Description d'expressions numériques en français

Matthieu CONSTANT

Abstract : This article aims at describing numerical noun phrases containing unit names as heads, such as the French *une distance de 30 mètres*. This description has been carried out in terms of a graph representation. First, a general model recognising a large number of utterances has been built. Then, in order to improve adequacy, specialised graphs for the most frequent units has been developed. At this stage, results are encouraging, but it has been shown that they could be improved in many ways.

Keywords: finite-state graph, number, numerical noun phrases, unit names.

Mots clés : expression numérique, graphe, nombre, noms d'unité.

1. Introduction

Les nombres se remarquent par leur fréquence très élevée dans les corpus de journaux. On les retrouve dans de multiples types d'expressions. Paradoxalement, très peu d'études linguistiques ont été menées dans ce domaine particulier. Dans cet article, nous nous consacrons à la description d'expressions numériques contenant une unité accompagnée d'un numéral : *une distance de 30 mètres*, par exemple. Ce type d'expression apparaît couramment.

Nous proposons une représentation des expressions numériques par graphes qui possède de nombreux intérêts. Tout d'abord, elle permet une recherche automatique dans les textes. Ensuite, le principe

✉ Matthieu CONSTANT , Laboratoire d'Automatique Documentaire et Linguistique
Université de Marne-la-Vallée, Bâtiment Copernic, 4e étage, Bureau 4B098
Tel : 01 60 95 77 38. Fax : 01 60 95 75 57 e-mail : mconstant@univ-mlv.fr

d'appel à des sous-graphes à l'intérieur d'un graphe permet une forte modularité. Par contre, il demande une organisation rigoureuse et soignée similaire à celle de la programmation informatique classique. Si cette approche est correctement utilisée, la représentation par graphes garantit une excellente lisibilité.

Cet article comporte quelques généralités sur les expressions numériques qui nous intéressent. Ensuite, nous définirons ce que l'on appelle un nombre puis une unité. À partir de là, nous développerons une représentation générale. Ensuite, constatant une amélioration possible de la qualité de la description, nous spécialiserons les graphes pour certaines unités. Enfin, nous présenterons les résultats obtenus par application de ces graphes à certains corpus et nous énumérerons les problèmes rencontrés et les améliorations possibles.

Les graphes ont été établis à l'aide du logiciel INTEX (SILBERZTEIN 1993) utilisant les dictionnaires DELAF et DELAC (COURTOIS 1990) et ont été appliqués à de gros corpus grâce au logiciel AGLAE (PAUMIER et CONSTANT 2000).

2. Généralités sur les expressions numériques

Il existe différents types d'expressions numériques telles que:

– les dates : *le 17 juin 2000*;
– les durées : *douze mois*;
– les expressions monétaires : *236 francs belges*;
– autres : *tel : 05 59 00 67 82*; *page 12*.

Dans notre étude, nous nous limiterons aux expressions numériques contenant une unité. Ces expressions contiennent deux éléments de base, un nombre et une unité, auxquels nous pouvons greffer un contexte droit et gauche. Grossièrement, nous pouvons envisager de représenter une expression numérique de la manière suivante :

(E + Contexte gauche) (Nombre) (E+Unité) (E + Contexte Droit)

Nous tenterons, dans cet article, de généraliser leur représentation afin d'en reconnaître le plus possible. Ensuite, nous spécialiserons les

représentations pour certaines unités très fréquentes afin d'obtenir des résultats plus fins. Nous essaierons d'allier quantité et qualité.

Afin d'améliorer la lisibilité de cet article, nous utiliserons le terme « expressions numériques » pour désigner les expressions qui nous intéressent.

3. Les nombres

Pour partir d'une bonne base, il faut, tout d'abord, définir précisément ce qu'est un nombre. Les nombres ont été catégorisés sous 6 formes différentes :

– les suites de chiffres : ***NombreExact*** (graphe 1 ci-dessous)

ex : *30* ; *124,4* ; *1 020.*

Dans le graphe 1, ***Chiffres*** est un chiffre 0, 1, ... ou 9. ***3Chiffres*** est un groupe de trois chiffres collés (sans espace). # signifie « pas d'espace ».

