Analyse et représentation d'expressions de mesure

3.1 Introduction

Dans cette partie, nous décrivons et représentons des expressions numériques et plus particulièrement des expressions de mesure[23], dans le but de les reconnaître automatiquement dans les textes de langue générale (quotidiens d'informations comme *Le Monde* et journaux de vulgarisation scientifique type *Science et Vie*)[24]. De précédentes versions de ce travail ont été publiées dans M. Constant (2000, 2002a).
Nous employons le terme expression de mesure pour toute séquence linguistique contenant la sous-séquence de base *Dnum Unité* (ex : *10 m*) où *Dnum* est un déterminant numérique cardinal. Nous distinguons deux types de mesure, les mesures « absolues » et les mesures « relatives » qui entrent respectivement dans les structures suivantes où le symbole *Ng* désigne un nom de grandeur :

(1) *N*0 *avoir un Ng de Dnum Unité* = : *la corde a une longueur de 10 m*
(2) *N*0 *Vsup Prép un Ng de Dnum Unité Prép*1 *N*1[25] = : *le couteau forme un angle de 10° avec la fourchette*

Les mesures absolues sont des mesures de caractéristiques ou propriétés (*Ng*) propres à l'argument *N*0, comme la taille ou la durée :

Max a une taille de 1,71 m
Le spectacle a une durée de trente minutes

Les mesures relatives mesurent une caractéristique ou propriété (*Ng*) de *N*0 par rapport à *N*1, comme la distance

[23] Des outils statistiques d'extraction d'expressions de mesures existent (R. Agrawal et R. Srikant, 2002).
[24] Nous n'étudions pas les discours techniques.
[25] *Prép* peut être vide.

Paris est à une distance de 600 km de Bordeaux

Nous intégrons également les expressions de pourcentage dans cette dernière catégorie, qui mesure l'inclusion d'un ensemble (*N*0) dans un autre (*N*1) :

Les étudiants représentent 10% de la population

Nous faisons aussi quelques remarques sur des séquences qui expriment une comparaison « relative » telles que dans la phrase suivante :

Max est deux centimètres plus grand que Luc

Dans cette section, nous décrivons en détail le comportement syntaxique de ces expressions dans des phrases élémentaires, puis dans les formes réduites de ces phrases. Nous en donnons des représentations simples sous la forme de grammaires locales et de tables syntaxiques. Cette étude a été réalisée sur le français et partiellement sur l'anglais. Nous attachons une grande importance à la réalité linguistique du contenu des textes. Notre travail est basé sur les études de M. Silberztein (1993) et A. Chrobot (2000) sur les déterminants numériques, de J. Giry-Schneider (1991) sur les phrases élémentaires représentant une mesure absolue[26] et de P. A. Buvet (1993, 1994) sur les déterminants nominaux.

3.2 Les composants élémentaires

3.2.1 Généralités

Nous avons défini une expression de mesure comme une séquence comportant la séquence *Dnum Unité* où *Dnum* désigne un déterminant numérique et *Unité* une unité de mesure. Nous souhaitons, dans un premier temps, décrire, de manière détaillée, chaque composant simple de cette séquence. Les déterminants numériques sont les plus variés et nous utilisons une typologie formelle. Il peut s'agir de :

- déterminants indéfinis au pluriel (*Dind-pl*) comme *des*, *plusieurs*, *quelques*, etc.
- déterminants numériques cardinaux simples ou composés écrits en lettres (ex : *douze* ; *quarante-trois*)
- séquences de chiffres arabes décrivant des nombres réels dans un format standard (ex : *1 905* ; *1,78*) ou dans un format scientifique (ex : *1,54.10+5*).
- déterminants nominaux de la forme *Det Nnum de* (=: *des milliers de*) où *Nnum* est un nom que l'on qualifiera de numérique comme *milliers*, *dizaines*, etc.

Nous étudions spécifiquement les trois derniers types. Après avoir examiné les différentes classes d'unités simples que nous utilisons, nous évoquons le cas des prédéterminants numériques qui sont essentiels pour une reconnaissance fine des mesures car ils introduisent de légères modifications sémantiques (ex : *à peu près dix ampères* ≠ *exactement dix ampères*).

Nous analysons également le schéma de phrase *Det Ng être de Dnum Unité* (=: *la longueur est de 30 m*) qui est commun aux deux structures de mesure que nous allons étudier. Cette étude va nous permettre de construire des graphes élémentaires de mesure à partir des graphes d'unités.

[26] Des travaux ont aussi été réalisés par A. Borillo (1985, 1998).

3.2.2 Graphes des déterminants numériques

3.2.2.1 Les déterminants numériques cardinaux écrits en lettres

Nous traitons brièvement ce point car les déterminants numériques cardinaux écrits en lettres ont déjà été étudiés et décrits sous la forme de graphes par M. Silberztein (1993) pour le français et A. Chrobot (2000) pour l'anglais. Dans cette section, nous reprenons les points importants de l'étude sur le français. Nous notons *DnumEnLettres* ce type de déterminants numériques, i.e. les nombres entiers[27] écrits en toutes lettres (borne supérieure : *un billiard*). L'utilisation des sous-graphes a un avantage indéniable car certaines séquences peuvent apparaître plusieurs fois dans un nombre (ex : *douze* dans *douze cent douze*). Certains termes comme *cent*, *mille*, *quatre-vingts* posent quelques problèmes orthographiques car :

- *mille* est invariable ;
- *cent* est au pluriel lorsqu'il est multiplié : *deux cents* ; mais il reste invariable lorsqu'il est suivi d'un autre nombre ou qu'il est utilisé comme centième [Larousse, 2002] : *trois cent vingt, deux mille deux cent* ;
- *Quatre-vingt* est au singulier lorsqu'il est suivi d'un nombre : *deux cent quatre-vingt-deux* ; mais il est au pluriel autrement : *mille quatre-vingts*. [Grevisse, 1975, paragraphe 406]

Ainsi, pour chaque type de nombres entiers (nombres inférieurs à cent, à mille, à un million, etc.), il est nécessaire de construire deux graphes : l'un décrivant ces nombres lorsqu'ils se trouvent dans la partie droite (ou finale) d'un nombre et l'autre décrivant ces nombres lorsqu'ils sont dans la partie gauche. Par exemple, *quatre-vingt(s)* a deux comportements suivant qu'il est à droite (*quatre-vingts*) ou à gauche (*quatre-vingt*) comme dans *quatre-vingt mille deux cent quatre-vingts*.
M. Silberztein (1993) n'a pas décrit les nombres se terminant par *million(s)* et *milliard(s)*. Ces nombres sont suivis de la préposition *de* : *cent vingt millions de*. Ce type de nombres rentre donc dans une structure différente : celle des déterminants nominaux (*Det N de*) décrits dans M. Gross (1986). Notons que nous n'avons pas traité le cas du décimal *un demi* suivi d'un tiret (ex : *une demi-heure*).

3.2.2.2 Nombres écrits en chiffres arabes

Nous notons ce type de déterminant *DnumEnChiffres*. Les nombres écrits sous la forme d'une suite de chiffres ont une syntaxe bien particulière qui diffère en français et en anglais. Une solution simple et naïve pour décrire les entiers naturels est de les représenter comme une suite de chiffres soudés d'au moins un élément. Cependant, cette représentation est trop simpliste. Les nombres avec plus de trois chiffres ne rentrent pas dans ce schéma : par exemple, en français, dans *1 298*, il existe un espace blanc obligatoire[28] entre le troisième chiffre en partant de la droite et le dernier chiffre à gauche ; en anglais, l'espace blanc est remplacé par une virgule (*1,298*). D'une manière générale, un espace blanc (une virgule en anglais) apparaît obligatoirement dans la séquence de chiffres tous les trois chiffres en partant de la droite. En français, l'ensemble des entiers naturels écrits en chiffres arabes peut donc être représenté par le graphe **NombreEntierEnChiffres** ci-dessous. Le graphe **3Chiffres** décrit une suite de trois chiffres soudés et le graphe **Chiffre** l'ensemble des chiffres arabes. Le symbole # indique que l'élément à sa gauche et celui à sa droite sont soudés l'un à l'autre ; autrement dit, tout espace blanc est interdit entre ces deux éléments. Cette description précise permet de lever, dans certains cas, l'ambiguïté qui existe naturellement entre un déterminant numérique et une date désignant une année écrite en chiffres. En effet, ce type de date est une

[27] Les nombres décimaux ne semblent pas pouvoir être écrits en toutes lettres.
[28] Parfois, c'est un point qui apparaît à la place de l'espace.

suite de chiffres collés. Ainsi, *2003* ne sera pas reconnu comme un déterminant numérique par notre grammaire. Par ailleurs, il est usuel d'utiliser des nombres décimaux écrits en chiffres. En français, la partie entière est séparée de la partie décimale par une virgule (*12,7* ou *3,896*). En anglais, c'est un point qui fait office de séparateur (*12.7* ou *3.896*). Dans ce cas, la partie décimale est une simple séquence de chiffres collés d'au moins un élément et est représentée par le graphe **PartieDecimale**. Un nombre quelconque écrit en chiffres arabes est alors reconnu par le graphe **DnumEnChiffres** regroupant les graphes **NombreEntierEnChiffres**[29] et **PartieDecimale**. Notons que ces deux parties (entière et décimale) sont obligatoirement collées à la virgule les séparant. Cela évite, par exemple, de reconnaître *10, 11* dans l'expression coordonnée *10, 11 ou 12 chaises*. Par ailleurs, les nombres peuvent être signés : ils peuvent avoir un + ou un – placé au début (à gauche).

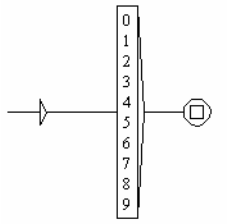

Figure 17 : Chiffre

Figure 18 : 3Chiffres

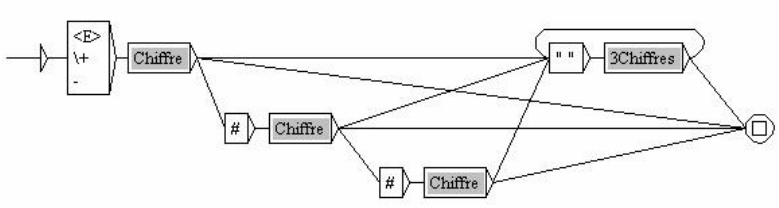

Figure 19 : NombreEntierEnChiffres

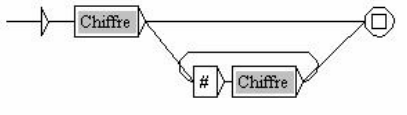

Figure 20 : PartieDecimale

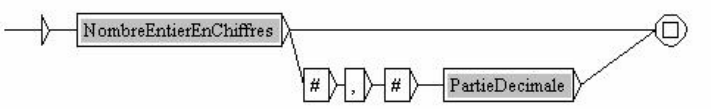

Figure 21 : DnumEnChiffres

Les nombres en notation scientifique possèdent aussi une syntaxe bien particulière. La partie entière ne comprend qu'un seul chiffre. La partie décimale est plus libre en fonction de la précision que l'on souhaite avoir. On ajuste ensuite ce nombre à l'aide d'une puissance de 10 (soit négative, soit positive) :

1,23.10E5 ou 1,23x10+5[30]
4,8x10-6 (dans *Science et Vie*)

Nous représentons de tels nombres dans le graphe **FormuleScientifique** ci-dessous :

[29] Le symbole \ devant + est un symbole de déspécialisation. Le symbole <*E*> est le mot vide.
[30] Nous ne regardons que des textes bruts sans tenir compte de l'enrichissement typographique : 1,23.10-6 serait plutôt écrit $1{,}23.10^{-6}$.

