Réflexions sur la localisation, l'étiquetage, la reconnaissance et la traduction d'expressions linguistiques complexes.

Cédrick Fairon, Jean Senellart

LADL, Université Paris 7
2, place Jussieu, 75251 Paris Cedex 05
fairon@ladl.jussieu.fr – senella@ladl.jussieu.fr
http://www.ladl.jussieu.fr

Abstract

We describe a process translating automatically time adverbs from English to French. The mechanism is based on the use of finite state transducers. We discuss about the different ways of locating complex linguistic sentences. In order to translate this sentences, we show that the description must be accurate and that the tagging process is not involved. The quality of the translation obtained in this application is an indirect proof of the validity of our description method. Besides, we show the limitations of local approaches to translation.

1. Introduction

L'objectif du traitement automatique du langage naturel, initialement orienté d'une manière très ambitieuse vers la traduction purement mécanique de textes quelconques, semble se recentrer vers l'accomplissement d'une description complète du langage. Cette description pouvant être basée sur une description manuelle systématique du lexique ou sur des procédures plus ou moins automatiques d'étiquetage de textes. Cette dernière activité est généralement présentée comme une étape indispensable à toute analyse ultérieure et dont la justification est la présence d'ambiguïté dans le lexique. Ces ambiguïtés empêchent en effet une correspondance directe entre les mots d'un texte et leurs catégories grammaticales. Les procédures associées affichent des taux de « reconnaissance » impressionnants de l'ordre de 99 %. Cette approche est cependant difficile à évaluer puisqu'il semble que l'analyse humaine (l'interprétation) de phrases n'ait pas comme point de départ un tel étiquetage[i]. Nous donnons ici une application associée à une description complète d'un phénomène : la traduction automatique d'adverbes de date. L'évaluation de la qualité du processus, et donc de la description est simple[ii] puisque le résultat est soit une traduction correcte, soit une traduction incorrecte. Cette application permet d'envisager une comparaison entre une description systématique du lexique et celle d'une approche basée sur des catégories grammaticales,

[i] Au contraire, l'attribution humaine d'étiquette semble être un processus intellectuel de très haut niveau dont la base est la compréhension complète de la phrase. La comparaison entre plusieurs systèmes n'est pas non plus évidente puisque le choix des catégories grammaticales est dépendant de la théorie/de l'application et n'a guère de justification théorique intrinsèque au langage.

[ii] Dans le cadre restreint de la traduction d'adverbes de date, cette évaluation est réaliste dans la mesure où il est possible d'apprécier finement, indépendamment du contexte, de la qualité et de la correction d'une traduction.

actuellement à la base de nombreux systèmes de traduction automatique. Nous montrerons en particulier que pour compléter notre objectif, nous n'avons utilisé, à aucun moment, de référence générale à des catégories grammaticales, et donc à un éventuel étiquetage.

2. Qu'est-ce que reconnaître une expression dans un texte ?

Reconnaître une expression dans un texte peut-être considéré :

- comme le résultat d'un processus : des parenthèses délimitent dans le texte des séquences reconnues (nous parlons alors de localisation)

- ou encore comme le processus lui-même (nous parlerons alors de description).

Si le processus s'arrête à cette étape de reconnaissance, il n'y a pas de différence entre les deux approches puisque seul le parenthésage importe. Par contre, si nous voulons utiliser le résultat de la reconnaissance dans un second traitement, alors la localisation ne sera pas suffisante, et il faudra reconsidérer l'ensemble de la procédure de description, ce que nous montrons ici sur l'exemple des groupes nominaux de la forme **déterminant-nom-DE-nom** (comme *le milieu de l'été*, *le garde du corps* ou *la stabilité des* prix) et en considérant comme second traitement une des applications possibles découlant de cette reconnaissance : la traduction automatique.

Supposons que nous soyons capables d'étiqueter de manière non ambiguë les groupes nominaux N_1 *de* N_2 (le rattachement du groupe prépositionnel *de* N_2 au nom étant un problème classique de l'étiquetage automatique). Pour que cet étiquetage ait un sens, il faut que l'attribution d'étiquettes à des séquences linguistiques permette de prolonger le calcul d'analyse de la phrase, en prenant ces étiquettes comme unité de base du calcul.