– les entiers en toutes lettres : ***NombreEnLettre***

ex : *quatre-vingt deux* ; *neuf cent douze* ; *dix-sept mille deux cent trente-trois.*

– les nombres approximatifs en toutes lettres : ***NombreImprecis*** (graphe 2 ci-dessous)

ex : *Quelques dizaines de milliards.*

Dans le graphe 2, ***MultSousMult*** représente des multiples et des sous-multiples tels que *milliard* ou *dixième.*

– les formules scientifiques : ***NotationScientifique***

ex : *1,2x10-5*

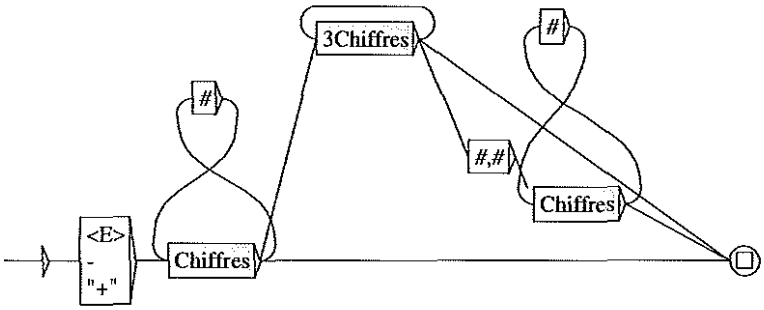
– les coordinations de nombres (en lettres et en chiffres) : ***AutreNombre***

ex : *cinq ou six* ; *12x4* ; *12 et 5* ; 45/2.

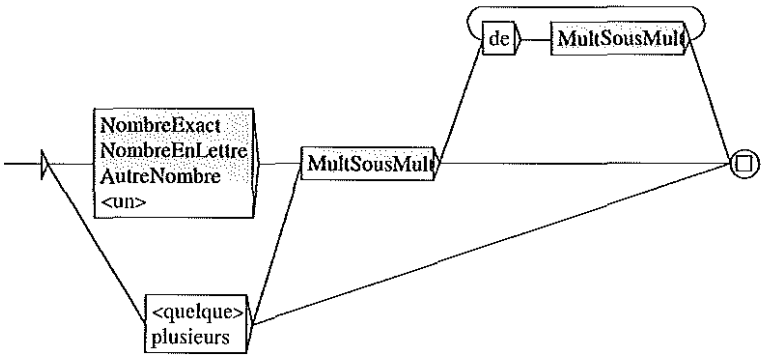
– les fractions en toutes lettres : ***Fractions***

ex : *deux tiers.*

Remarque : <E> est l'élément vide.



Graphe 1 : NombreExact



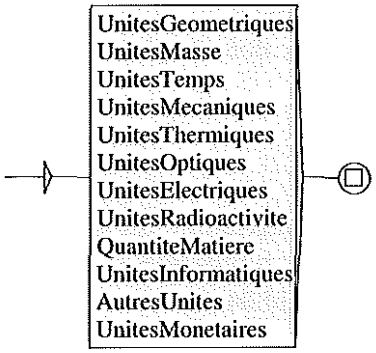
Graphe 2 : NombreImprecis

4. Les unités

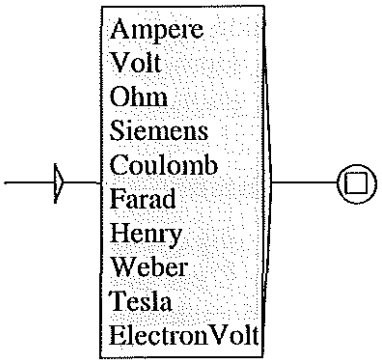
4.1. Unités simples

Les nombres ayant été définis, il faut maintenant définir les unités. Dans un premier temps, nous avons répertorié les unités de mesure en toutes lettres comme *millimètre* puis ces mêmes unités sous forme d'abréviations (*mm* pour *millimètre*). Enfin nous avons étendu notre liste aux unités monétaires comme *dollar australien*.

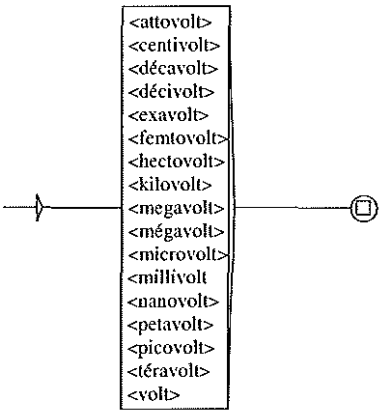
Les unités de mesure étant catégorisées par domaine scientifique, nous avons logiquement gardé cette classification dans les graphes. Prenons l'exemple simple de *millivolt*. *Millivolt* est une sous-unité de *volt* qui fait partie des unités électriques.