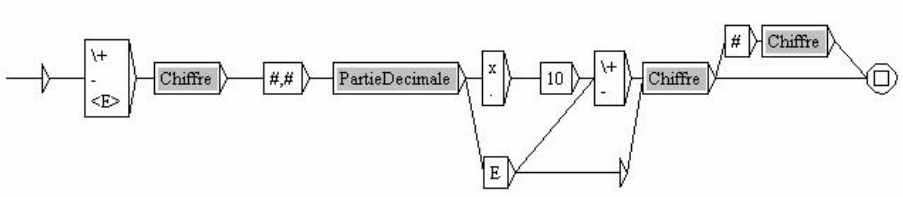

Figure 22 : FormuleScientifique

3.2.2.3 Les déterminants nominaux numériques

Nous regardons maintenant les déterminants nominaux de la forme *Det Nnum de* où *Nnum* est un nom que l'on qualifiera de numérique, tel que *milliers, millions, milliards* :

Dix millions de personnes sont allergiques à la poussière

Notons que le déterminant numérique *dix* avant *millions* peut aussi s'écrire en chiffres. *Det* peut également être un déterminant indéfini pluriel :

10 millions de personnes sont allergiques à la poussière
Des millions de personnes sont allergiques à la poussière

Il est également possible d'utiliser le modifieur numérique *demi* :

Un demi million de personnes sont allergiques à la poussière

Nous définissons les *Nnum* à partir de sa possibilité d'occurrence dans :

- des multiples exacts de 10 : *million(s), milliard(s), billion(s)*

  Cette planète est à 17 milliards d'années-lumière de la Terre

- des sous-multiples[31] de 10 : *dixième(s), centième(s), milliardième(s)*

  Ce robot a une précision d'un millionième de centimètre

- d'autres termes : *millier(s), centaine(s), cinquantaine(s), douzaine(s), dizaine(s)*, etc.

  Marie attend une vingtaine d'amis cette semaine

Il est possible d'insérer un ensemble restreint de modifieurs dans notre séquence comme dans :

Paul a perdu une (E + bonne + petite) dizaine de kilos

Par ailleurs, on peut combiner les structures *Det Nnum de* afin de former des structures plus complexes de la forme *Det Nnum de (Nnum de)**. C'est le cas dans les exemples ci-dessous :

1,67 milliardième de milliardième de milliardième de kg. (revue *Science et Vie*)

[31] Il est également possible d'utiliser des fractions du type *deux tiers* :
Ce robot a une précision de deux tiers de centimètre

La dernière phrase ci-dessus montre que l'on peut combiner différents types de *Nnum*. D'autre part, il existe des variantes semi-figées de ces séquences qui ont la structure *des N et des N de* :

Marie a gagné des centaines et des centaines d'amis dans cette affaire.

Notons que la contraction de la préposition *des* en *de* est également possible :

La galaxie est constituée de millions et de millions d'étoiles.

Les deux noms de la structure doivent être identiques et les déterminants sont obligatoirement *des* :

* *Léa a acheté des dizaines et des centaines de chiens*
* *Léa a acheté plusieurs dizaines et plusieurs dizaines de chiens*

Nous avons rassemblé toutes ces expressions dans le graphe **DetNnumDe**[32] :

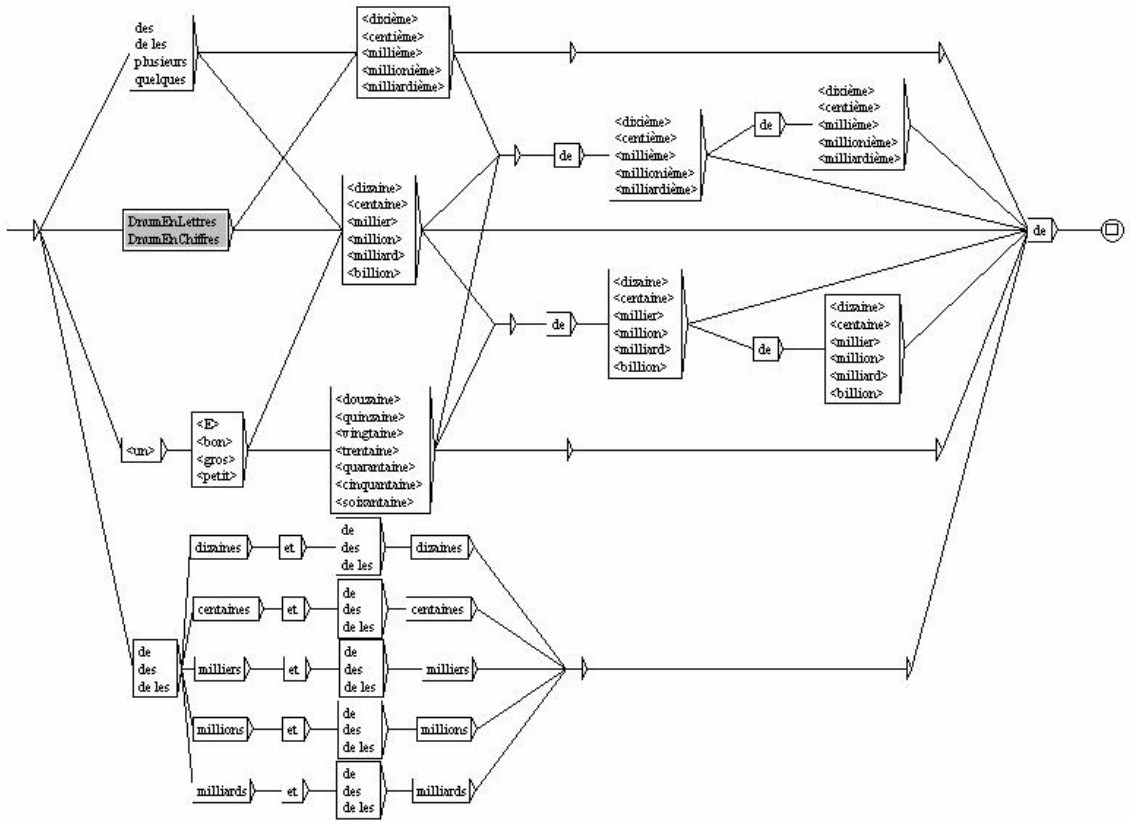

Figure 23 : DetNnumDe

[32] Ce graphe n'est pas cyclique contrairement à ce que l'on aurait pu penser. En effet, nous décidons de limiter le nombre de répétitions de la séquence *Nnum de* car les séquences trop longues sont difficilement compréhensibles par les lecteurs. Par ailleurs, certaines informations comme les fractions ne sont pas représentées.

Remarque générale sur les nombres :

Nous regroupons tous les graphes représentant des déterminants numériques dans le graphe **Dnum**.

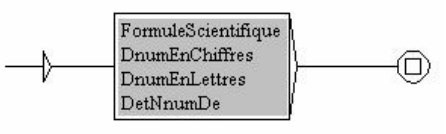

Figure 24 : Dnum

Il existe des combinaisons de ces déterminants numériques formant, par exemple, des approximations sous la forme d'intervalles (*entre 2 et 3 mètres* ; *de 7 à 9 kg*). Cependant, comme nous le verrons ultérieurement leur comportement ne peut être étudié de manière locale, mais dans le cadre d'une phrase élémentaire.

3.2.2.4 Les prédéterminants numériques

Dans les textes, on constate la présence de prédéterminants numériques (M. Gross, 1977) qui se trouvent avant ou après la séquence *Dnum N* (dans notre cas, *Dnum Unité*) comme *presque, environ, exactement, à peu près*:

Il y a 45 enfants environ
Max a mangé à peu près 30 fruits
Marie a (presque + environ + exactement) 10 ans
Marie a 10 ans (environ + très exactement)

Ces mots modifient l'interprétation de la valeur du déterminant numérique. La distribution des prédéterminants situés avant la séquence *Dnum N* (*PreDnum*) et des prédéterminants situés après la séquence *Dnum N* (*PreDnumPost*) n'est pas la même :

*Luc possède (à peu près + environ + *ou presque + presque) 30 voiliers*
*Luc possède 30 voiliers (?à peu près + environ + ou presque+*presque)*

Les prédéterminants *PreDnumPost* ne peuvent apparaître entre le déterminant et le nom :

** Il y a 45 (environ + très exactement) enfants*
** Marie a 10 (environ + très exactement) ans*

Cette contrainte n'est cependant pas toujours vrai comme le montre la phrase suivante (pas très naturelle) :

? Il y a quelques dizaines environ de voitures à boites automatiques[33]

Il est possible d'avoir à la fois un prédéterminant *PreDnum* et un prédéterminant *PreDnumPost* comme dans la phrase :

Paul a couché avec environ mille femmes au total

[33] Le symbole *?* signifie que la phrase qu'il précède n'est pas très naturelle.

Il est très facile de représenter ces deux ensembles sous la forme de graphes comme montré ci-dessous[34] (graphes **PreDnum** et **PreDnumPost**), puis de les incorporer dans notre structure de base.

