MATHÉMATIQUES ET SCIENCES HUMAINES

DENIS MAUREL

Description par automate des dates et des adverbes apparentés

Mathématiques et sciences humaines, tome 109 (1990), p. 5-16.

<http://www.numdam.org/item?id=MSH_1990__109__5_0>

Math. Inf. Sci. hum. (28e année, n°109, *1990*, p.5-16).

DESCRIPTION PAR AUTOMATE DES DATES ET DES ADVERBES APPARENTÉS

Denis MAUREL[1]

RÉSUMÉ - *Cet article présente une utilisation de la théorie des automates finis pour la reconnaissance automatique des dates (et des adverbes apparentés à une date) dans un texte. Une introduction porte sur la nécessité d'une étude syntaxique précise du lexique et sur l'utilisation des automates en informatique linguistique. Puis, l'article propose la construction d'automates pour les dates, les heures et certains compléments circonstanciels de temps, avant de discuter sur les tables de transitions et la forme des bases de données correspondantes. Enfin, une conclusion précise quelques applications de cette étude.*

ABSTRACT - Describing by Automata Dates and Adverbs Related to a Date

This article presents a use of finite automata theory for automatic recognizing of French dates (and adverbs related to a date) in a text. The introduction deals with the necessity of a precise syntactic study of vocabulary and with using automata in linguistic information processing. Then the article describes automata construction for dates, hours and some time adverbial complements before arguing about tables of transitions and about forms of corresponding data bases. Lastly, a conclusion precises practical applications of this study.

1. INFORMATIQUE ET LINGUISTIQUE

Le développement actuel de l'informatique dans tous les domaines et son ouverture à un public de plus en plus large impose une amélioration sensible de la communication homme-machine. Or le langage le plus adapté à la communication, à la fois par sa richesse et par le nombre de ses utilisateurs, est bien sûr la langue naturelle. C'est pourquoi les perspectives commerciales des industries de la langue sont souvent citées comme le grand marché de cette fin de siècle [1]. La plupart des systèmes commercialisés de communication homme-machine (par exemple les systèmes d'interrogation de bases de données ou les correcteurs orthographiques) reposent sur l'utilisation de mots-clefs appartenant à un vocabulaire figé, sans considération de leur environnement syntaxique. Il s'ensuit par conséquent une perte d'informations et une arrivée de réponses ne correspondant pas au souhait de l'utilisateur. La nécessité d'une étude syntaxique précise de chaque entrée du lexique apparaît donc comme une condition impérative de la réussite d'une telle amélioration (voir par exemple J.L. Jayez [2], L. Danlos [3] et la description systématique des constructions syntaxiques du français par M. Gross [4][5][6]).

---

[1] Travail réalisé dans le cadre du Laboratoire d'Automatique Documentaire et Linguistique (Université Paris 7) et du Centre d'études et de recherches en Informatique linguistique.

Il s'agit donc d'utiliser, dans une analyse automatique de texte, des informations sur la structure syntaxique d'une phrase ou d'une partie de phrase. Pour cela, il faut disposer d'une base de données importante en taille et assez souple pour permettre facilement toute sorte de corrections, ajouts ou retraits. De plus, la présentation des données doit être facilement accessible de la part d'un non informaticien. Toutes ces qualités se retrouvent [7] dans la théorie des automates finis (un automate [8] est un graphe comportant des sommets, appelés états, et des flèches, appelées transitions). L'idée d'utiliser des automates dans le domaine linguistique a été avancée par C.F. Hockett [9] ainsi que par N. Chomsky et G.A. Miller [10] qui en ont aussi souligné les insuffisances dès qu'il est nécessaire d'introduire l'idée de récursivité dans la grammaire, c'est-à-dire finalement dès que le domaine considéré prend de l'importance. Comme notre étude restera limitée, nous ne retiendrons des automates que leurs qualités. Un exemple - parmi d'autres - d'utilisation des automates en linguistique informatique appliquée au Français est donné par par J. Courtin [11][12].

Cet article présente la construction d'un automate fini dans le but d'obtenir une première approche de la structure temporelle d'un texte. Celle-ci est véhiculée principalement par le temps des verbes et par des adverbes ou des compléments de temps. Laissant de côté l'analyse des verbes, l'étude qui suit abordera en première partie la construction d'automates pour les dates proprement dites, les horaires et les adverbes de temps. Puis seront abordés les compléments circonstanciels de temps. Dans une deuxième partie, il sera traité des tables de transitions qui complètent au niveau lexical la description syntaxique par automate. Enfin, la conclusion proposera une réflexion sur l'insertion d'automates (tels que l'automate réalisé) dans un cadre plus vaste.