Graphe 3 : unités simples en toutes lettres



Graphe 4 : unités électriques



Graphe 5 : Volt

Remarque : nous écrivons les unités sous leur forme canonique, *<téravolt>* par exemple, en supposant qu'elles se trouvent déjà dans les dictionnaires que nous appliquerons. Cela allège les graphes où l'on écrit toutes les formes : *téravolt* et *téravolts* dans notre exemple.

4.2. Unités complexes

À partir de ces unités simples, il est possible de générer des unités complexes telles que *mm/s* ou *francs par minute*. Nous avons développé un graphe générant toutes les unités complexes possibles contenant une ou plusieurs unités simples. Il décrit plus d'unités qu'il n'en existe réellement. Par exemple, l'expression *pascal par hectopascal*

est décrite alors qu'elle ne devrait a priori jamais apparaître dans un texte. Notre approche étant celle de l'analyse, nous négligerons cette propriété du graphe.

5. Représentation générale

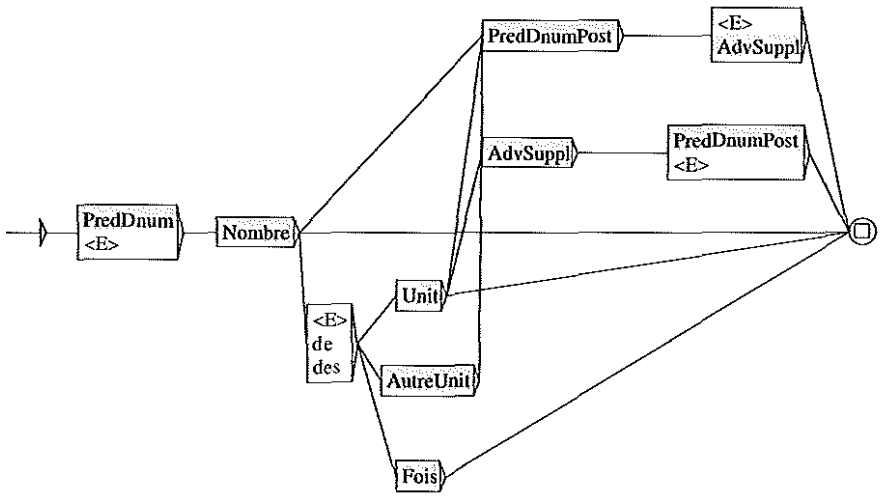
Le premier objectif est de reconnaître le maximum d'expressions numériques. C'est pourquoi nous les avons représentées sous une forme très générale valable dans la quasi-totalité des cas.

5.1. Expressions simples

Pour commencer, nous pouvons écrire une expression numérique sous sa forme la plus simple

(Nombre) (E+Unité)

À cette représentation, nous pouvons alors greffer des graphes d'adverbes et prédéterminants, construits par Maurice Gross, modulant l'expression à droite et à gauche (voir graphe 6).



Graphe 6 : expression numérique simple

PredDnumPost liste les prédéterminants se trouvant après le déterminant numérique (***Dnum***) : *de moyenne* ; ***PredDnum*** les prédéterminants avant le *Dnum* : *environ*. Enfin, ***AdvSuppl*** reconnaît des

adverbes supplémentaires tels que *en plus*. Le sous-graphe ***Fois*** reconnaît les expressions telles que *fois plus élevé*. ***Unit*** reconnaît les unités décrites précédemment. ***AutreUnit*** décrit les structures comme *chances sur quelques milliards*. Nous obtenons ainsi des expressions telles que *en gros 10 000 volts en plus de moyenne*.

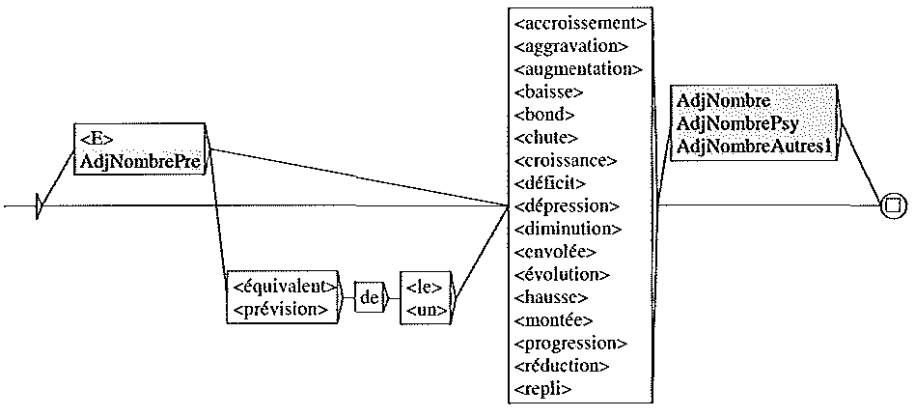
5.2. Expressions complexes

Les expressions simples décrites précédemment peuvent être utilisées au sein d'expressions plus complexes. Elles peuvent être reliées par des conjonctions de coordination et des prépositions sous différentes formes comme :