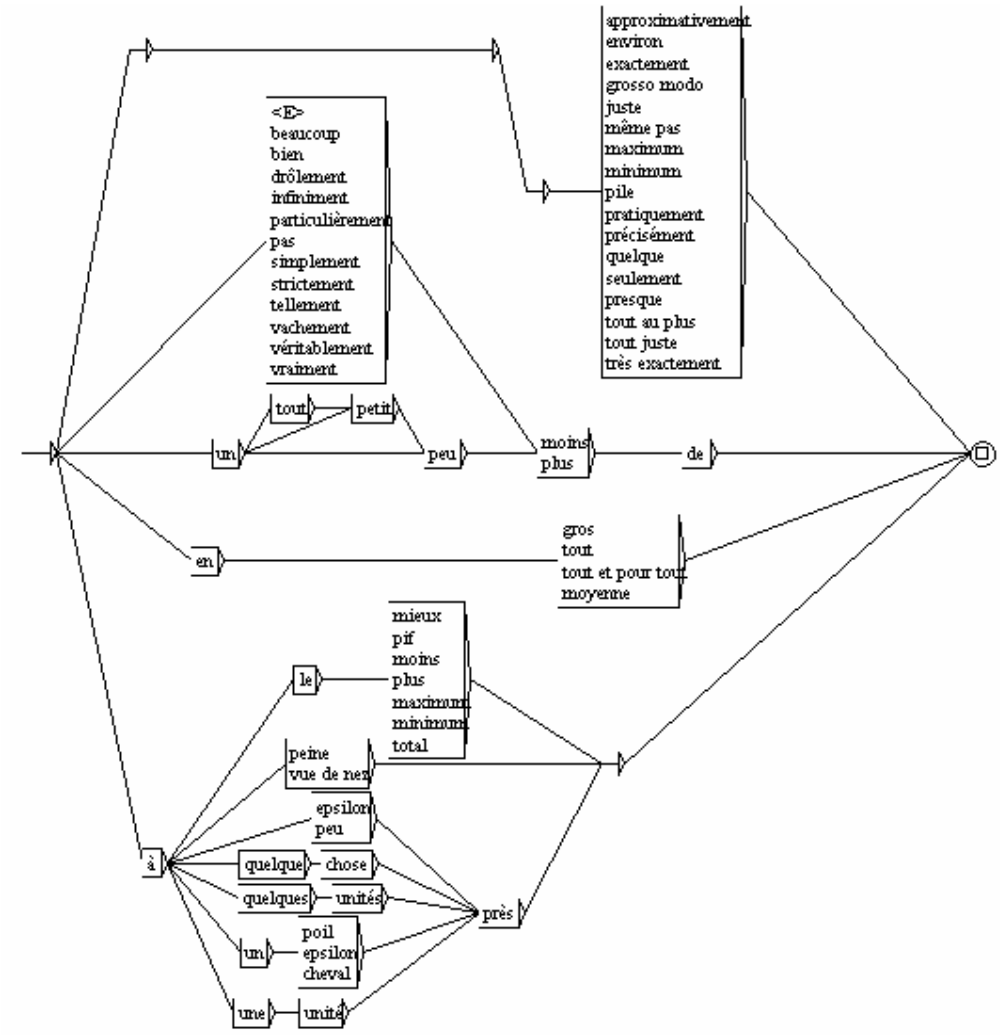

Figure 25 : PreDnum

[34] La base de ces graphes nous a été fournie par M. Gross.

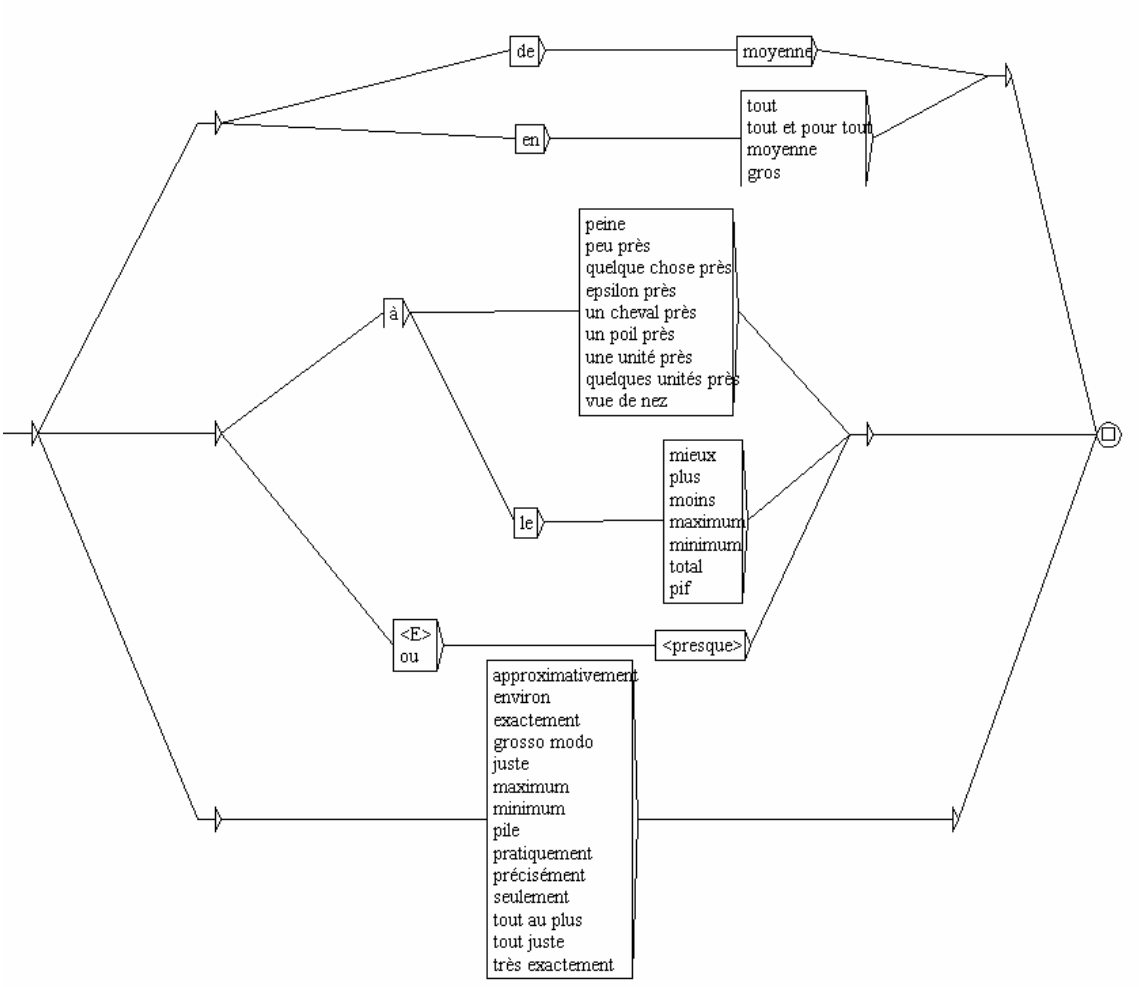

Figure 26 : PreDnumPost

Cependant, le comportement syntaxique des prédéterminants est plus complexe que cela et ils ne peuvent être uniquement étudiés de manière locale. Il faut les considérer dans des phrases simples comme dans M. Gross (1977). Par exemple, ils peuvent jouer le rôle d'adverbes car certains peuvent s'insérer n'importe où dans les phrases :

Max a (au total + à un poil près + ?environ) dépensé 400 euros
*(Au total + A un poil près + *Environ), Max a dépensé 400 euros*

Nous reviendrons sur ce phénomène ultérieurement lorsque nous examinerons nos différents schémas de phrase représentant des mesures.

3.2.3 Graphes élémentaires de mesure

3.2.3.1 Les unités

Dans cette section, nous répertorions les unités de mesure. Nous utilisons la classification scientifique : étant donné une unité de base (*mètre* ; *m*), nous regroupons, dans une même classe (ou graphe), ses multiples et sous-multiples (ex : *kilomètre* ; *km* ; *millimètre* ; *mm*). Nous divisons chaque classe en deux : les unités écrites en toutes lettres (*millimètre*,

centimètre) et les symboles des unités (*mm*, *cm*). Nous donnons ci-dessous les graphes **Metre** et **Metre_abr**. Chaque unité écrite en toutes lettres est mise entre angles : par exemple, *<millimètre>* est l'ensemble {*millimètre*, *millimètres*}. Cela signifie que l'on considère que les unités ont été décrites dans les dictionnaires que nous allons utiliser pour appliquer nos grammaires. Les symboles sont, quant à eux, écrits entre guillemets car ils doivent être reconnus tels quels dans les textes : la séquence *"cm"* interdit les variantes en majuscules, c'est-à-dire *CM*, *cM*, *Cm*.

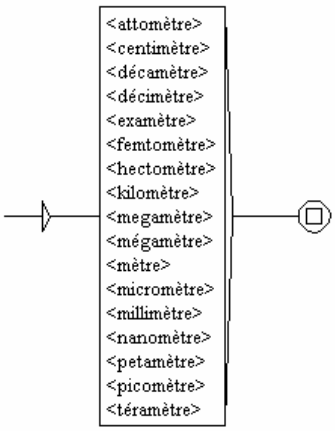

Figure 27 : Metre

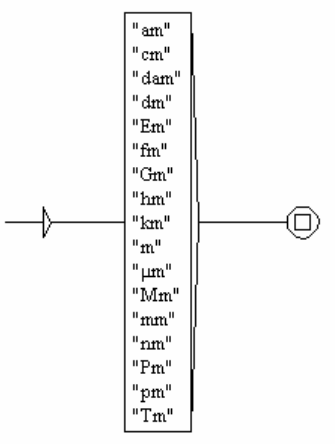

Figure 28 : Metre_abr

Nous répertorions ci-dessous les 20 graphes que nous avons construits à l'aide du dictionnaire *Larousse* :

> Ampere, Are, Bit, Calorie, DegreCelsius, ElectronVolt, Gramme, Hertz, Joule, Kelvin, Litre, Livre, Metre, Mile, Mille, Newton, Octet, Seconde, Volt, Tonne

A chacun de ces graphes, nous associons le graphe des symboles des unités correspondants dont le nom se termine par *_abr*. La plupart des graphes ne présentent aucune difficulté et leur contenu est facilement construit manuellement (voire automatiquement) sur le même modèle que nos deux exemples.
Le graphe **DegreCelsius** décrit les différents types de degrés pour mesurer une température : *degré Celsius*, *degré Fahrenheit*, *degré Kelvin*. Les symboles décrits dans **DegreCelsius_abr** sont *°C*, *°F* et *°K*. Le graphe **Mille** correspond aux milles nautiques. Il n'existe pas de graphe *_abr* associé. Nous avons considéré que les unités de temps n'étaient pas suffisamment bien représentées par le graphe **Seconde** (*seconde*, *milliseconde*, …). Nous avons donc construit un graphe **Ndiv-temps** répertoriant les noms désignant des divisions du temps : *an*, *année*, *trimestre*, *mois*, *jour*, *heure*, *minute*, etc.

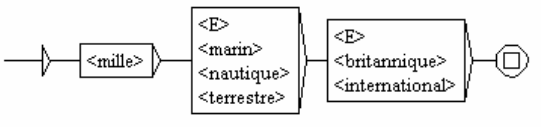

Figure 29 : Mille

Au vu de cette liste, il est clair que nous n'avons pas répertorié toutes les unités simples existantes. Mais, nous considérons que cela est suffisant pour notre étude.