L'automate présenté ici a servi à la réalisation d'un programme de reconnaissance automatique des adverbes de date du Français [13]. En guise d'illustration, voici quelques exemples, extraits du journal *Le Monde*, où les séquences de mots qui sont écrites en italique sont celles qui ont été reconnues par le programme mis au point :

Il a fait défection *le 7 novembre 1981, un mois et une semaine* avant la proclamation...

A deux mois environ du sixième anniversaire de la guerre irano-irakienne, que lui-même a déclenchée *le 22 septembre 1980*, le président Saddam Hussein a appelé *le 2 août* les dirigeants iraniens...

Les crédits à la consommation ont continué d'augmenter rapidement... *au cours des trois premiers mois de l'année.*

La Cour s'était retirée pour délibérer *jeudi en milieu de matinée*... [Elle] avait choisi *la première heure du trentième jour* du procès. *Dimanche à 1 heure du matin*... M. Hachemi Zammel prononce le verdict...

2. CONSTRUCTION DE L'AUTOMATE

2.1. *Dates*

Les trois automates des figures 1 à 3 contiennent la mention du jour, le premier avec l'article défini *le* ou l'adjectif démonstratif *ce*, le deuxième avec l'article indéfini *un* et le troisième sans article, ni adjectif démonstratif [14].

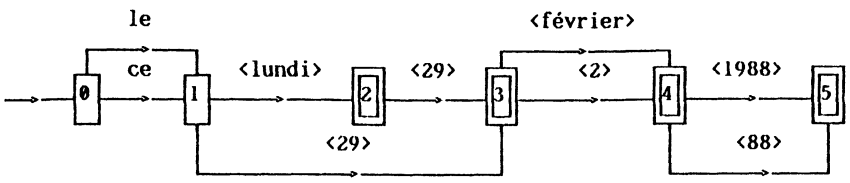

Figure 1

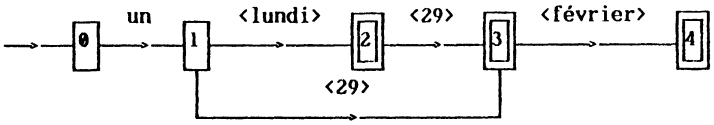

Figure 2

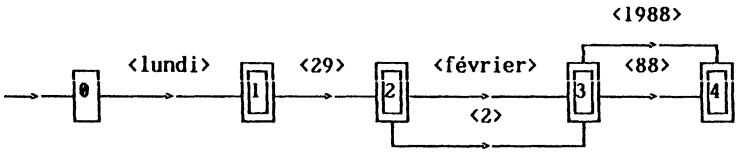

Figure 3

Bien sûr, les exemples utilisés sur ces figures pour étiqueter les différentes transitions représentent en fait des classes sémantiques (ou, plus loin, syntaxiques) : l'étiquette *<lundi>* est équivalente aux étiquettes de la classe {*dimanche, lundi, mardi, mercredi, jeudi, vendredi, samedi*}. De même, l'étiquette *<29>* remplace les étiquettes de la classe {1^{er}, *1, 2..., 31*}, l'étiquette *<février>*, celles de la classe {*janvier, février, mars, avril, mai, juin, juillet, août, septembre, octobre, novembre, décembre,* 7^{bre}, 8^{bre}, 9^{bre}, 10^{bre}}, l'étiquette *<2>*, celles de la classe {*1, 2..., 12*}, et les étiquettes *<1988>* et *<88>*, respectivement celles des classes {*100, 101..., 1988, 1989...*} et {*1, 2..., 98, 99*}. Bien sûr, il faut tenir compte du fait que tous ces nombres peuvent aussi être écrits en toutes lettres ou en chiffres romains.

L'automate suivant (figure 4) ne concerne plus que les mois et les années avec les mêmes conventions d'étiquetage.