Un des calculs possibles sur ces catégories est la traduction. Supposons que nous disposions d'une procédure de traduction notée T. Dans l'hypothèse d'utilité de l'étiquetage, cette procédure devrait s'exprimer ainsi :

$$T(N_1 \textit{ de } N_2)=T_{\text{N de N}}(T(N_1),T(N_2))$$

C'est à dire que la procédure T ne dépend pas de la nature des éléments lexicaux N_1 et N_2, mais uniquement du fait qu'ils sont tous les deux « étiquetés » *N*. Intuitivement (ou naïvement) pour traduire le schéma *N de N* en anglais nous pouvons donc, dans cette hypothése, supposer que $T_{\text{N de N}}(T(N_1),T(N_2)) = T(N_1)$ of $T(N_2)$.

Considérons les quelques exemples du tableau suivant. Nous y avons fait figurer des séquences *N de N* extraites de corpus français, avec leurs «traductions » suivant le schéma précédent, et leurs «vraies » traductions. Nous avons indiqué à l'aide de (*) , (?*) et (?) plusieurs degrés d'acceptabilité des traductions (Nous supposons que nous pouvons proposer des traductions indépendantes du contexte, ce qui est évidemment incorrect, mais permet de simplifier l'argumentation sans la modifier).

	$T_{\text{N de N}}$	« vraie traduction »
Milieu de l'été	middle of the summer	at midsummer
Membre du gouvernement	member of the government	
Méchanceté de l'équipe	nastiness of the team	

Secret de polichinelle	???	open secret
Garde du corps	(*) guard of the body	bodyguard
lieu du meurtre	(*?) place of the murder	scene of the murder
		the place where X was murdered
loi de juin	(?) law of June	June's law
Projet d'un an	(?) project of one year	one year project
critique des professionnels	(*) criticism of the professionals	criticism FROM professionals
stabilité des prix	(?) stability of price	price stability
stabilité des impôts	(*?) stability of the taxation	taxation stability
		stability IN taxation

Tableau 1

Traduction des structures N_1 de N_2 suivant le schéma $T(N_1)$ of $T(N_2)$. Les déterminants ne sont pas indiqués car trop dépendants du contexte. La troisième colonne indique une traduction préférentielle (qui dépend cependant du contexte, ce qui n'est pas vraiment le problème ici)

- Les trois premières entrées de la table correspondent à trois cas de figure où la règle de traduction $T_{N\ de\ N}$ fonctionne correctement.
- Les trois entrées suivantes correspondent à trois mots composés en français. Traduire chacune de ces entrées nécessite l'utilisation d'un dictionnaire de mots composés puisque par définition le calcul de sens (et donc de la traduction) de ces entrées ne suit pas de règle générale.
- Viennent ensuite deux entrées comportant une référence à une date et à une durée : dans les deux cas le schéma de traduction est particulier (propre à l'entrée) *June's law* et *one year project.*
- Enfin, les trois derniers exemples montrent la nécessité d'utiliser une préposition différente de *of* sans aucune justification générale.

Nous pouvons déduire de ces exemples que la règle $T_{N\ de\ N}$ n'a rien de général. Considérer que les exemples où elle ne s'applique pas sont des «exceptions » qu'il faut décrire autrement pourrait être une solution, ce qui est d'ailleurs le cas actuellement dans plusieurs systèmes de traduction automatique qui utilisent des listes de mots composés. Pour ces mots cependant, l'étiquetage préalable n'aura servi à rien puisque la solution adoptée consiste associer une séquence sans étiquette du français à une séquence en anglais. D'autre part, la notion d'exception s'effondre quand on procède à une énumération systématique : le nombre de cas où la règle ne s'applique pas est comparable à ceux où la règle s'applique. D'une manière indirecte, la preuve de cette assertion est apportée par le fait qu'un mot sur deux d'un texte écrit fait partie d'une expression composée (Senellart 1998) non nécessairement nominale. De plus, dans les derniers exemples du tableau, il est difficile de parler de mots composés : *la critique de* N_2 se traduira d'une manière très générale par *critism from* N_2 . Dans ce cas encore, l'étiquetage préalable n'apporte pas la factorisation souhaitée.