Nous décidons quand même d'ajouter un graphe (**Nmonnaie**) regroupant toutes les unités de monnaie plus le graphe des symboles (**Nmonnaie_abr**). Certains noms de monnaies peuvent être regroupés dans des sous-graphes (**Dinar** pour les différents types de dinars, **Dollar** pour les différents types de dollars, **Franc**, **LivreSterling**, etc.). Nous donnons le graphe **Dollar** dans la figure ci-dessous :

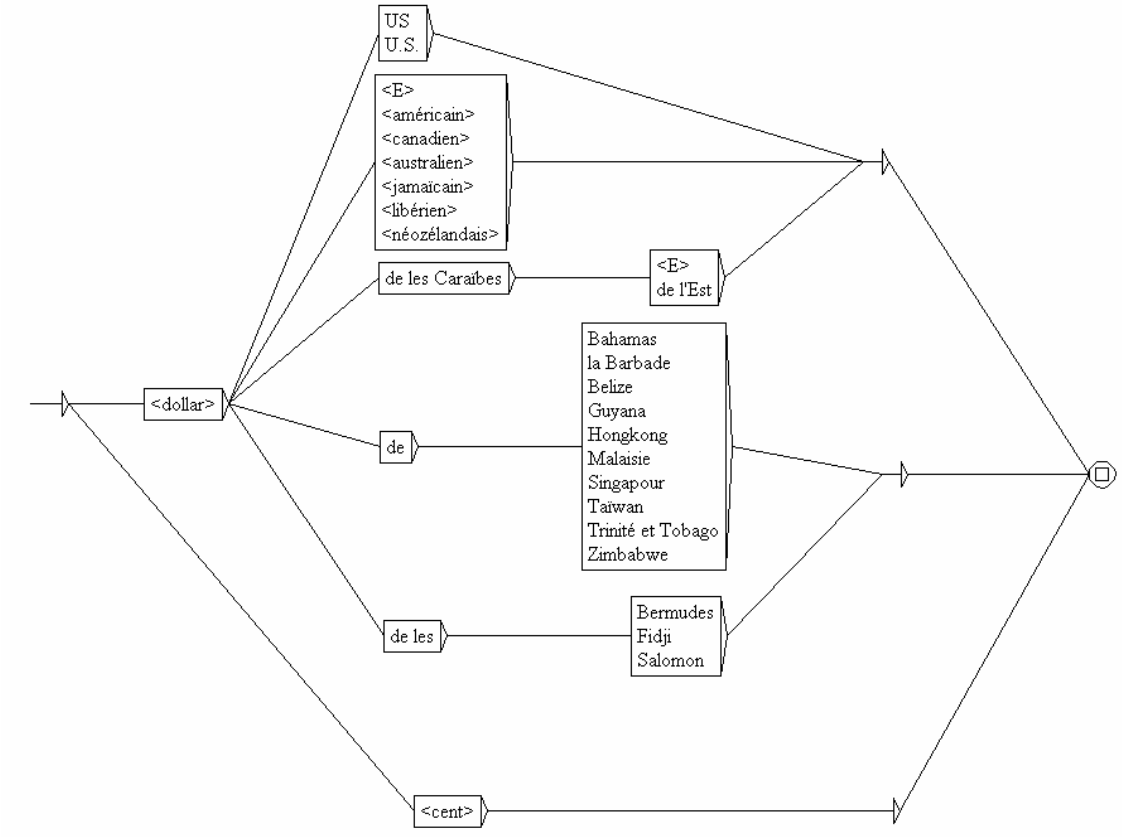

Figure 30 : Dollar

Nous avons également construit le même genre de graphes pour les unités en anglais pour lesquelles nous avons utilisé le dictionnaire en-ligne se trouvant à l'URL www.unc.edu/~rowlett/units/. Nous donnons ci-dessous le graphe décrivant la classe des unités dont l'unité de base est *gram* (gramme). Les symboles de monnaies ont un comportement différent des autres unités car ils sont toujours situés avant *Dnum* (ex : *£10*).

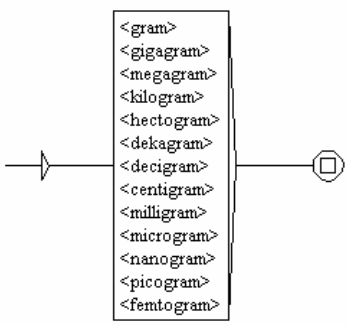

Figure 31 : Gram

3.2.3.2 Les contraintes internes à *Dnum Unité*

La séquence *Dnum Unité* comprenant deux composants indépendants est une facilité théorique d'écriture que l'on s'est donné. Dans les faits, bien que cette indépendance se vérifie souvent, elle n'est pas toujours vraie. En effet, plusieurs points remettent en cause cette représentation. Tout d'abord, il semble exister des contraintes stylistiques entre *Dnum* et *Unité*. Par exemple, la combinaison (*DnumEnLettres + DetNnumDe) Unité_Abr* n'est pas naturelle alors que les autres sont tout à fait acceptables : **(dix + quelques dizaines de) m ; 10 (m + mètres) ; (dix + quelques dizaines de) mètres*. Nous illustrons cette règle sous la forme d'un graphe représentant la séquence formée d'un déterminant numérique et d'une unité métrique.

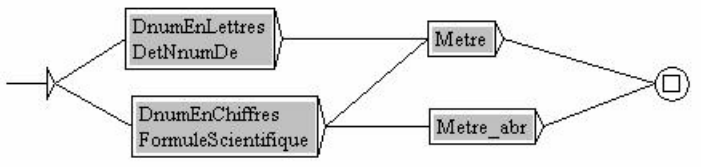

Figure 32 : Règle stylistique

Ensuite, les déterminants numériques ne sont pas toujours connexes et peuvent se diviser en deux parties entre lesquelles vient se greffer l'unité, comme le montrent les exemples ci-dessous :

Max a un retard de huit minutes (trente + et demi) par rapport à son emploi du temps
Marie a sauté 5 m 60

Cette contrainte est représentée par le graphe ci-dessous [35] . On utilise le graphe **NombreEntierEnChiffres** car tout nombre décimal est interdit (** 2,5 m 12*). Notons que l'espace blanc entre l'unité et le nombre entier en chiffres qui la suit est respecté pour éviter la confusion avec les *mètres carrés* (*m2*) ou *mètres cubes* (*m3*) :

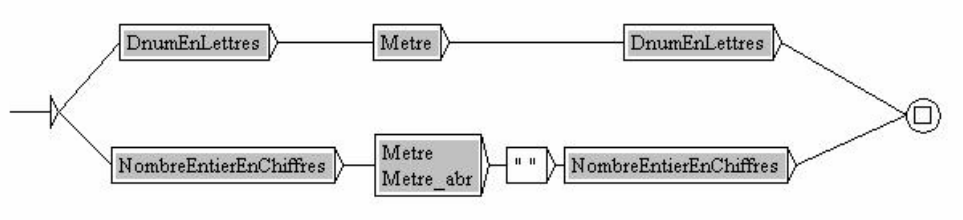

Figure 33 : Non-connexité des déterminants

Enfin, une séquence *Dnum Unité* peut parfois commuter avec une suite de plusieurs *Dnum Unité*. Cette suite a une syntaxe qui lui est propre car elle dépend de la classe d'unités utilisée. Cette forme sert essentiellement à ajouter une précision à la mesure :

Léa a eu droit à trois heures, dix minutes et douze secondes de sueurs froides
*Luc a exactement parcouru 10 kilomètres et (30 mètres + *30 secondes)*

Nous représentons un exemple (très partiel) de telles séquences dans le graphe suivant :

[35] Le graphe présenté est partiel car il ne contient pas compte d'expressions telles que *trois mètres et demi*.

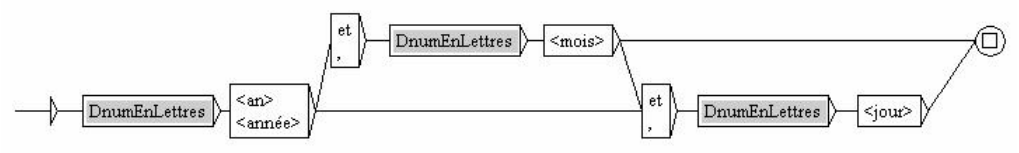

Figure 34 : exemple de syntaxe propre à une unité

Ainsi, pour chaque classe d'unités, il est nécessaire de regrouper la séquence *Dnum Unité* en un seul graphe **GNmesure-unite**, en tenant compte des remarques précédentes et de la présence potentielle de prédéterminants. Nous donnons ci-dessous un exemple d'un tel graphe pour la classe d'unités **Metre**. Le graphe **DnumMetre-precis** reconnaît des suites *Dnum Metre* selon une syntaxe spécifique (ex : *10 kilomètres et 300 mètres*)[36]. Ce dernier graphe a été conçu de telle manière qu'il ne reconnaisse pas des séquences comme *10 kilomètres et 3 000 mètres* (cf. graphe ci-dessous). Les graphes **DnumEnLettres1-99** et **DnumEnLettres1-999** représentent des nombres écrits en toutes lettres respectivement de *un* à *quatre-vingt dix-neuf* et de *un* à *neuf cent quatre-vingt dix-neuf*. Les graphes **NombreEntierEnChiffre1-99** et **NombreEntierEnChiffre1-999** décrivent des nombres entiers écrits en chiffres allant respectivement de *1* à *99* et de *1* à *999* (plus 0).

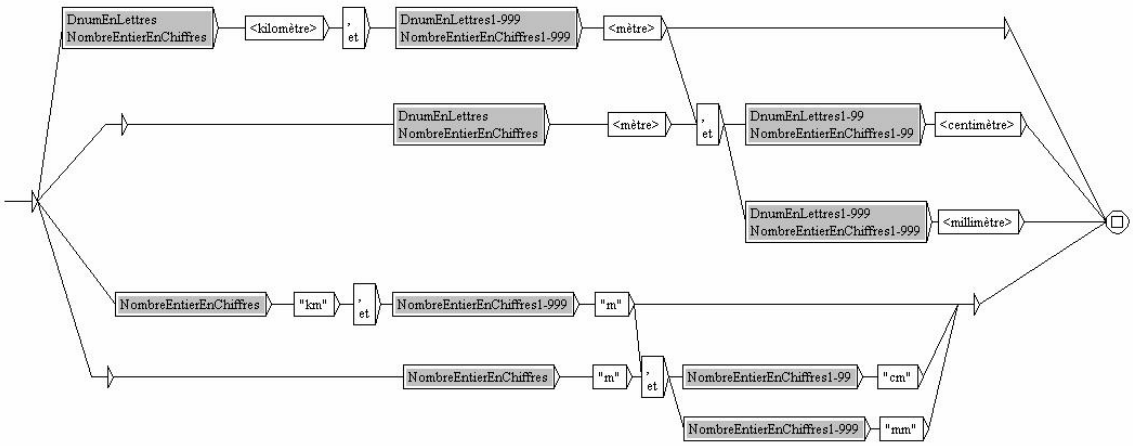

Figure 35 : DnumMetre-precis

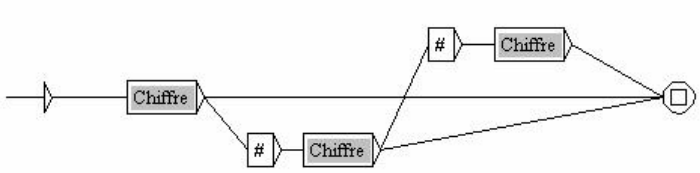

Figure 36 : NombreEnChiffres1-999

[36] D'une manière générale, ce type de graphe est nommé selon le modèle suivant : *DnumUnite-precis*.