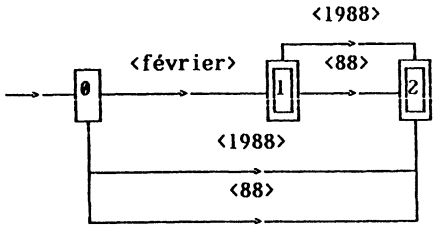

Figure 4

Une étude linguistique plus poussée des formes prises par une date montrerait qu'il est souvent possible d'insérer, dans ces quatre automates, soit un *modifieur* (*<Modif>*), soit une apposition. La classe des *modifieurs* comprend des adjectifs (par exemple *prochain*), des locutions adverbiales (par exemple *à venir*) et des participes passés (par exemple *écoulé*). La séquence suivante, où le signe + désigne la disjonction exclusive, présente une date suivie d'un tel modifieur :

le lundi 29 février (prochain + à venir)

La classe des appositions comprend pour sa part uniquement des noms de temps (*<Ntps>*), comme par exemple *matin* dans la séquence :

le lundi matin 29 février

L'ensemble de ces considérations permettent la construction de l'automate de la figure 5 à partir des quatre automates précédents.

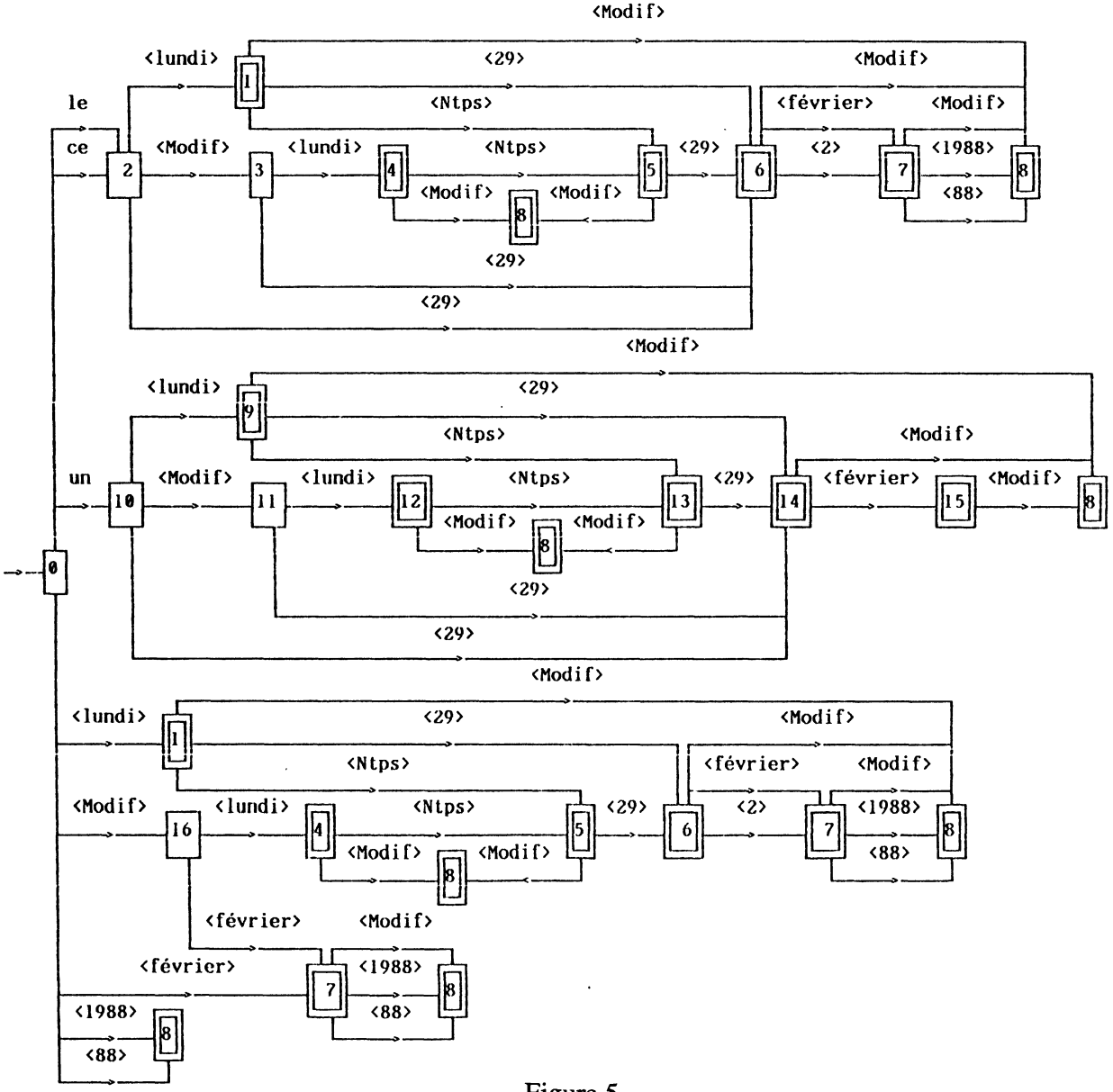

Figure 5

2.2. *Heures*

La description de l'heure est donnée par un sixième automate (figure 6). L'étiquette <*8*> est équivalente aux étiquettes de la classe {*0, 1, 2, 3...,24*} et les étiquettes <*5*> et <*7*>, à la classe {*0, 1, 2, 3...,59*}. De même l'étiquette <*midi*> correspond à la classe {*midi, minuit*} et les étiquettes <*heures*>, <*minutes*> et <*secondes*> sont respectivement équivalentes aux étiquettes des classes {*heure, heures, h*}, {*minute, minutes, mn*} et {*seconde, secondes, s*}.

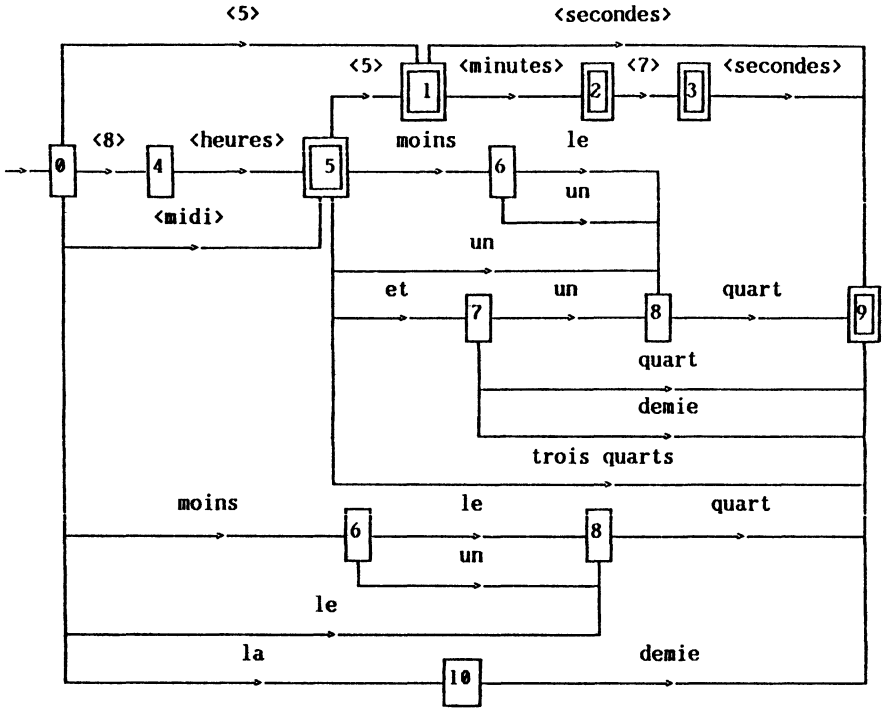

Figure 6