Avec en tête l'objectif de la traduction, on ne peut dire qu'une expression n'est décrite que si le processus de description est unique pour chaque expression de comportement linguistique différent. Dans le cas précédent, la description projette toutes les structures *N de N* sur une structure unique comme si tous les groupes nominaux de cette forme là avaient les mêmes propriétés syntaxiques. En particulier, on remarque que les expressions du tableau entrent dans des constructions élémentaires différentes : *les professionnels font la critique de* N_2, *les*

prix ont une certaine stabilité, etc. Cette information est présente dans des bases de données construites au LADL et appelées lexique-grammaire. Une reconnaissance basée sur des dictionnaires de mots composés et le lexique-grammaire permettra donc d'avoir effectivement une règle propre pour chaque unité reconnue et donc de lui associer une traduction.

Nous venons de montrer que décrire les groupes (N de N) ne peut se faire par une procédure d'étiquetage général, ce qui remet en cause l'utilité de l'étiquetage de ces groupes. Dans la description des adverbes de date que nous utiliserons, chaque chemin propre devra donc correspondre à une seule et unique construction linguistique. Pour prendre un exemple, considérons les deux expressions *pendant la (nuit+période) du 24 au 25 janvier* . Ces deux expressions ne correspondent pas à la même construction, la preuve en est leurs traductions respectives :

during the (night+*term) of Tuesday the 24th

during the (*night+term) from 24th to 25th January

La ressource décrivant ces deux expressions sera donc différente malgré l'apparente équivalence.

2. Traduction des adverbes de date (anglais → français)

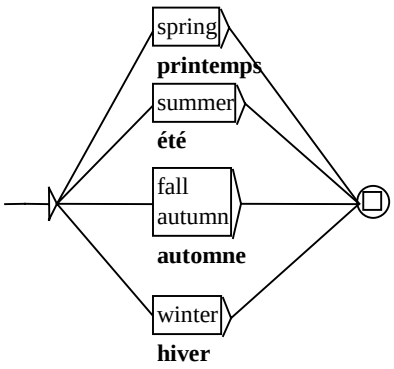
Le processus de traduction que nous décrivons maintenant est une application simple et complète de mécanismes formels utilisables par exemple dans le système INTEX (cf. Silberztein 1993), mais elle est aussi une preuve de la validité de la description.

2.1 Aspects techniques

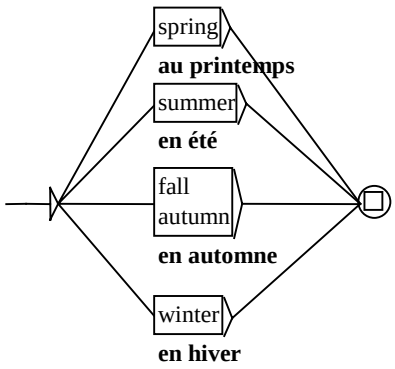
La description sur laquelle nous nous sommes basés est une description faite par M. Gross et D. Maurel (pour le français) développée sous la forme de graphes (des automates finis récursifs). Cette description a été réalisée pour le français et l'anglais avec un objectif commun : la recherche de motifs linguistiques dans des corpus informatisés. Pour ce faire, les graphes doivent viser l'exhaustivité. Autrement dit, toutes les expressions de date valides doivent être représentées dans les graphes pour qu'il soit ensuite possible de les repérer automatiquement dans les textes.

Quand un des chemins du graphe correspond à une séquence de mots dans un texte, on dit que cette séquence est reconnue. Notre travail a été d'associer pour chacune des séquences reconnues une sortie dans la langue cible. Nous avons donc utilisé ces graphes en tant que transducteurs (pour une description détaillée de l'utilisation des transducteurs, cf. Silberztein, 1998). Dans la mesure où ces transducteurs permettent également le déplacement de séquences reconnues, nous parlerons de « transducteurs étendus » (Senellart 1998).