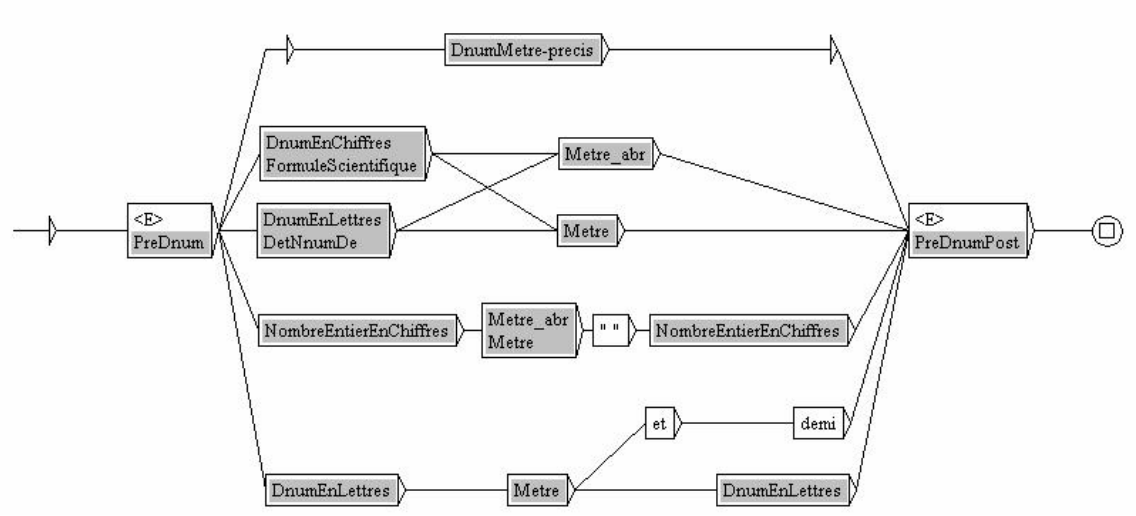

Figure 37 : GNmesure-metre

Dans la suite, nous emploierons le terme générique *GNmesure* pour représenter la séquence *Dnum Unité.*

3.2.3.3 Contraintes entre *Ng* et *Unité*

Dans le schéma de phrase *Det Ng être de Dnum Unité* (=: *la longueur est de 30 m*), chaque *Ng* sélectionne un ensemble restreint d'unités homogènes :

- soit des unités simples comme *largeur* qui sélectionne le *mètre*, ses multiples (*kilomètre*) et ses sous-multiples (*millimètre*), plus d'autres unités comme le *mille nautique* et le *mile*.

- soit des unités complexes (des combinaisons d'unités simples) comme *vitesse* qui sélectionne des combinaisons d'unités de mesure de longueur (*mètre, mile*) et de temps (*heure*) : *kilomètres à l'heure*.

Sur la base de l'étude de J. Giry-Schneider (1991), nous avons systématiquement examiné les noms *Ng* entrant dans nos deux structures de base et nous avons associé à chacun un ensemble de graphes représentant les unités sélectionnées par *Ng*. L'étude nous a conduit à décrire 17 classes d'unités qui sont explicitées dans le tableau ci-dessous. Chaque ligne correspond à une classe. La première colonne donne, pour chaque classe, le nom du graphe de type *GNmesure* qui sera automatiquement construit à partir du contenu de la classe. La deuxième colonne correspond à l'ensemble des noms des graphes[37] décrivant les unités écrites en toutes lettres d'une classe (graphes du type *Unite*). La troisième colonne correspond à l'ensemble des noms des graphes décrivant les symboles des unités d'une classe (graphes du type *Unite_abr*). La dernière colonne correspond à l'ensemble des noms de graphes du type *DnumUnite-precis* associés à une classe d'unités.

[37] Par convention, les noms des graphes sont toujours précédés du symbole '*:*'.

A	B	C	D	E
nom graphe	Unite	Unite_abr	DnumUnite-precis	Exemples
GNmesure-longueur	:Metre+:Mile+:Mille	:Metre_abr+:Mile_abr	:DnumMetre-precis+:DnumMile-precis	*15 mètres*
GNmesure-masse	:Gramme+:Livre+:Tonne	:Gramme_abr+:Livre_abr+:Tonne_abr	:DnumGramme-precis	*15 grammes*
GNmesure-temperature	:DegreCelsius+:Kelvin	:DegreCelsius_abr+:Kelvin_abr	-	*15 °C*
GNmesure-force	:Newton	:Newton_abr	-	*13 kN*
GNmesure-population	:Habitant	-	-	*mille habitants*
GNmesure-energie	:Joule+:Calorie+:ElectronVolt	:Joule_abr+:Calorie_abr+:ElectronVolt_abr	-	*124 kJ*
GNmesure-intensite-elec	:Ampere	:Ampere_abr	-	*0,2 A*
GNmesure-informatique	:Bit+:Octet	:Bit_abr+:Octet_abr	-	*56 ko*
GNmesure-temps	:Seconde+:Ndiv-temps	:Seconde_abr+:Ndiv-temps_abr	:DnumNmesure-temps-precis	*deux minutes*
GNmesure-tension	:Volt	:Volt_abr	-	*110 V*
GNmesure-monetaire	:Nmonnaie	:Nmonnaie_abr	:DnumMonnaie-precis	*cinq dollars*
GNmesure-vitesse	:Nmesure-vitesse	-	-	*120 km/h*
GNmesure-surface	:Nmesure-surface	:Nmesure-surface_abr	:DnumNmesure-surface-precis	*10 hectares*
GNmesure-volume	:Nmesure-volume	:Nmesure-volume_abr	-	*43 m3*
GNmesure-densite-pop	:Nmesure-densite-pop	-	-	*10 habitants au km2*
GNmesure-frequence	:Nmesure-frequence	:Hertz_abr	-	*50 Hz*
GNmesure-angle	:Nmesure-angle	:Nmesure-angle_abr	-	*dix radians*

Table 2 : classes d'unités

La plupart du temps, les graphes sélectionnés correspondent à des unités simples décrites dans la section précédente. D'autres représentent des combinaisons d'unités comme pour le nom *vitesse* qui sélectionne des unités complesxes combinant des unités métriques et de temps. En physique, une unité de vitesse est la « division » d'une unité métrique par une unité de temps : *centimètres par heure*, *km/s*. Dans le langage courant, il existe des variations moins « rigoureuses » telles que *kilomètres à l'heure* ou *kilomètres-heure*. Le premier exemple peut même être réduit à *à l'heure* comme dans l'exemple suivant :

Max roule à une vitesse de 80 kilomètres à l'heure
Max roule à une vitesse de 80 à l'heure

Mais cela n'est valable qu'avec le nom *heure* :

Ce météorite a une vitesse de 1,2 km à la seconde
**Ce météorite a une vitesse de 1,2 à la seconde*

Nous donnons ci-dessous le graphe **Nmesure-vitesse** représentant ce type d'unités. Dans ce graphe, nous autorisons des expressions plus exotiques telles que *miles par an*. Nous ne divisons pas cette unité en deux comme auparavant (unité en toutes lettres ; symboles). Le graphe **Nmesure-longueur** est l'union des graphes **Metre**, **Metre_abr**, **Mile**, **Mile_abr** et **Mille**. Le graphe **Nmesure-temps** est l'union des graphes **Ndiv-temps**, **Ndiv-temps_abr**, **Seconde** et **Secondes**.

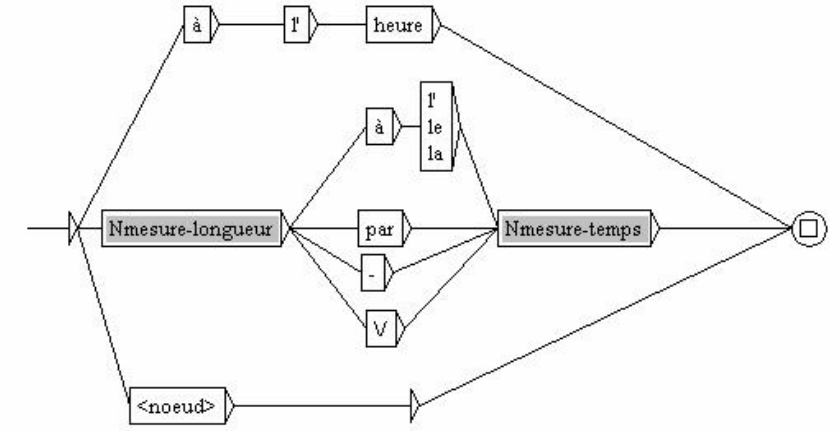

Figure 38 : Nmesure-vitesse

Les noms désignant une surface (*aire*, *surface*, *superficie*) sélectionnent deux types d'unités :

- la combinaison d'une unité métrique de longueur accompagnée soit d'un 2 soudé si l'on a un symbole, soit du modifieur *carré* si l'unité le précédant est écrite en toutes lettres (ex : *m2*, *mètre carré*) ;

- des unités de surface simples comme *are*, *hectare* symbolisées par *a* et *ha*.

Dans ce cas, on sépare les symboles des unités écrites en toutes lettres et on construit les deux graphes **Nmesure-surface** et **Nmesure-surface _abr** suivants :

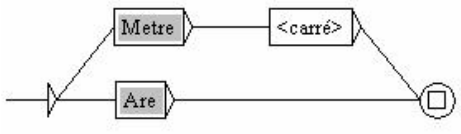

Figure 39 : Nmesure-surface

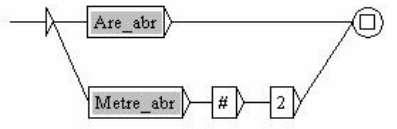

Figure 40 : Nmesure-surface_abr

A cette classe, on associe également le graphe **DnumNmesure-surface-precis**. En effet, on peut exprimer une mesure de surface à l'aide de la multiplication de deux mesures de longueurs (**GNmesure-longueur**) comme suit :

Marie a acheté un champ d'une surface de (70 m x 120 m + cent mètres sur trente).

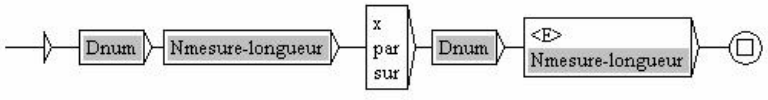

Figure 41 : DnumNmesure-surface-precis

Il en est de même pour les unités sélectionnées par *volume*. Le graphe formé est l'union de deux types d'unités :

- des séquences comprenant une unité métrique de longueur suivie d'un 3 collé ou du modifieur *cube* ;

- Les unités dérivées de *litre* se trouvant dans les graphes **Litre** et **Litre_abr**.

Nous synthétisons ces unités dans les graphes **Nmesure-volume** et **Nmesure-volume_abr**.

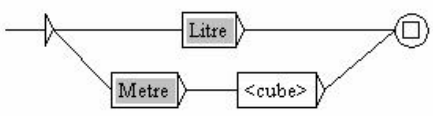

Figure 42 : Nmesure-volume

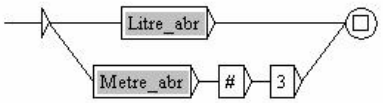

Figure 43 : Nmesure-volume_abr

Ci-dessous nous donnons le graphe **Nmesure-densite-pop** décrivant les unités sélectionnées par le nom *densité (démographique + de population)* et reconnaissant des expressions telles que *habitants au km2*.