Ces six premiers automates présentent un grand nombre de possibilités combinatoires, qui illustrent bien la diversité des formes qu'il faut reconnaître lorsqu'on cherche à couvrir l'ensemble du langage naturel. Par exemple, il est remarquable que la langue française dispose de plusieurs expressions synonymes (toutes attestées par la littérature) pour désigner le quart d'une heure :

A sept heures moins (le + un) quart

Il était midi (un + et un + et) quart

2.3. *Adverbes de temps apparentés à une date*

L'automate de la figure 7 reconnaît les adverbes de temps en lien avec une date (<*Advtps*>) qu'il est possible de classer schématiquement en deux groupes : les adverbes qui impliquent l'idée d'une répétition (par exemple *quotidiennement*) et ceux qui désignent une date (par exemple *aujourd'hui*). Les premiers apparaissent toujours seuls, alors que les seconds peuvent être précédés d'une préposition (<*Prép*>) et suivis d'une apposition par un nom de temps, comme par exemple dans la séquence :

dès demain matin

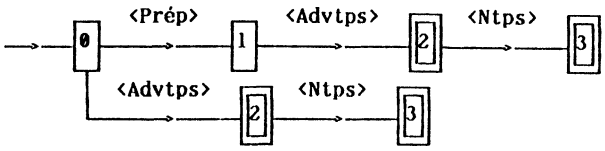

Figure 7

2.4. *Compléments circonstanciels de temps*
La simple lecture d'un article de journal prouve que la reconnaissance des dates et des adverbes de temps précédents est encore totalement insuffisante à la compréhension de la structure temporelle d'un texte. En effet d'autres expressions, les *compléments circonstanciels de temps*, font référence à une date, parfois avec une aussi grande précision, parfois de manière plus floue. La date de coréférence peut d'ailleurs être une date du texte, mais aussi la date de parution du journal ou du rapport, ou la date d'élocution dans le cadre du langage parlé.