– *de* (ES) [1] *à* (ES) ;
– *entre* (ES) *et* (ES) ;
– *à partir de* (ES).

Nous obtenons ainsi des expressions telles que *entre en gros quelques centaines et quelques milliers de volts environ*.

Il est aussi possible d'ajouter des groupes nominaux (*GN*) spécifieurs à gauche de ces expressions. Le but est de trouver des noms assez généraux valables pour toutes les unités. Nous nous sommes aperçu que les noms désignant une évolution s'adaptent parfaitement. Ils sont décrits dans le graphe 7 ci-dessous.

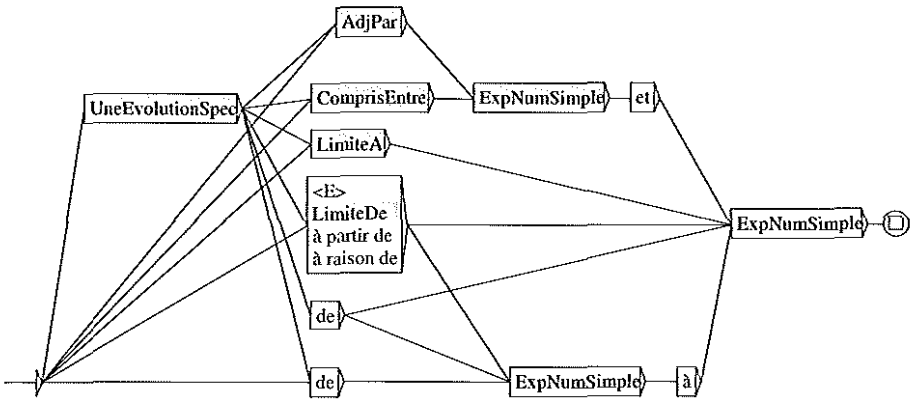


Graphe 7 : GN d'évolution

[1] (ES) =Expression Simple.

Il est important de noter que nous avons greffé à droite et à gauche quelques graphes d'adjectifs spécifiques aux nombres (construits par Maurice Gross). Ainsi, ce graphe reconnaît des expressions telles que *un accroissement spectaculaire*. Par ailleurs, nous avons ajouté d'autres *GN* spécifieurs ayant pour base des noms comme *record* ou *différence*, tous accompagnés par les adjectifs décrits précédemment.

La synthèse des deux structures proposées est exposée dans le graphe 8 ci-dessous représentant les expressions numériques complexes. Il reconnaît des expressions telles que *une chute de la précision comprise entre 10 et 20 micromètres environ*. Le sous-graphe ***UneEvolutionSpec*** désigne les *GN* spécifieurs. ***LimiteA*** reconnaît les expressions comme *supérieur ou égal à*. ***LimiteDe*** décrit les structures telles que *inférieure de*. Enfin, ***AdjPar*** représente des expressions du type de *borné par*.



Graphe 8 : expressions numériques complexes

Remarque : les expressions complexes exposées ci-dessus peuvent encore être complexifiées en les reliant par les conjonctions de coordination *ou*, *contre*, *au lieu de*, etc.