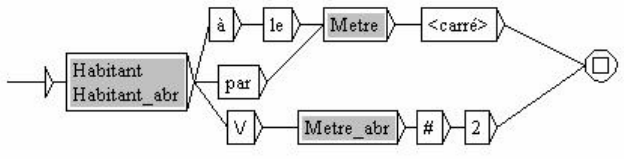

Figure 44 : Nmesure-densite-pop

L'unité scientifique traditionnellement associée à la fréquence est le *hertz*. Ainsi, *fréquence* sélectionne les unités décrites par le graphe **Hertz**. Cependant, dans le langage courant, c'est beaucoup plus libre et cela peut être n'importe quel groupe nominal comptable (sans déterminant) suivi par une préposition (ou le symbole '/'), un déterminant optionnel et une unité de temps :

Le moteur a une fréquence de trois tours par seconde
La machine a une fréquence de trente poulets (à la + /) minute

Ainsi, nous avons besoin d'une description complète d'un groupe nominal. Cependant, pour des raisons évidentes de clarté ce groupe nominal ne peut être trop long car il doit être suivi d'une séquence exprimant le temps. Ainsi, nous limitons notre groupe nominal sans déterminant à la séquence maximale suivante *Adj N Adj de Det N*. Les unités associées à *fréquence* sont représentées dans le graphe **Nmesure-frequence** où les symboles <*A*>, <*N*> et <*DET*> désignent respectivement un adjectif, un nom et un déterminant.

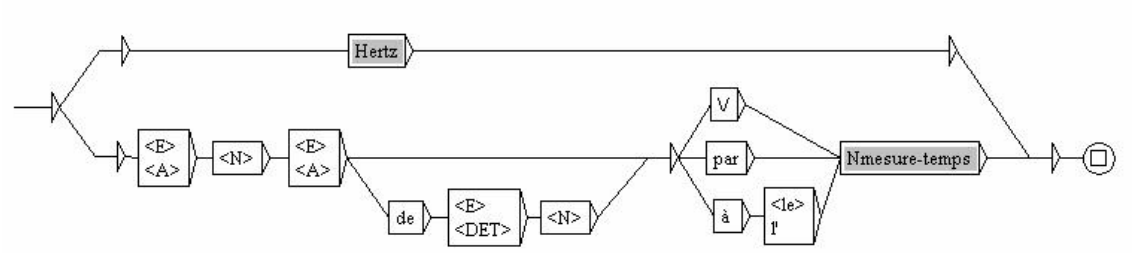

Figure 45 : Nmesure-frequence

Le graphe **Nmesure-angle** contient les unités *radian, degré* et leurs sous-multiples comme *minute*. Les symboles sont dans **Nmesure-angle_abr** : °, *rad*, etc.

Certains *Ng* sont ambigus comme *tension* qui désigne soit une *tension électrique* soit une *tension artérielle*. Cette ambiguïté est levée si l'on regarde les unités sélectionnées par

chacun : *Volt* pour *tension électrique* et *<E>* pour *tension artérielle*. Par ailleurs, *longueur* est aussi ambigu : il peut désigner

- soit une durée comme dans :

  Ce spectacle a une longueur de 2 heures

- soit une mesure métrique comme dans :

  Cette corde a une longueur de 25 mètres

Il en est de même pour *poids* qui est soit une force (nom standard en physique), soit une masse comme il est courant de l'utiliser dans la vie de tous les jours. La vie courante ne permet pas de distinguer les notions de force et de masse.

Les noms sélectionnant une unité monétaire ont un comportement particulier par rapport aux autres en ce sens qu'ils autorisent l'effacement de l'unité :

Ce cadeau a une valeur de 2,50 (euros + ?E)

3.2.3.4 Processus de génération des graphes *GNmesure*

La table décrivant les classes d'unités peut être vue comme une table syntaxique (cf. chapitre sur le lexique-grammaire). Chaque ligne correspond à une entrée lexicale (un nom de classe) qui est donnée dans la première colonne (ex : *GNmesure-longueur* pour la première ligne). Les colonnes (autres que la première) contiennent certaines propriétés de la classe : colonne B, l'ensemble des graphes d'unités en toutes lettres de la classe ; colonne C, l'ensemble des graphes des symboles (s'ils existent) ; colonne D, l'ensemble des graphes du type *DnumUnite-precis* (s'ils existent). Chaque élément de la table est soit un élément lexical (noms des graphes) soit un booléen (+ pour vrai ; - pour faux)[38].
Pour chaque entrée lexicale de la table, le but est de construire un graphe qui décrit tous les groupes nominaux de mesure de type *GNmesure* dans lesquelles rentrent cette entrée. Nous utilisons la méthode d'E. Roche (1993, 1994) qui consiste à utiliser un graphe patron qui représente l'ensemble des structures potentielles des entrées de la table. Un graphe patron est associé à une classe d'éléments lexicaux ; il est paramétré de façon à pouvoir être adapté à chaque élément lexical en fixant la valeur des paramètres. Chaque élément d'information des tables (c'est à dire une propriété ou plus simplement une colonne de la table) est représenté par une variable dans le graphe patron. Pour chaque ligne *i* de la table à convertir *T*, on réalise une copie du graphe patron. Puis, pour chaque transition (ou boîte) de ce graphe, nous effectuons les opérations suivantes pour chacune des variables *@j* contenues dans cette transition :

- si $T(i,j) = +$, on remplace *@j* par le mot vide <E> ;
- si $T(i,j) = -$, on supprime la transition, coupant ainsi le chemin reconnaissant la structure correspondant à la colonne *j* ;
- sinon, on remplace *@j* par l'information lexicale contenue dans $T(i,j)$.

Par cette méthode, nous générons les graphes de type *GNmesure* associés à notre table. Nous utilisons le graphe patron ci-dessous. La variable *@B* correspond aux informations contenues

[38] Cette table ne contient pas de + ; les – signifient qu'il n'existe pas de graphes du type défini par la colonne.

dans la colonne B, la variable *@C* correspond aux informations contenues dans la colonne C, etc. Pour chaque entrée le graphe généré a pour nom le contenu de la colonne A (*@A*).

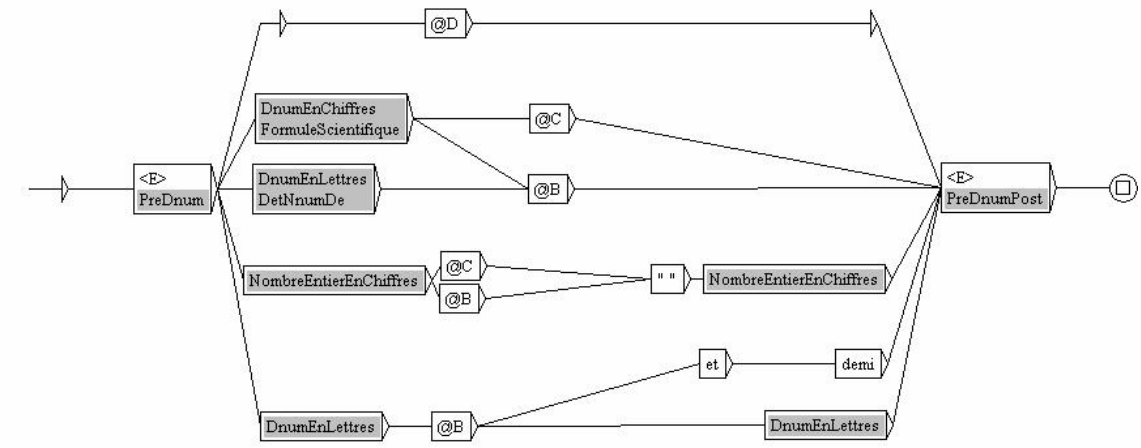

Figure 46 : Graphe patron pour générer les graphes du type *GNmesure*

Pour l'entrée *GNmesure-vitesse*, nous obtenons le graphe **GNmesure-vitesse** ci-dessous. La variable *@B* est remplacée par l'information lexicale : *Nmesure-vitesse* (nom du graphe précédé de *:*). Les boîtes contenant les variables *@C* et *@D* sont supprimées, éliminant ainsi les chemins reconnaissant des structures interdites.

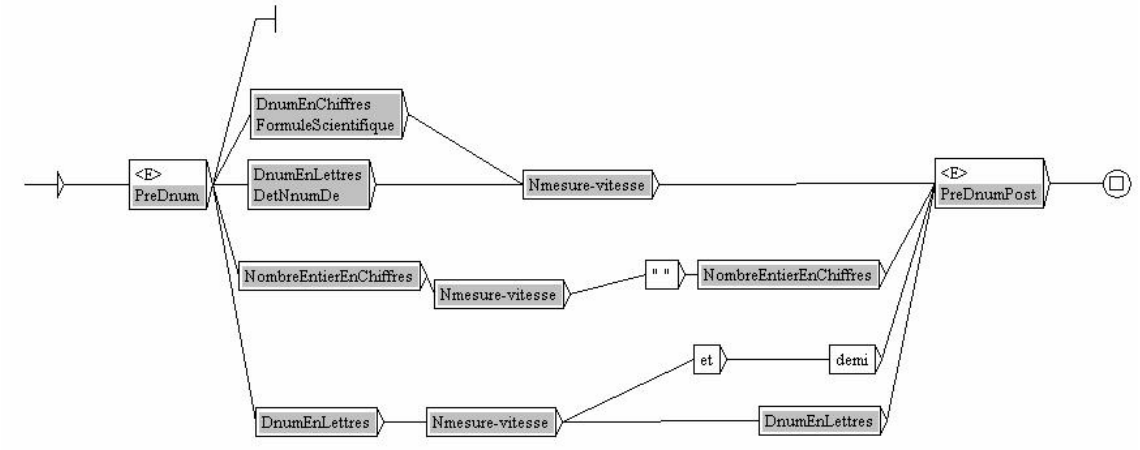

Figure 47 : GNmesure-vitesse

3.2.4 Quelques variantes

3.2.4.1 Quelques variantes simples

On constate que nos phrases peuvent être étendues à d'autres schémas de phrase :

(a) *N*0 *être Adj à N*1 *= : la longueur est supérieure à 10 mètres*
(b) *N*0 *être Prép N*1 *= : le poids est de l'ordre de 10 kilos*
(c) *N*0 *être Vpp à N*1 *= : la tension est limitée à trente volts*
(d) *N*0 *V Prép N*1 *= : la fréquence atteint 55 Hz*

La structure (a) contient des adjectifs appropriés aux mesures tels que *égal* et *supérieur*. Ils sont décrits dans le graphe **Adj-numA**.

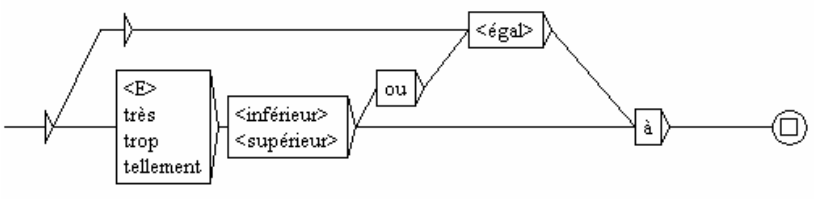

Figure 48 : Adj-numA

La structure (b) comprend des prépositions composées (pour la plupart) qui peuvent aussi être vues comme des prédéterminants (ex : *jusqu'à* , cf. M. Gross, 1977). Ces prépositions sont décrites dans le graphe **PreDnumPrep**.