Bien sûr, la construction de ces compléments circonstanciels de temps ne peut être réalisée en dehors de l'analyse de phrases complètes, car les formes autorisées varient suivant le contexte. Plus précisément, c'est le verbe qui en détermine le sens et la forme. Une étude systématique de tous ces compléments a donc été réalisée à partir d'un verbe choisi comme classifieur des dates et des adverbes apparentés : le verbe *avoir lieu.* L'utilisation de ce verbe tient principalement à ce qu'il admet différents types de compléments (dates proprement dites, répétitions ou durées) :

La réunion a lieu lundi
La réunion a lieu tous les lundis
La réunion a lieu toute la journée de lundi

Les compléments circonstanciels sont des groupes nominaux en général prépositionnels. L'automate de la figure 8 reconnaît les plus simples d'entre eux qui sont constitués de la combinaison d'une préposition, d'un article ou d'un adjectif démonstratif, d'un adjectif numéral cardinal (*<Dnum>*) et d'un nom de temps. Il s'agit par exemple de séquences comme :

dans les trois jours
à partir de ce week-end
durant la semaine
par mois
pendant des années
deux décennies

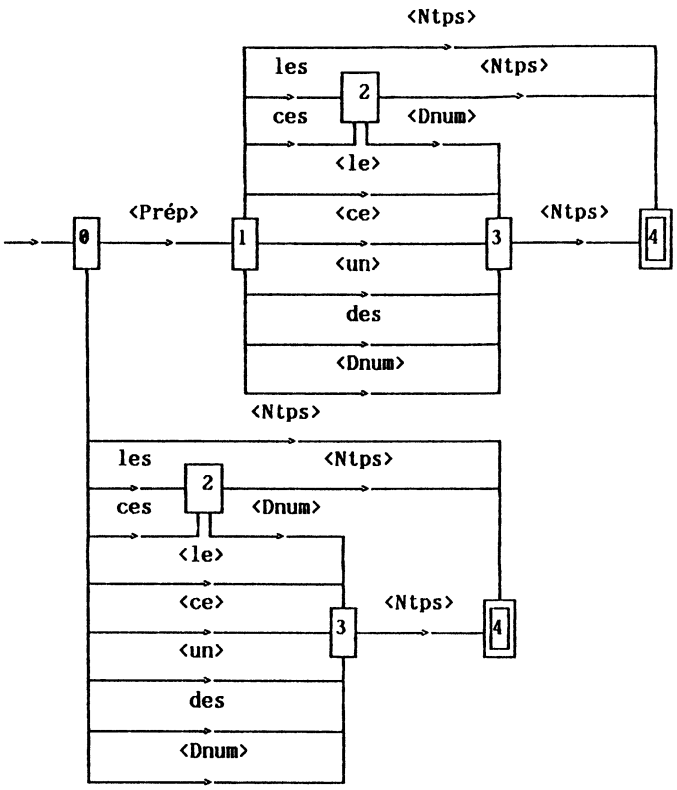

Figure 8

Bien sûr, dans cet automate, les étiquettes *<le>*, *<ce>* et *<un>* désignent respectivement les classes {*le, la, l'*}, {*ce, cet, cette*} et {*un, une*}. Comme pour les dates, il est possible d'insérer dans ces groupes nominaux des modifieurs ou des appositions, comme par exemple :

dans les trois derniers jours
à partir du lendemain matin

L'automate de la figure 9 traduit ces différentes possibilités.

3. AUTOMATE FINAL

Pour être plus complet, il faudrait ajouter que les dates, les heures, les adverbes de temps et les compléments circonstanciels de temps comportent souvent un *déterminant* (*<Dét>*). La classe des déterminants contient des adverbes (par exemple *exactement*), des adjectifs (par exemple *certain*) et aussi des groupes nominaux (par exemple *la fin de*) [6]. Par exemple :

après exactement trois jours
durant certaines semaines
vers la fin de l'année
vers le début du lundi 29 février 1988