L'utilisation de cette bibliothèque de graphe pour la traduction a nécessité certains aménagements. Par exemple, dans le graphe *ByDate.grf* (qui décrit les expressions du genre *by the middle of winter, not later than summer*) il a été nécessaire de créer en parallèle au graphe *Season.grf* (suffisant pour traduire le premier exemple) un nouveau graphe *PrepSeason.grf* permettant d'ajouter la préposition adaptée (*au printemps* vs. *en hiver, en été, en automne)* et nécessaire pour traduire le deuxième exemple : *not later than summer → pas plus tard qu'en été.* Autrement dit, la catégorie **saison** n'est pas suffisante pour la traduction : il manque de l'information syntaxique.



Graphe x. *Season.grf*



Graphe x. *PrepSeason.grf*

Les transducteurs sont appliqués en mode de remplacement : à chaque fois qu'un des chemins du graphe est identifié dans un texte, la sortie associée à ce chemin est insérée dans le texte à la place du texte de la langue source. Pour faciliter la lecture des résultats, nous recopions également entre crochets la séquence originale. Ce recopiage est effectué en recourant au principe des transducteurs à mémoire tels qu'ils sont décrits par (Silberztein 1998). Cette technique permet de recopier en sortie la séquence qui a été reconnue en entrée dans le graphe.

Le principe général envisage l'association d'une traduction pour chacun des chemins du graphe principal. Un tel principe est cependant très coûteux et relativement artificiel. Il s'agirait dans ce cas de prendre un à un les chemins du graphe principal et de leur associer à chacun une traduction. Nous avons associé à chaque chemin de chaque sous-graphe des traductions locales. Cette position s'accorde avec le principe de factorisation des expressions dans les graphes qui a pour objectif de simplifier le travail de constitution de la bibliothèque.

Nos transducteurs sont appliqués de manière itérative. Au stade actuel de note description, nous avons besoin de trois passes pour passer de la langue source à la langue cible.

La traduction est assurée par la première passe. Dans le même temps, des marqueurs graphiques sont insérés pour identifier les mots ou syntagmes qui doivent être permutés. Au terme de cette première étape, *the following day* a été transcrit en *le [suivant] [jour]*. Cette particularité est évidemment propre à la paire de langue traitée. D'autres adaptations seraient sans doute à envisager dans le cas de traductions envisageant d'autres paires de langues.

La deuxième passe a pour fonction de permuter les séquences qui ont été marquées par [ ] lors de la première étape.

Enfin une troisième passe relativement simple prend en charge la mise au point de quelques problèmes d'élision *(le mois DE avril),* de flexion *(LE TOUT PREMIER semaines),* et de contraction *(DE LE dernier jour).* La question des contractions peut être peut être résolue par l'application d'un transducteur simple, mais l'élision et la flexion doivent être traitées par des procédures extérieures à INTEX.

2.2 Les limites de l’approche locale et de la traduction automatique

Le travail de traduction effectué localement sur les adverbes se base sur le principe implicite qu’il est possible de traduire ces expressions indépendamment du contexte. Il est évident qu’une telle supposition est fausse dans le cadre général. Le cas de figure rencontré le plus couramment est du type : *during Sunday’s game.* Dans cet exemple, l’adverbe de date reconnu est *during Sunday*. L’analyse correcte devrait être *during (Sunday’s game)*. Cette

catégorie d'adverbe de date n'est pas descriptible puisqu'on aurait aussi pu trouver *during the game won by the Heat*, c'est à dire un adverbe de date libre. Le fait d'opérer localement (c'est à dire, sans avoir reconnu la structure syntaxique de la phrase) n'est cependant pas théoriquement incorrecte si on travaille sur une structure d'automate de texte. Toute expression reconnue donne lieu à une nouvelle dérivation (chemin parallèle) dans l'automate de la phrase. Cette dérivation ne sera effectivement prise en compte que si on peut trouver un chemin complètement reconnu entre le début et la fin de la phrase passant par celle-ci. L'automate de texte suivant montre cette situation pour une phrase d'exemple. Dans ce cas, la seule analyse allant du début à la fin de la phrase ne passe pas par l'adverbe de date reconnu. Seule la volonté de revenir à la fin de l'analyse à une forme linéaire entraîne l'apparition d'analyses erronées, il en est de même du phénomène d'ambiguïté (Senellart 1999).