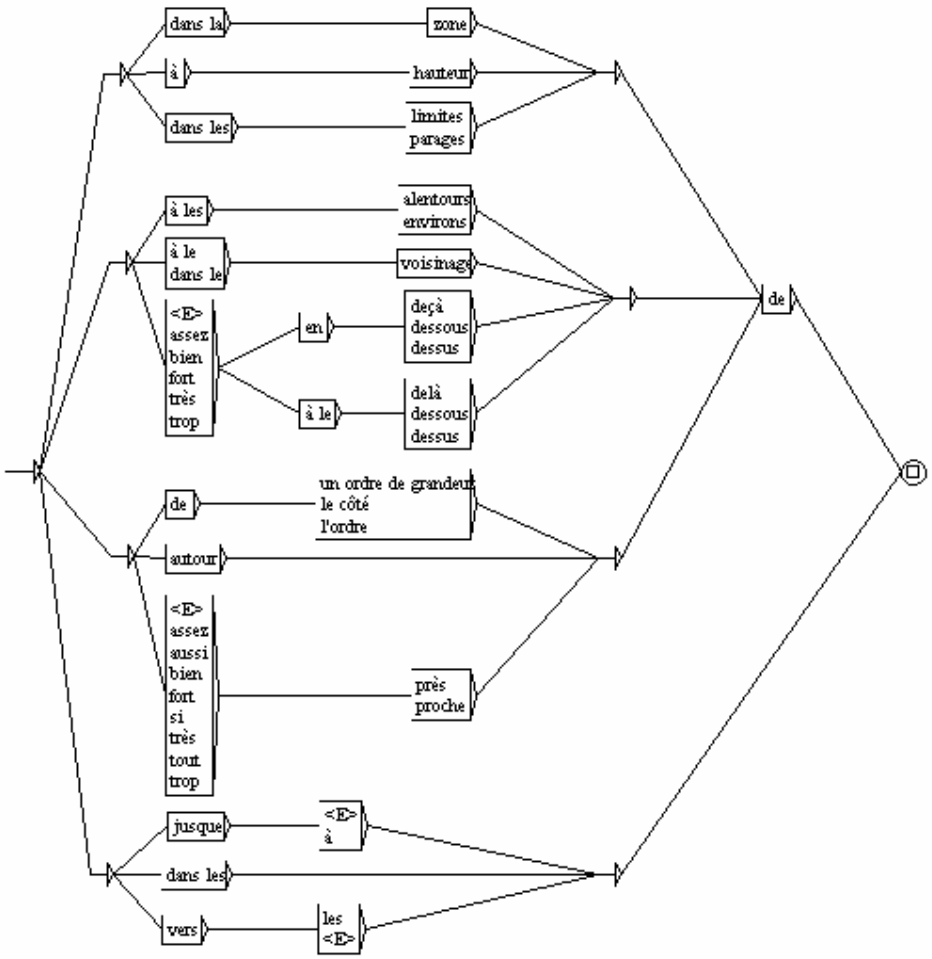

Figure 49 : PreDnumPrep

La structure (c) comporte des verbes au participe passé *Vpp* : *la tension est limitée à trente volts*. Cette dernière phrase est la forme passive de

On limite la tension à trente volts

Ces passifs ont un sens statif, même avec agent. Nous décrivons quelques *Vpp* de ce type dans le graphe **Vpp-numA**.

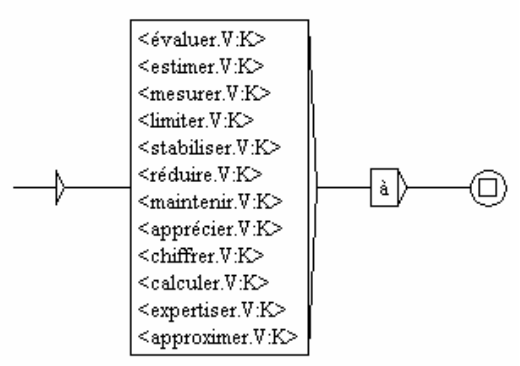

Figure 50 : Vpp-numA

Enfin, il existe quelques verbes appropriés aux mesures. Ils ont tous un aspect statique avec quelques nuances sémantiques : *atteindre, s'élever à, se monter à*.

3.2.4.2 Quelques variantes complexes

Dans cette section, nous nous intéressons à des combinaisons plus complexes que dans les sections précédentes. Nous regardons d'abord le comportement de nos phrases de mesure lorsqu'on leur applique des coordinations. Nous examinons aussi en détail les structures *de … à …* et *entre … et …* qui, dans notre cas, désignent des approximations de mesures sous la forme d'intervalles. Nous montrons qu'elles ne peuvent être décrites localement de manière complète : le contexte minimal est une phrase.

3.2.4.2.1 Les coordinations

Prenons les deux phrases simples de mesure suivantes :

Max a une taille de 1,70 m
Luc a une taille de 1,80 m

Ces phrases peuvent être coordonnées à l'aide de la conjonction *et* :

Max a une taille de 1,70 m et Luc a une taille de 1,80 m

Nous réduisons cette phrase complexe à une forme factorisée, en plusieurs étapes, à l'aide de l'adverbe *respectivement* :

- regroupement des sujets :

  Max et Luc ont respectivement une taille de 1,70 m et une taille de 1,80 m

- factorisation de *taille* (*N*) et de la préposition *de* :

  Max et Luc ont respectivement une taille de 1,70 m et (de + E) 1,80 m

- factorisation de l'unité lorsqu'elle est similaire

  Max et Luc ont respectivement une taille de 1,70 et 1,80 m

Cette dernière étape n'est pas réalisable lorsque les unités ne sont pas les mêmes :

La première et la deuxième ont respectivement une tension de 30mV et 3V

Cette analyse est aussi valable pour les noms entrant dans la deuxième structure étudiée. Nous avons les deux cas suivants

Pierre est à une distance de 30 km de Paris et Pierre est à une distance de 50 km de ta ville
= Pierre est à une distance de 30 et 50 km respectivement de Paris et de ta ville

Pierre est à une distance de 200 km de Paris et Paul est à une distance de 45 km de Paris
= Pierre et Paul sont respectivement à une distance de 200 et 45 km de Paris

La conjonction *ou* marque une approximation de la valeur numérique du déterminant (sous la forme d'un choix entre plusieurs valeurs) :

Max a cinq ou six ans.

Nous pouvons interpréter *cinq ou six* comme un déterminant composé, mais cela peut poser des problèmes. En effet, la phrase ci-dessus est équivalente aux deux phrases suivantes :

Max a cinq ans ou six ans.
Max a cinq ans ou Max a six ans.

On peut aussi avoir des phrases du type :

Le tuyau est à une distance de 90 cm ou 1 m du mur.
Le tuyau est à une distance de 90 cm du mur ou le tuyau est à une distance de 1 m du mur.

Pour plus de détails sur la coordination dans les groupes nominaux, nous suggérons au lecteur de se référer à C. Domingues (2001).

3.2.4.2.2 La structure entre … et …

Revenons à notre phrase de base *Det Ng être de Dnum Unité* :

Cette longueur est de 15 m.

Il est possible d'exprimer une approximation en utilisant un intervalle de mesures grâce à la structure *être compris entre … et ….* On remarque que *compris* peut être effacé.

Ce Ng être (E + compris) entre $Dnum_1$ $Unité_1$ et $Dnum_2$ $Unité_2$
= : cette longueur est (E + comprise) entre 90 cm et 1,10m

Comme pour les coordinations, $Unité_1$ et $Unité_2$ peuvent être factorisées si $Unité_1 = Unité_2$:

Det Ng être (compris + E) entre $Dnum_1$ (E + Unité) et $Dnum_2$ Unité
= : la longueur est (comprise + E) entre 90 (E + m) et 110 m

La réduction à un groupe nominal par relativation puis réduction de la relative donne des séquences telles que :

Une longueur (E + comprise) entre 90 (E + m) et 110 m

Lorsque l'on utilise une variante de *être de*, comme *être supérieur à* (cf. section précédente), on observe un comportement un peu différent à cause de la présence de la préposition *à* : la préposition *à* est obligatoirement effacée. Les phrases obtenues ne sont pas très naturelles.

La température est supérieure à dix degrés Celsius.
? La température est supérieure (E + à) entre dix et quinze degrés Celsius*

Par ailleurs, la variante en *être PreDnumPrep* est peu naturelle :

La tension est (de l'ordre de + dans les + à hauteur de) 15 V
*?*La tension est (de l'ordre d' + dans les + à hauteur d') entre 14 et 15 V*

Pour les variantes avec des verbes tels que *atteindre*, on a :

*La diamètre atteint (*comprise + E) entre 90 (E + m) et 110 m*

Les structures *entre Dnum et Dnum Unité* sont beaucoup plus fréquentes que *entre Dnum Unité et Dnum Unité*. Ne tenir compte que des expressions trouvées dans les corpus aussi grands soient-ils n'est pas suffisant. La première intuition à partir des expressions du corpus est de considérer *entre Dnum et Dnum* comme l'équivalent d'un déterminant numérique comme montré dans l'analyse ci-dessous :

(Le diamètre) (atteint) ((entre 90 et 110) m)

Nos graphes permettent d'analyser des formes plus rares telles que *entre 90 cm et 1,10m*. L'exemple ci-dessus est analysé comme suit :

(Le diamètre) (atteint) (entre (90) et (110 m))

3.2.4.2.3 La structure de … à …

Il existe une autre structure permettant d'exprimer une approximation de mesure sous la forme d'un intervalle : c'est la séquence *de … à …*. Nous modifions notre phrase de départ en la remplaçant par :

Det Ng être de $Dnum_1$ $Unité_1$ à $Dnum_2$ $Unité_2$
=: ? cette température est de 10 à 15 degrés

Notre exemple ci-dessus est réductible au groupe nominal :

cette température (qui est + E) de 10 degrés à 15 degrés

Comme pour *entre… et …*, lorsque l'on utilise les variantes *être Adj à*, la préposition *à* est interdite. Néanmoins, l'exemple ci-dessous montre que l'utilisation de telles phrases n'est pas très naturelle :

*La longueur est supérieure (?E + *à) de 10 m à 15 m*

Avec les *PreDnumPrep*, on observe l'effacement obligatoire de la préposition *de* :

*La longueur est (à hauteur de + de l'ordre de + ?vers les) (E + *de) 10 à 15 m*

Avec les verbes sans préposition, la préposition *de* n'est pas obligatoire :

La température atteint (de + E) 10 degrés à 15 degrés

La structure d*e … à …* est naturellement ambiguë. Son interprétation dépend du verbe de la phrase élémentaire dans laquelle elle se trouve. En effet, elle peut exprimer une évolution et non une approximation sous la forme d'un intervalle comme avec le verbe *passer* :

Le prix du pain est passé de 65 à 70 centimes

Dans cette phrase, le prix initial du pain est de 65 centimes ; à l'état final, il est de 70 centimes. Dans beaucoup de cas, cette ambiguïté est localement impossible à lever comme avec le verbe *augmenter* :

La tension entre ces deux points de la ligne a augmenté de 10 (E + V) à 15 V

Il existe deux analyses :

La tension entre ces deux points de la ligne a augmenté d'une valeur de 10 (E + V) à une valeur de 15 V
La tension entre ces deux points de la ligne a augmenté d'une valeur de 10(E + V) à 15 V

3.2.4.2.4 Le tiret

Il existe par ailleurs d'autres structures combinant des nombres exacts et désignant une approximation de valeur sous la forme d'un intervalle. La plus simple est l'emploi du tiret entre deux nombres :

L'intensité du courant sur cette ligne est de 150-200 ampères

Ces phrases paraissent plutôt orales qu'écrites. Il peut exister des problèmes d'interprétation car le nom-unité *ampères* est effacé entre 150 et le tiret :

L'intensité du courant sur cette ligne est de 150 ampères-200 ampères

Cela est confirmé par la phrase :

Ce chemin fait 800 mètres-1 kilomètre.