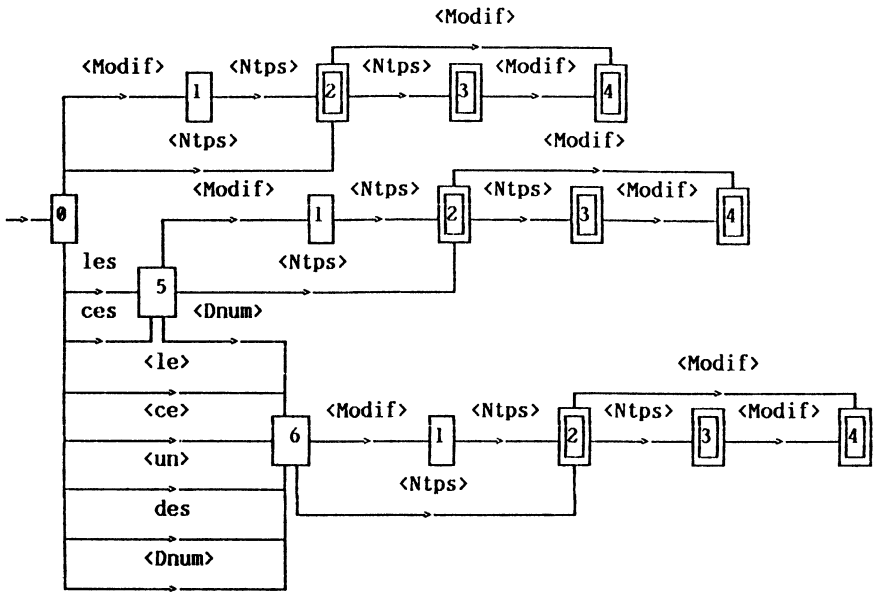

Figure 9

Il faudrait donc compléter les automates des figures 5, 6 et 9 par l'éventuel parcours de l'automate de la figure 10 :

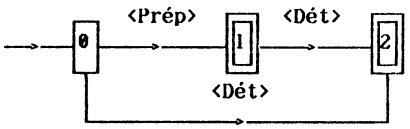

Figure 10

C'est le rassemblement de ces quatre automates en un seul qui fournira l'automate final, utilisé pour reconnaître les exemples du paragraphe 1. Cette reconnaissance exige parfois plusieurs passages successifs (en particulier pour les conjonctions de coordination et les compléments de nom). Cet automate final se présente donc sous la forme suggérée par la figure 11.

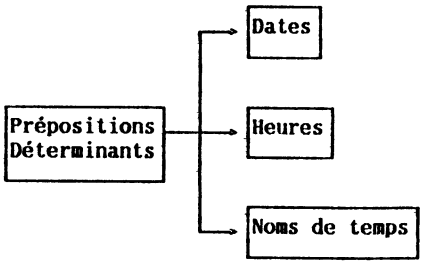

Figure 11

4. TABLES DE TRANSITIONS

Pour permettre à la fois la maintenance et l'adaptation à telle ou telle application, les automates qui viennent d'être présentés l'ont été sous une forme simple, compréhensible à la fois des linguistes et des informaticiens. Pour cela, les étiquettes désignant les transitions sont réduites à un minimum de symboles : des classes sémantiques (par exemple : *<Ntps>*) ou syntaxiques (par exemple : *<Prép>*). Mais il est évident que l'utilisation de ces classes ne peut garantir l'existence des transitions correspondantes pour chaque terme de la classe concernée.

Par exemple, les deux séquences : *en semaine* et * *en mois* correspondent au chemin constitué des transitions étiquetées *<Prép>* et *<Ntps>* de l'automate de la figure 8, mais la première est correcte alors que la seconde ne l'est pas (ce qui est symbolisé, dans les exemples cités, par une étoile). Etant donné la richesse et la diversité de la langue naturelle, il est probable que si on voulait constituer des classes véritablement homogènes pour toutes les propriétés lexicales, syntaxiques ou sémantiques rencontrées, on arriverait au terme de l'étude avec des classes comprenant... un ou deux éléments !

A titre de comparaison, un travail de grande envergure a été réalisé au Laboratoire d'Automatique Documentaire et Linguistique (L.A.D.L.) sur les verbes du Français : cent dix propriétés syntaxiques de trois mille verbes ont été étudiées de manière exhaustive. En prenant comme principe de constitution d'une classe de verbes l'égalité complète des dites propriétés, il en est ressorti deux mille classes différentes ! [5].