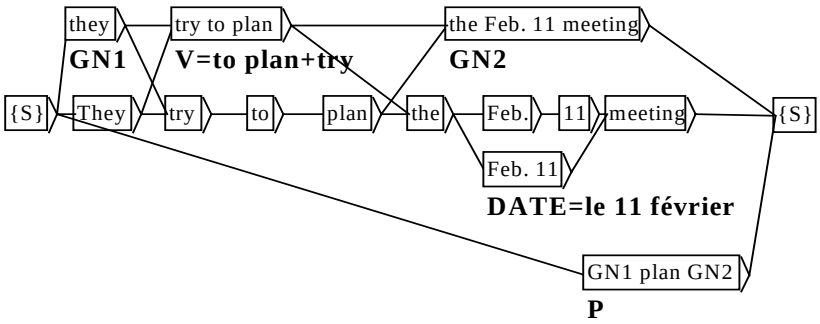


Figure x. *Automate du texte*

De plus certaines expressions sont ambiguës. Cette ambiguïté peut être réelle (un traducteur humain aurait lui-même à faire un choix arbitraire) ou liée au fait que l'analyse est locale et ne prend pas en compte les considérations sémantiques. Dans les exemples suivant, la préposition *over* a deux traductions différentes :

May I stay over friday → Puis-je rester jusqu'à vendredi

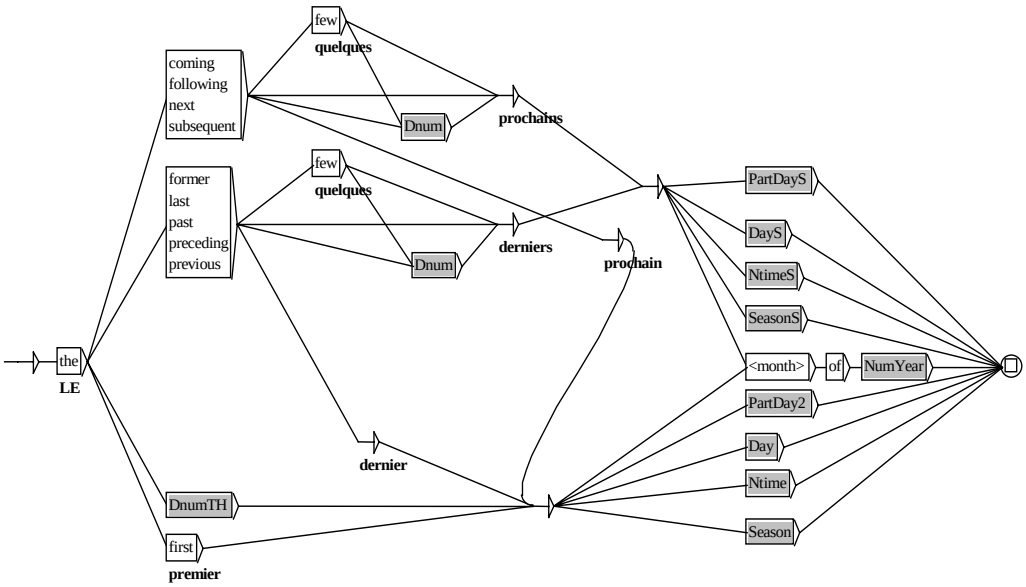
It took place over friday → Cela a eu lieu tout le vendredi

Le choix automatique de l'une ou l'autre traduction devrait tenir compte du temps du verbe principal et de sa nature. Dans de tels cas d'ambiguïté, nous proposerons les deux traductions en sortie, car il n'est pas possible de trouver une traduction «passe-partout» au sens où l'entend M. Salkoff (Salkoff, à paraître). Dans notre système, *May I stay over friday* sera donc traduit par *May I stay (jusqu'à+tout le) vendredi*.