6. Spécialisation pour certaines unités

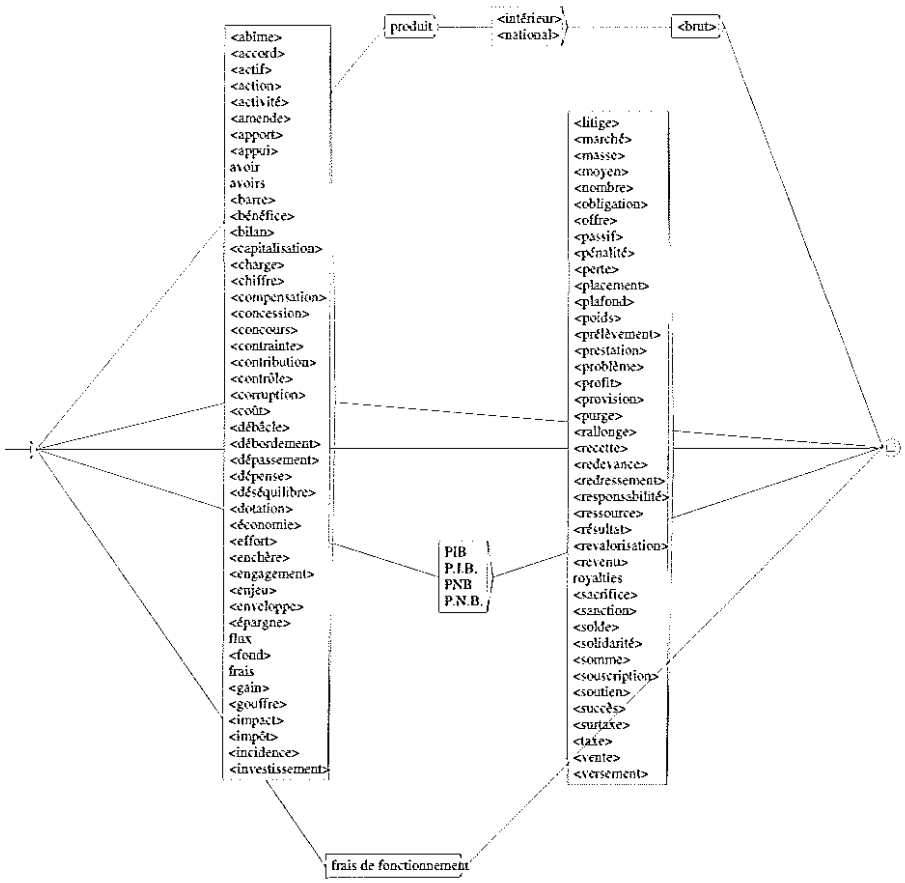
En appliquant les graphes généraux précédents sur de petits textes, nous avons observé une excellente efficacité sur les articles scientifiques comme ceux de *Science et Vie*. Cependant, malgré les résultats

encourageants observés sur les articles d'ordre général comme ceux du *Monde*, il nous a semblé qu'une amélioration était envisageable. En effet, certaines unités comme les unités métriques ou monétaires reviennent très fréquemment. Il serait donc extrêmement bénéfique d'affiner les expressions générales pour ces unités. Par exemple, *un budget de 15 millions de francs* est un type d'expression qui apparaît très souvent dans les corpus. La description stricte des expressions numériques ne permet de reconnaître que *15 millions de francs*. Ainsi, l'information peut être précisée. Une représentation plus spécialisée pour les expressions monétaires permettrait d'identifier l'expression entière.

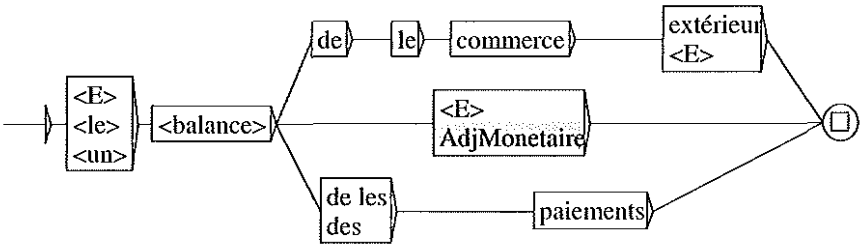
Dans cette étude, nous nous sommes limité aux expressions contenant des unités métriques et monétaires et des vitesses.

Les groupes nominaux spécifieurs de chaque unité doivent être décrits. Nous avons utilisé une méthode extrêmement simple. Pour chaque unité concernée, nous avons recherché tous les noms susceptibles d'être inclus dans des expressions contenant cette unité comme *investissement* (unités monétaires) ou *longueur* (unités métriques) (voir graphe 9 ci-dessous). Puis nous leur avons greffé à droite et à gauche des adjectifs spécifiques aux nombres (cf. graphe 7).