Bien sûr, il serait possible (théoriquement) d'étiqueter les transitions non pas par des classes de mots, mais par les mots eux-mêmes. Ainsi la transition étiquetée *en* arriverait sur un état d'où partirait une transition étiquetée *semaine*, mais pas de transition étiquetée *mois*. Il est à noter cependant que la représentation d'un travail aussi complexe que la comparaison terme à terme de chaque mot par un automate compliquerait certainement tant la mise au point que la maintenance d'un tel système, sans parler de la simple présentation des résultats ! C'est pourquoi il a paru plus judicieux de constituer un automate à partir de certaines classes sémantiques ou syntaxiques et de réserver pour un deuxième temps la vérification terme à terme de la validité de la reconnaissance par l'automate. Celle-ci opère à partir d'une base de donnée qui contient les informations nécessaires sous la forme de tables de transitions. Ces tables sont des matrices binaires indicées par tous les éléments des classes concernées. Un signe + à l'intersection d'une ligne et d'une colonne correspondra à l'existence des transitions correspondantes, un signe - au contraire. Par exemple, une fois reconnu par l'automate le chemin *<Prép> <Ntps>*, les séquences en (*semaine* + **mois*) entraînent la consultation de la table dont voici un extrait :

A.....semaine C.....été
B.....mois D.....siècle

<Prép> <Ntps>	ABCD
dans	----
en	+-+-
par	++++

C'est le signe + à l'intersection de la première colonne (*semaine*) et de la deuxième ligne (*en*) qui autorise la première séquence et c'est le signe - à côté qui refuse la deuxième. Dans cet extrait de table sont aussi répertoriés l'acceptation des séquences :

en été + par (semaine + mois + été + siècle)

et le refus des séquences :

* *dans (semaine + mois + été + siècle) + en siècle*

Pour des séquences plus complexes, plusieurs tables sont consultées successivement. Par exemple la séquence :

vers la fin du dernier mois de l'année 1988

correspond à une double lecture de l'automate (avec décomposition de *du* en *de le*) :

<Prép> < Dét> <le> <Modif> <Ntps>
de
<le> <Ntps> <Modif>

et va exiger la consultation de quatre tables pour vérifier la compatibilité de :

dernier et *mois* *année* et *1988*	dans la table	*<Ntps> <Modif>*
vers et *la fin de*	dans la table	*<Prép> <Dét>*
la fin de et *mois*	dans la table	*<Dét> <Ntps>*
mois et *année*	dans la table	*<Ntps> de <Ntps>*

C'est seulement alors que cette séquence sera véritablement reconnue.

Bien sûr, la constitution de cette base de données représente un tel travail linguistique qu'il est possible de s'interroger sur sa nécessité. Cependant celle-ci possède un autre débouché où elle se révèle sans conteste indispensable : il s'agit de la génération de texte en langue naturelle. En effet, s'il peut paraître (en première approximation) négligeable de savoir, dans la reconnaissance de texte que *en* ne peut être suivi immédiatement de *mois*, cette connaissance devient absolument indispensable en génération de texte, sous peine d'aboutir à un charabia incompréhensible. La recherche en *informatique linguistique* s'oriente donc vers la constitution de telles bases de données, accessible à la fois pour reconnaître, interpréter, traduire ou générer un texte, car toutes ces questions sont bien plus liées qu'il ne le semblaient de prime abord [15].

Peut-être est-il intéressant d'un point de vue épistémologique de citer ici le principe de départ de M. Gross (et du L.A.D.L.). Ayant remarqué que : "toute construction théorique a toujours été précédée par un long travail d'accumulation systématique de données" et que "les chercheurs se sont toujours efforcés de combler les trous qui pouvaient se présenter dans leurs données avant d'avancer une règle générale", M. Gross souhaite l'élaboration de règles linguistiques non plus à partir d'observations isolées, comme c'est le cas actuellement, mais à partir de cette accumulation de données, afin de déboucher sur une véritable *science* de la syntaxe.

5. CONCLUSION

L'automate réalisé présente de façon claire et concise un grand nombre de possibilités combinatoires. Le dictionnaire comporte à ce jour cinq cent quatre-vingts entrées, l'automate est composé de deux cent soixante-quatorze transitions se rapportant à quatre-vingt-trois tables. De plus, malgré la grande quantité d'informations emmagasinées, les modifications restent des opérations simples. Il est tout à fait possible d'ajouter ou de retrancher une transition à l'automate, d'augmenter ou de réduire les classes sémantiques ou syntaxiques qui servent d'étiquettes ou encore de modifier une table de transition. Cette souplesse d'utilisation des bases de données est un des grands avantages de la représentation d'une partie de la langue sous la forme d'automates. Cette méthode semble donc véritablement opérationnelle pour le traitement de la langue naturelle qui échappe le plus souvent à tout recensement exhaustif.