2.3 Application

Les exemples que nous présentons ci-dessous sont tirés d'un corpus de 12 Mo composés de textes téléchargés sur le site du New-York Times à la date du 4/11/1998. Ce corpus correspondant environ à 2 millions de mots.

Le graphe *TheFollowingNtime.grf* est appelé par un graphe parent qui distingue les prépositions susceptibles d'apparaître devant cette séquence (*over, for, in*, etc.). C'est le graphe parent que nous avons appliqué sur notre corpus pour obtenir les exemples présentés ci-dessous.



Graphe x. *TheFollowingNtime.grf*

It has only gotten worse **pendant LE six derniers mois** < in the last six months >.

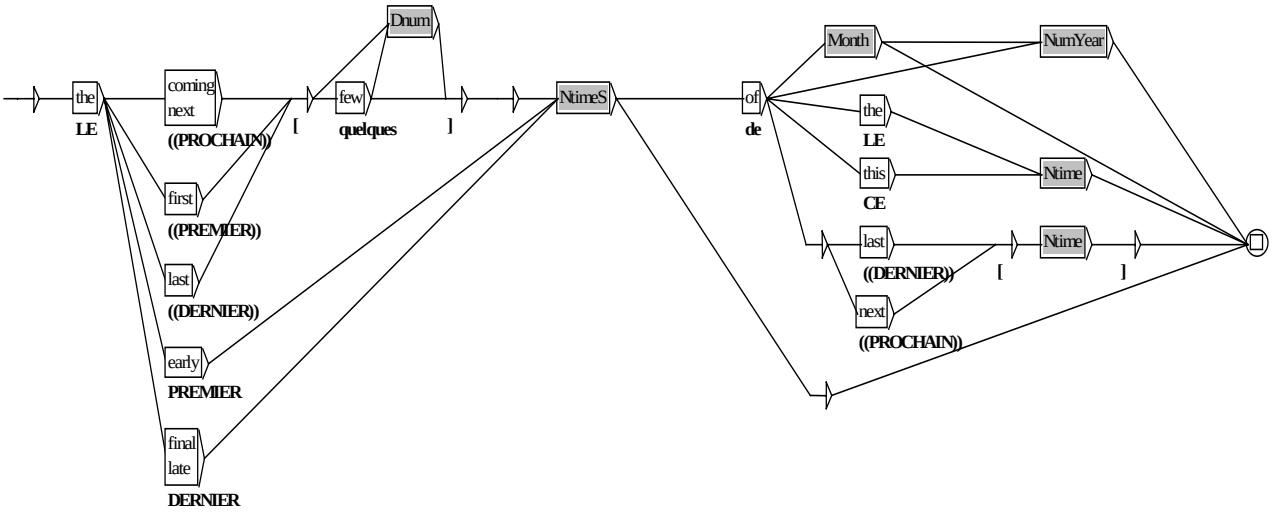
{S} LE troisième nuit <on the third night> the commandos hunkered down near the village.

Analysts predict that fares will keep falling at least **pendant LE quelques prochains mois** < for the next few months >.

I get charged **pendant LE premier minute** < for the first minute > but after I press the Aplio button.

Austria equalized **à LE dernier minute** <in the last minute > against everyone's favourite team Cameroon.

It reflects the stunning acceleration of all these political trends **au cours de LE dernier quart de siècle** < over the past quarter century >.



Graphe x. *TheNextTimesOf.grf*

Like **dans LE PREMIER années** < in the early years >, he would do wild things, like challenging R. Reagan to a duel.

"**LE [DERNIER] [quelques] jours** < The last few days >," says taxi driver Boonlarb Srikam, "I've noticed all that".

The imperial family was killed **dans LE [PREMIER] heures de le [juillet] [17] 1918** < in the early hours of July 17, 1918 >.