Pour certains noms, nous nous sommes aperçu qu'il était possible d'affiner encore plus. Ainsi, nous avons développé des graphes spéciaux pour quelques noms fréquents comme *budget*. Ainsi, nous pouvons reconnaître des formes telles que *budget annuel 1999 pour le financement des villes de plus de 10 000 habitants*. Ces graphes étant pour la plupart difficilement lisibles tant ils sont riches, il nous a paru préférable de montrer un graphe simple. Le graphe 10 ci-dessous décrit les structures ayant comme nom de base *balance*. En greffant ces graphes aux *GN* spécifieurs généraux (cf. graphe 7), nous obtenons des expressions comme *une augmentation de capital de 2,4 millions de francs*. Par ailleurs, il est possible de mettre les *GN* spécifieurs spécialisés à droite des unités afin de reconnaître des formes telles que *baisse de 78 F de la redevance*. De même, nous avons mis, à droite des unités, des structures spécifiques à chaque unité telles que *à l'ouest de (de la frontière de la ville de Nville)* pour les unités métriques.



Graphe 9: Noms monétaires

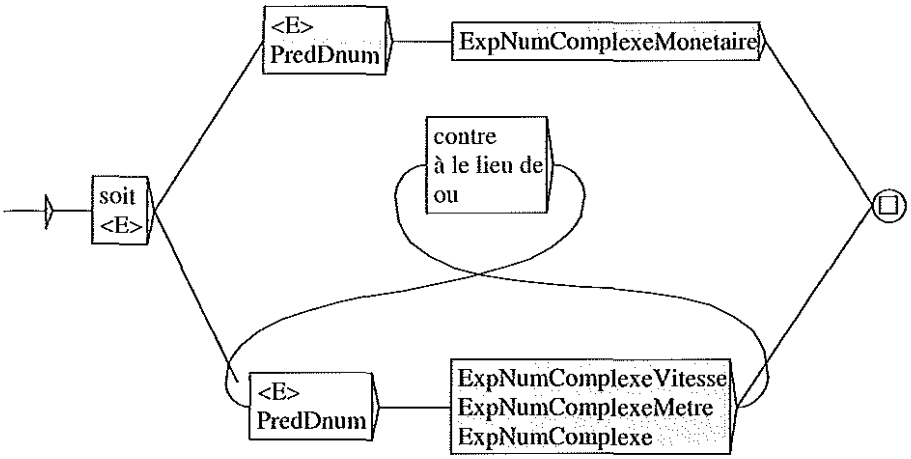


Graphe 10 : balance

Remarque : comme l'ensemble des *GN* spécifieurs pour les unités monétaires est inclus dans celui des *GN* spécifieurs pour les pourcentages, nous avons, temporairement, inclus le symbole % dans les unités monétaires.

7. Résultats, interprétations et améliorations possibles

Nous allons dans cette partie exposer les résultats obtenus. Nous avons appliqué le graphe général graphe 11 ci-dessous au texte du *Monde* pour l'année 1994 (130 Mo) et à quelques articles de *Science et Vie*. Dans ce paragraphe, nous avons classé les résultats obtenus en fournissant quelques explications quand nécessaire, puis nous avons énuméré les différentes erreurs et améliorations possibles.



Graphe 11 : graphe général

7.1. Résultats

Nous classons les expressions reconnues selon deux catégories différentes :

– déterminants numériques (*Dnum*)
– groupes nominaux (*GNnum*)

7.1.1. *Déterminants numériques*

Cette catégorie ne contient aucune unité. Elle ne représente que les expressions numériques les plus basiques. Cependant, il nous apparaît utile de nous attarder sur le sujet.

Tout d'abord, nous pouvons diviser cette catégorie selon deux situations : (*a*) *Dnum* suivi d'un *GN*; (*b*) *Dnum* suivi d'un *GN* effacé.

(*a*) Dnum suivi d'un *GN*

Les exemples (1a) et (1b) sont des plus classiques et n'ont pas besoin d'être expliqués :

(1a) *des centaines de mutilés à vie*
(1b) *3 285 700 demandeurs de emploi*[1]

Par contre, l'exemple (1c) est plus intéressant. Nous constatons que *trois* est suivi du pronom *autres* qui est lié au groupe nominal *grandes unités de volontaires arméniens* qui pourrait le remplacer (1c-bis). Il serait sans doute utile d'élaborer dans le futur une méthode automatique liant le pronom *autre* à son groupe nominal correspondant, ce qui n'est pas une mince affaire.