Cependant, si les résultats sont satisfaisants dans un cadre limité, une telle étude ne peut être totalement indépendante du reste de l'analyse de la phrase. Quelques exemples le montreraient facilement. Cela traduit simplement l'impossibilité de décrire l'ensemble de la langue sous la forme d'un automate fini [16]. D'ailleurs, l'automate réalisé ici s'est limité à la description un certain nombre de compléments circonstanciels de temps, ceux dont la syntaxe reste suffisamment simple. En particulier, il n'a pas été question de propositions relatives ou propositions incises. Ce même critère est à l'origine de l'élimination complète de cette étude des propositions subordonnées circonstancielles de temps.

La recherche en *informatique linguistique* s'est donc orientée vers la constitution de *lexiques-grammaires* complets, associant une étude syntaxique précise à chaque entrée du lexique. L'étude des verbes par le L.A.D.L. en est un exemple. Néanmoins, certaines parties du discours, comme par exemple les dates, constituent un domaine particulier, le plus souvent limité et emprunt d'une certaine régularité lexicale et syntaxique, ce qui les rend propices à une description formelle sous la forme d'un automate. Une telle description présente l'avantage de proposer à un analyseur syntaxique une interprétation probable dès le début de l'analyse, ce qui éviterait une dispersion en de multiples hypothèses.

Diverses applications de cette étude sont possibles. Par exemple, l'utilisation de cet automate dans le cadre de l'interrogation de bases de données permettrait des questions relatives aux dates sous une forme proche du langage naturel. Autre exemple, dans le cadre de la traduction automatique utilisant des modules de transfert [17], les séquences représentant une date pourraient, une fois reconnues par l'automate, être traduites directement, sans passer par les différentes représentations, ce qui ferait gagner du temps au système. Une implémentation de cet automate dans le programme EUROTRA de traduction entre les différentes langues de la C.E.E. devrait avoir lieu dans le courant de l'année 1990.

BIBLIOGRAPHIE

[1]. ABBOU A., MEYER T., LEFAUCHEUR I., *Les industries de la langue, les applications industrielles du traitement de la langue par les machines*, Paris, Daicadif, 1987.

[2]. JAYEZ J. H., *Compréhension automatique du langage naturel, le cas du groupe nominal en français*, Paris, Masson, 1985.

[3]. DANLOS L., *Génération automatique de textes en langues naturelles*, Paris, Masson, 1986.

[4]. GROSS M., *Grammaire transformationnelle du français, syntaxe du verbe*, Paris, Larousse, 1968, Malakoff, Cantilène, 1986 (réimpression).

[5]. GROSS M., *Méthodes en syntaxe*, Paris, Hermann, 1975.

[6]. GROSS M., *Grammaire transformationnelle du français, syntaxe du nom*, Paris, Larousse, 1977, Malakoff, Cantilène, 1986 (réimpression).

[7]. PERRIN D., "Automates et algorithmes sur les mots", *Annales des télécommunications*, 44, n°12, (1989), 20-33.

[8]. PERRIN D., SAKAROVITCH J., "Automates finis", *Editions scientifiques de l'UAP*, 1, (1988), 35-51.

[9]. HOCKETT C. F., "Two models of grammatical description", *Word*, 10, (1954), 210-233.

[10]. CHOMSKY N., MILLER G. A., "Introduction to the formal analysis of natural languages", *Handbook of Mathematical Psychology*, vol. 2 (1963), 269-322, New-York, Wiley, Paris, Gauthier-Villars, 1968 (traduction française).

[11]. COURTIN J., "Organisation d'un dictionnaire pour l'analyse morphologique", *IRMA* (1973), Grenoble.

[12]. COURTIN J., *Algorithmes pour le traitement interactif des langues naturelles*, Grenoble, Thèse d'Etat, 1977.

[13]. MAUREL D., *Reconnaissance de séquences de mots par automate, adverbes de date du Français*, Paris, Thèse de Doctorat (Université Paris VII), 1989.

[14]. MAUREL D., "Grammaire des dates, étude préliminaire à leur traitement automatique", *Linguiticae Investigationes*, 12, (1988), n°1, 101-128.

[15]. PITRAT J., *Textes, ordinateurs et compréhension*, Eyrolles, Paris 1985.

[16]. CHOMSKY N., *Aspects of the Theory of Syntax*, Cambridge, MIT press, 1965.

[17]. DANLOS L., "Génération automatique de textes en langue naturelle", *Annales des télécommunications*, 44, (1989), n°12, 94-100.