{s} **LE [PROCHAIN] [dix] années** < for the next ten years >, the main tasks of the euro will remain internal ones.

Beijing's leaders realize that **LE [PROCHAIN] [quelques] années** < the next few years > are going to be tough-and that's good news.

Ces quelques exemples reflètent l'état de la traduction avant le passage du graphe de permutation. Au passage de ce dernier, les groupes marqués par des doubles parenthèses et ceux marqués par des crochets sont permutés. *LE [PROCHAIN] [quelques] années* devient alors *LE quelques PROCHAIN années*. La troisième passe prend ensuite en charge les élisions et flexions, comme nous l'avons expliqué ci-dessus.

Conclusion

La procédure décrite ici conduit à une traduction de qualité des adverbes de temps. Comme nous l'avons vu, cette traduction connaît certaines limites qu'il n'est pas possible de franchir. Ces limites n'étant pas liées à la description.

Aucune des procédures utilisées (mis à part la procédure de flexion) ne recourt à des catégories traditionnelles. Les plus « grandes » catégories d'objets syntaxiquement équivalents ont au maximum une dizaine d'items (les jours de la semaine, les mois). Nous avons effectué le même travail sur d'autres classes d'expressions linguistiques comme les titres de personnalité et les groupes nominaux décrivant des noms propres. Le problème général de traduction ne résume pas à de telles classes. Cependant, dans un contexte d'extraction d'information dans un document en langue étrangère, ces procédures permettent d'obtenir certaines informations complexes (cf. Senellart 1998). De plus, une telle méthodologie est applicable pour toutes les expressions composées figées (dont les dates font partie).

La qualité de la traduction dépend de celle de la description comme nous l'avons montré. Le travail d'adaptation de la description pour un but de traduction prend en compte les particularités de la langue cible. Nous n'aurions pas eu besoin de différencier *printemps* et *hiver* dans nos automates si la préposition associée était la même comme c'est le cas en anglais (*in summer*…). L'objectif de séparation des objets syntaxiquement distincts est donc lié non seulement à la langue source mais aussi à la langue cible. Une description idéale (s'il en est une) devrait passer par une représentation formelle universelle. Les tentatives de formalisation effectués dans un domaine aussi numérique que les dates (Bestougeff 1988) montrent qu'il n'est pas évident que cette représentation existe (à l'aube = ?, à midi ≠ à 12h00) : la structure de stockage idéal restant donc le langage naturel avec les inconvénients cités. Cela implique que le travail de description/traduction devra être adapté pour toute autre paire de langue.

Enfin, dans la même optique, il n'est vraisemblablement pas envisageable d'inverser simplement nos transducteurs pour effectuer l'opération de traduction inverse. L'ensemble des séquences générées par le « traducteur » étant un sous-ensemble non structuré des séquences possibles de la langue cible.

Références

Bestougeff H. Ligozat. (1988), *Outils logiques pour le traitement du temps, de la linguistique à l'intelligence artificielle* (Masson).

Gross M. (1986), *Grammaire transformationnelle du Français – Syntaxe de l'adverbe*, Vol. 3 (Cantilène).

Gross M. (1997), *La traduction automatique a 50 ans*, Pour la Science, Numéro Spécial.

Maurel D. (1989), *Reconnaissance des séquences de mots par automates. Adverbes de date du français*. Thèse de docyorat, Université Paris 7.

Salkoff M. (à paraître), *A Formal Comparative Grammar of French and English on Translation Principles.*

Senellart J. (1998), *Locating Noun Phrases with Finite State Transducers*. Coling-ACL'98, pp.1212-1220.

Senellart J. (1999), *Localisation d'expressions inguistiques complexes dans de gros corpus*. Thèse de doctorat, Université Paris 7.

Silberztein M. (1993*), Dictionnaires électroniques et analyse automatique de textes : le système INTEX* (Masson).

Silberztein M. (1999), Transducteurs pour le traitement automatique de textes, Travaux de Linguistique, N°37, pp.127-142.