(1c) *En 1914, les Russes mirent sur pied quatre grandes unités de volontaires arméniens et trois autres en 1915.*
(1c-bis) *En 1914, les Russes mirent sur pied quatre grandes unités de volontaires arméniens et trois grandes unités de volontaires arméniens en 1915.*

(*b*) *Dnum* suivi d'un *GN* effacé

Cette structure ressemble à celle de l'exemple (1c), à l'exception près qu'il n'y a pas de pronom remplaçant le groupe nominal effacé. Ce *GN* est simplement supposé. L'exemple (2) pourrait être transformé en (2-bis) De même que précédemment, une méthode automatique de recherche des *GN* effacés devrait être élaborée.

(2) *La France compte 3 285 700 demandeurs de emploi, soit 3 200 de plus que en octobre.*
(2-bis) *La France compte 3 285 700 demandeurs de emploi, soit 3 200 [demandeurs de emploi] de plus que en octobre.*

[1] Forme normalisée du nom composé *demandeur d'emploi.*

7.1.2. *Groupes nominaux numériques*

Les expressions reconnues de ce type sont les plus significatives. Le résultat de l'application du graphe « généraliste » ***ExpNumComplexe*** à *Sciences et Vie* montre l'efficacité d'une généralisation sur les expressions scientifiques. En voici quelques exemples variés :

(3a) *environ 2 000 km2*
(3b) *autour de une valeur moyenne de 130 GeV*
(3c) *1,67 milliardième de milliardième de milliardième de kg*
(3d) *à raison de 66 milliards par centimètre carré et par seconde*
(3e) *de 50 à 80 millions de mètres cubes à le maximum*
(3f) *mille à dix mille fois plus élevées*
(3g) *1,6 x 10-27 kg*
(3h) *une chute quotidienne de 1 % à 2 %*

Par ailleurs, la spécialisation des graphes affine efficacement les expressions monétaires (4) et métriques (5). Voici quelques résultats de l'application du graphe 9 au *Monde 94*.

(4a) *une centaine de kilomètres à l'est (de Sanaa)*
(4b) *3 200 mètres de altitude*
(4c) *une distance de environ 10 à 15 kilomètres*
(5a) *un endettement de l'ordre de 8 milliards de marks*
(5b) *un montant de une douzaine de millions de francs par an*
(5c) *10 F à le lieu de 30 F*
(5d) *seulement 20 % des crédits à les petites et moyennes entreprises*

7.1.3. *Autres*

Bien que dans la grande majorité des cas, nous obtenions des *Dnum* ou des *GNnum*, nous avons aussi quelques adjectifs numériques tels que (6). Nous ne nous attarderons pas sur le sujet car, pour les décrire correctement, il suffira d'écrire leur grammaire :

(6) *un totem de granit haut de mille mètres seulement*

7.2. Erreurs et améliorations possibles

Il subsiste de nombreuses erreurs et améliorations possibles que nous avons répertoriées.

7.2.1. *Les mots composés*

Certaines expressions reconnues semblent aberrantes telles que (7). Par un système de concurrence efficace, ces erreurs pourraient être évitées. En effet, *avions F-16* pourrait être considéré comme un mot composé (COURTOIS 1990) et être prioritaire par rapport à *–16 de l.*

(7) *des avions F-16 de l'armée ont photographié (...)*

7.2.2. *Les dates*

Dans notre représentation, les dates ne sont pas prises en compte. Cependant, elles s'introduisent dans les expressions numériques sous de multiples formes. Tout d'abord, les incises de dates telles que (8) sont extrêmement fréquentes. Le problème est assez facile à gérer car ce sont les mêmes types d'incises qui reviennent ; de plus, il existe des graphes de *Dates* élaborés par Maurice Gross, qu'il serait très facile d'insérer dans nos graphes. L'exemple (8) devrait alors être reconnu sous la forme (8-bis).

(8) *Déjà tombé de 10 % en 1986 à 5,1 % en 1991, le taux de chômage (...)*

(8-bis) *Déjà tombé de 10 % (en 1986) à 5,1 % (en 1991), le taux de chômage (...)*

Il existe une autre amélioration possible. En supposant que les dates soient reconnues par les graphes de Maurice Gross, nous pouvons tenter d'affiner la description des nombres. En effet, les nombres écrits en chiffres semblent posséder une syntaxe bien particulière. Ils sont regroupés par groupes de trois chiffres collés entre eux. Les groupes, quant à eux, sont séparés par un espace : *1 345 678* par exemple. Dans l'exemple (9), l'expression *5754 a* est reconnue car *5754* est considéré comme un nombre et *a* comme l'abréviation de *ares*. Or, d'après la syntaxe standard, *5754* n'est pas un nombre, donc, l'expression n'est pas reconnue. De même, (10) n'est pas pris en compte :

(9) *leur année 5754 a débuté le 16 septembre*

(10) *la forte expansion du secteur crédit de l'établissement entre 1988 et 1991*

Remarque : dans le cadre de la description des dates, les suites de quatre chiffres consécutifs sont des numéraux d'année, ils sont alors reconnus comme tels.

