoqué*** *la panne de plusieurs serveurs, a frappé l'Internet mardi*>.

r = **II** (V_1 est de type $Oper_1$)
Nous avons pris des mesures concrètes, ***en apportant*** <***qui apportent***> de *premières réponses aux attentes que vous avez formulées.* | *L'occident a fini par imposer des sanctions,* ***en freinant*** *gravement* <***qui ont*** *gravement* ***freiné***> *l'assistance humanitaire.*

3 Conclusion

Par la couverture des phénomènes linguistiques, nos règles de paraphrasage, en particulier nos règles sémantiques, vont au-delà de celles ordinairement prises en charge par les systèmes de TALN. Il est donc légitime de se demander à quel point leur automatisation et intégration à de tels systèmes est réaliste en ce moment. (Nous prenons pour acquis que c'est quelque chose de désirable, voire nécessaire, pour tout système de production automatique de textes qui se veut performant.) N'étant pas spécialiste de TALN, nous ne sommes pas en mesure de donner une réponse précise à cette question. Il s'agit certainement d'un défi de taille, puisque les règles de paraphrasage que nous envisageons présupposent le recours à des formalismes complexes — les représentations sémantiques, incluant l'information communicative, et les définitions lexicographiques poussées — formalismes qui ne sont pas complètement maîtrisés eux-mêmes. Cependant, on a déjà vu des implémentations robustes du système de paraphrasage lexico-syntaxique (certaines ont été mentionnées au début de l'article), lui aussi assez complexe. Il y a donc lieu de croire qu'on ne tardera pas à voir des implémentations des règles de paraphrasage sémantiques.

Remerciements

Un grand merci à Igor Mel'čuk pour ses commentaires judicieux. Mes remerciements vont également à mes assistantes de recherche, Alexandra Tsedryk et Stephanie Doyle.

Références

APRESJAN, J., BOGUSLAVSKIJ, I., IOMDIN. L., LAZURSKIJ, A., SANNIKOV, V., SIZOV, V. & TSINMAN. L. (2003). ETAP-3 Linguistics Processor : A Full-fledged NPL Implementation of the MTT. In: KAHANE, S. & NASR, A., eds, *Actes de la Première conférence internationale sur la théorie Sens-Texte*. Paris, École Normale Supérieure, juin 16-18 2003; 279-288.

BANNARD, C. & CALLISON-BURCH, C. (2005). Paraphrasing with Bilingual Parallel Corpora. In: *Proceedings of ACL 2005*. Ann Arbor, 25-30 June 2005; 597-604.

BARZILAY, R. & MCKEOWN, K. (2001). Extracting Paraphrases from a Parallel Corpus. In: *Proceedings of ACL/EACL 2001*. Toulouse, July 6-11 2001; 50-58.

DORR, B., DREEN, R., DEVIN, L., DAMBOW, O., DARWELL, D., HABASH, N., HELMREICH, S., HOVY, E., MILLER, K., MITAMURA, T., REEDER, F. & SIDDHARTHAN, A. (2004). Semantic Annotation and Lexico-Syntactic Paraphrase. In: *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora*, LREC, Portugal.

DANLOS, L. (2006). Discourse Verbs and Discourse Periphrastic Links. In: *Proceedings of Constraints in Discurse '06*, Maynooth, Ireland.

KAHANE, S. (2003). The Meaning-Text Theory. In: AGEL, V., EICHINGER L.M., EROMS, H-W., HELLWIG, P., HERINGER H.J. & LOBIN, H., eds. *Dependency and Valency. An International Handbook of Contemporary Research*, vol. 1. Berlin/New York: De Gruyter; 546-570.

KAHANE, S. & MEL'ČUK, I. (2006). Les sémantèmes de causation en français. In: HAMON, S. & AMY, M., eds. *La cause : approche pluridisciplinaire. Linx* 54 ; 247-292.

KITTREDGE, R. (2002). Paraphrasing for Condensation in Journal Abstracting. *Journal of Biomedical Informatics* 35: 4; 265-277.

IORDANSKAJA, L., KITTREDGE, R. & POLGUÈRE, A. (1991). Lexical Selection and Paraphrase in a Meaning-Text Generation Model. In: PARIS, C. L., SWARTOUT, W. R. & MANN, W. C., eds. *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Boston: Kluwer; 293-312.

LAVOIE, B., KITTREGDE, R., KORELSKY, T. & RAMBOW, O. (2000). A Framework for Machine Translation and Multilingual NLG Systems Based on Uniform Lexico-Structural Processing. In: *Proceedings of the 6th Conference on Applied Natural Language Processing*. Seattle.

MEL'ČUK, I. (1974). *Opyt teorii lingvističeskix modelej Smysl ⇔ Tekst.* Moskva: Nauka.

MEL'ČUK, I. (1992). Paraphrase et lexique: la théorie Sens-Texte et le Dictionnaire explicatif et combinatoire. In: MEL'ČUK, I. *et al.* (1984, 1988, 1992, 2000), *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I-IV*. Montréal: Presses de l'Université de Montréal; 9-59.

MEL'ČUK, I. (2001). *Communicative Organization in Natural Language.* Amsterdam/Philadelphia : Benjamins.

MILIĆEVIĆ, J. (2007a). *La paraphrase. Modélisation de la paraphrase langagière*. Bern: Peter Lang.

MILIĆEVIĆ, J. (2007b). Semantic Equivalence Rules in Meaning-Text Paraphrasing. In: WANNER, L., ed., *Selected Lexical and Grammatical Issues in Meaning-Text Theory. In Honour of Igor Mel'čuk.* Amsterdam/Philadelphia: Benjamins; 267-297.

NAZARENKO, A. (2000). *La cause et son expression en français*. Paris : Ophrys.

ZHAO, SH., WANG, H. & LIU, T. (2009). Extracting Paraphrase Patterns from Bilingual Parallel Corpora. *Natural Language Engineering* 15/4: 503-526.

ŽOLKOVSKIJ, A. & MEL'ČUK, I. (1967). O semantičeskom sinteze. *Problemy kibernetiki* 19: 177-238. [Traduction française : Sur la synthèse sémantique (1970). *TA Informations* 2: 1-85].