7.2.3. *Les expressions syntaxiquement fausses*

Quelques expressions prises sans contexte paraissent correctement reconnues dans (11), *de 50 à 52 à l'heure*. Cependant, remises dans le contexte, elles sont incorrectes. En effet, *à l'heure actuelle* est un adverbe figé qu'il suffira de reconnaître. Pour l'instant, ce genre d'erreurs, n'est pas traité et le sera, dans le futur, à l'étape de la levée d'ambiguïté.

(11) *De 50 à 52 à l'heure actuelle*

7.2.4. *Petites améliorations*

Les langues naturelles étant en perpétuelle évolution, il est nécessaire de fréquemment mettre à jour nos graphes comme pour les exemples (12). (12a) est l'exemple typique d'un petit affinement : *taxes professionnelles* est un nom composé qu'il suffit de reconnaître par application du dictionnaire DELAC (COURTOIS 1990) ou par amélioration des listes représentant les GN monétaires (cf. 6). Notons aussi l'ambiguïté de (12a) : *entre 20 et 25 millions de taxes* signifie soit *un grand nombre de taxes*, soit *entre 20 et 25 millions (de+en+sous forme de) taxes*. (12b) montre le besoin d'étendre la description spécialisée pour les expressions contenant un pourcentage.

(12a) rapporter à la commune entre 20 et 25 millions de taxes professionnelles

(12b) *un taux de croissance annuel moyen de 8,1 %*

Par ailleurs, le cas *Dnum N sur Dnum* de (13) est courant. Nous pourrions ajouter un graphe comme :

(Nombre) (E+<A>)(<N>)(E+<A>) (sur) (Nombre).

(13) *(Deux jeunes sur trois) souhaitent se marier.*

7.2.5. *Ajout de verbes et d'adjectifs*

Il est aussi possible d'améliorer la représentation en élargissant aux verbes (14) et aux adjectifs (15) :

(14) *il faut (ajouter les 110 milliards de francs de l'emprunt Balladur).*
(15) *chômeurs (âgés de plus de cinquante-cinq ans)*

8. Conclusion

Nous avons représenté des expressions numériques sous forme de graphes. Nous les avons modélisées en deux temps :

- une formulation aussi générale que possible pour couvrir le plus grand nombre de cas ;
- une formulation spécialisée pour certaines unités très fréquentes afin d'affiner les expressions reconnues.

Nous avons vu qu'il existait de nombreuses améliorations à réaliser. Cette étude n'a pas la prétention de décrire toutes les expressions numériques, mais elle présente une base pour un plus vaste chantier. En effet, ces graphes reconnaissent des séquences de manière anarchique linguistiquement parlant. Elles pourraient être réorganisées selon des catégories grammaticales et sémantiques. Pour l'instant, nous ne reconnaissons, en gros, que les déterminants et les groupes nominaux. Dans le futur, il faudrait élargir aux structures verbales et, ainsi, aux phrases complètes. La traduction automatique étant l'un des objectifs visés, il pourrait être utile d'appliquer la même démarche aux expressions numériques de l'anglais. Ainsi, la correspondance entre les deux représentations constituerait la base d'un système de traduction simple et efficace.

Bibliographie

COURTOIS (Blandine) : 1990, « Dictionnaires électroniques du français, *Langue française*, 87 (Paris : Larousse), p. 11–22.

GROSS (Maurice), « The use of finite automata in the lexical representation of natural language », in GROSS (M.) et PERRIN (D.) eds., *Electronic Dictionaries and Automata in Computational Lin-*

guistics, Lecture Notes in Computer Science 377 (Paris : Springer-Verlag), p. 34–50.

GROSS (Maurice) : 1997, « The construction of local grammars », in ROCHE (E.) et SCHABES (Y.) eds., *Finite-state language processing* (Cambridge, Massachusetts : A Bradford Book), p. 329–354.

PAUMIER (Sébastien), CONSTANT (Matthieu), *A corpora processing software : AGLAE.* [http://www.ladl.jussieu.fr]

SILBERZTEIN (Max) : 1993, *Dictionnaires électroniques et analyse automatique des textes. Le système INTEX* (Paris : Masson).