3.2.4.2.5 *Remarques*

- Nouvelle notation

Dorénavant, dans les graphes, pour décrire la séquence *Dnum Unité*, nous employons les termes *GNmesure* et *GNmesureFinal* qui sont aussi les noms génériques des graphes que nous utilisons. Dans *GNmesure*, l'unité est optionnelle alors qu'elle est obligatoire dans *GNmesureFinal* (ex : *entre GNmesure et GNmesureFinal*).

- Problèmes stylistiques de la séquence *Dnum Unité*

Comme nous l'avons mentionné, un nombre en lettres ne peut pas être suivi d'une unité sous la forme d'un symbole alors qu'un nombre en chiffres peut être suivi par n'importe quel type d'unités : *deux mètres* ; **deux m* ; *2 m* ; *2 mètres*. Par ailleurs, les combinaisons complexes requièrent une certaine homogénéité dans le choix des types de déterminants numériques et d'unités. Par exemple, il semble difficilement concevable d'avoir la séquence suivante : *entre 4 et cinq mètres*. Nos graphes ne tiennent pas compte de cette dernière règle par souci de clarté et de simplicité. Ce choix peut causer quelques rares erreurs de reconnaissance dans les textes.

- Ambiguïté des combinaisons complexes

Il y a une ambiguïté dont nous n'avons pas tenu compte dans nos graphes. En effet, la séquence *entre 4 et 5 millions de dollars* est interprétée comme *entre 4 dollars et 5 millions de dollars*. Mais il existe une autre interprétation qui est, dans la quasi-totalité des cas, la bonne (à cause du point précédent : homogénéité des déterminants numériques). Il faut considérer que c'est la séquence *millions de dollars* qui a été factorisée et non *dollars*. La séquence précédente doit alors être analysée comme *entre 4 millions de dollars et 5 millions de dollars*.

- Récursivité dans les combinaisons complexes

Les combinaisons complexes en *entre … et …* et *de … à …* sont théoriquement récursives. En effet, la règle récursive suivante semble pouvoir s'appliquer : *Dnum* → *entre Dnum et Dnum*. Cependant, le nombre de niveaux est très étroitement limité : l'effacement du *N* =: *valeur* est interdit.

> *?* La tension est entre [de (10 V) à 31 V] et [de (4 kV) à (5 kV)]*
> *La tension est entre une valeur de 10 à 30 V et une valeur de 4 à 5 kV*

Notre choix de ne pas décrire récursivement ce type de séquences paraît donc fondé. Notre représentation est donc équivalente à un automate fini.

3.3 Représentation des mesures absolues

3.3.1 Généralités

La structure élémentaire qui nous intéresse représente l'expression de la mesure (absolue) d'une caractéristique ou propriété intrinsèque (désignée par *Ng*) d'un élément *N0* :

*N*0 *avoir un Ng de Dnum Unité*
Le bateau a une longueur de 15 mètres

La première phrase indique que *le bateau a une longueur* (i.e. *N*0 a une caractéristique *Ng*) et puis que *cette longueur est de 15 mètres* (i.e. mesure de la caractéristique *Ng*).
Nous avons étendu les résultats de J. Giry-Schneider (1991) à un plus grand nombre de noms et de propriétés syntaxiques associées. Nous regardons aussi l'application de cette analyse à la reconnaissance automatique de ce type d'expressions dans des textes.

3.3.2 Ng composés

Dans un premier temps, nous avons sélectionné un ensemble de caractéristiques *Ng* qui entrent dans cette structure de phrase. Nous prenons les plus courantes : longueur, poids, coût, température, etc. Au total, nous en avons sélectionné une quarantaine. Nous regardons également quelques noms du domaine économique comme *loyer* pour l'exemple, mais ceci aurait nécessité une étude bien plus approfondie. Nous renvoyons aux travaux de M. Gross (1997) et de T. Nakamura (à paraître) sur le domaine de la bourse. Les noms simples sont, par exemple, les noms *longueur*, *poids*, *vitesse*, *force*, *coût*, etc. Plus des deux-tiers ont cette forme. D'autres sont des mots composés. Nous pouvons les classer en plusieurs classes selon leur structure interne (G. Gross, 1996) :

- une classe NA (nom suivi d'un adjectif) : *tension électrique* ; *tension artérielle* ; *densité démographique* ; *pression atmosphérique* ; *intensité lumineuse* ; *intensité électrique* ; *puissance énergétique* ; *loyer mensuel*

- une classe NDN (nom suivi de la préposition *de* puis d'un nom) : *taille de chaussure* ; *pointure de pied(s)* ;

- une classe complexe NDNA : *taille de mémoire (vive + cache + virtuelle)*

La quasi-totalité de ces noms composés se retrouvent dans les textes sous une forme simple résultant d'un effacement d'un ou plusieurs composants du mot composé. En général, c'est la partie à droite du premier nom qui est effacée comme par exemple dans :

Max a une tension (artérielle + E) de 10
Paris a une densité (démographique + E) de 10 000 hab/km2
Luc a une pointure (de pieds + E) de 43

Certains noms composés du domaine de l'informatique (ex : *taille de mémoire vive*) ont des formes ambiguës particulières. La préposition *de* peut être effacée si l'on supprime l'adjectif :

Cette machine a une taille de mémoire vive de 128 Mo
*Cette machine a une taille mémoire (E + *vive) de 128 Mo*

Le nom de tête *taille* peut aussi disparaître dans la séquence d'origine ; l'effacement de l'adjectif y est toujours possible :

Cette machine a une mémoire (E + vive) de 128 Mo

Notons pour finir que *loyer mensuel* se comporte différemment et ne peut être traité qu'au niveau de la phrase (et non localement comme pour les précédents). En effet, l'adjectif *mensuel* peut à la fois être effacé et transformé en adverbe (*mensuellement*) pouvant s'insérer n'importe où dans la phrase.

Marie a un loyer (mensuel + E) de 1 000 euros
Mensuellement, Marie a un loyer de 1 000 euros

Il en est de même pour d'autres noms dénotant des flux journaliers, mensuels, annuels, etc. : *débit, flux*, … Ce comportement est impossible pour les autres noms composés de la classe NA :

Max a une tension (artérielle + E) de 10
** Artériellement, Max a une tension de 10*

Paris a une densité (démographique + E) de 10 000 hab/km2
? Démographiquement, Paris a une densité de 10 000 hab/km2*

Par la suite, nous décidons de ne pas traiter les noms tels que *loyer*.

3.3.3 Propriétés distributionnelles, lexicales et transformationnelles

Nous étudions maintenant les variations lexicales et les transformations que peut subir notre phrase de base :

N0 avoir un Ng de Dnum Unité

3.3.3.1 Distribution du sujet

La distribution du sujet dépend du nom. Nous distinguons trois types de sujets : les groupes nominaux humains (*Nhum*), les groupes nominaux concrets (*Nconc*) et les groupes nominaux prédicatifs (*Npred*). Par exemple, le nom *durée* ne sélectionne ni les sujets humains et ni les sujets concrets :

*(Le spectacle + *Paul + *La corde) a une durée de dix minutes*

Le nom composé *tension artérielle* ne sélectionne que les noms humains alors que *taille* interdit les noms prédicatifs :

*(*Le spectacle + Paul + *la corde) a une tension de 12*
*(*Le spectacle + Paul + la corde) a une taille de 2 m*

La distribution du sujet permet de lever l'ambiguïté du nom *longueur*. En effet, l'une des deux entrées a la même distribution du sujet que *durée*, alors que l'autre a la même distribution que *taille*.

*(Le spectacle + *Paul + *la corde) a une longueur de dix minutes*

*(*Le spectacle + La baleine + la corde) a une longueur de 20 m*

3.3.3.2 Verbes supports

D'abord, le schéma de phrase précédent est un cas particulier du schéma de phrase ci-dessous :

*N*0 *Vsup Prép un Ng de Dnum Unité*

Le verbe support et la préposition associée peuvent connaître des variations lexicales (*avoir, faire, être (à + de), compter, contenir*, etc.) :

Cet immeuble a une hauteur de 150 mètres
= Cet immeuble fait une hauteur de 150 mètres (*Vsup =*: *faire*)

La salle des fêtes a une température de 30°C
= la salle des fêtes est à une température de 30°C (*Vsup =*: *être à*)

Ce spectacle a une longueur de deux heures
= Ce spectacle est d'une longueur de deux heures (*Vsup =*: *être de*)

L'agglomération parisienne a une population de dix millions d'habitants
= L'agglomération parisienne compte une population de dix millions d'habitants (Vsup =: *compter*)

Ces aliments ont une énergie de 10 kJ
= Ces aliments contiennent une énergie de 10 kJ (*Vsup = : contenir*)

Cependant, cette variation dépend du *Ng*. Une étude systématique est nécessaire :

*Max (a + fait + est de + *est à[39] + *comporte) une taille de 2 m*
*Cette propriété (a + fait + ?est de + *est à + ?comporte) une surface de 2 hectares*
*Cette salle (a + fait + ?*est de + est à + *comporte) une température de 17°C*
*Ces aliments (ont + *font + *sont de + *sont à + comportent) une énergie de 10 kJ*
*Ce bus (a + *fait + *est de + est à + *comporte) une vitesse de 100 km/h*

Notons que l'utilisation de *température* avec le verbe support *faire* est autorisée dans une phrase au sujet impersonnel de la forme :

*Il faire Dnum Unité Loc N*0
= : Il fait 10°C dans cette salle

Ce phénomène semble marcher pour *pression* et *hygrométrie* mais il n'existe pas pour les autres noms :

** Il fait une longueur de 100 m (sur + dans) le bateau*
** Il fait une tension de 50 kV (sur + dans) cette ligne*

[39] Il existe un emploi de *être à* qui fonctionne dans ce cas mais il dénote un état temporaire dans une évolution : *Dans sa phase de croissance, Max est (déjà + E) à une taille de 1,70 m.*

3.3.3.3 Permutations

Nous regardons maintenant les transformations que peuvent subir les phrases dont la structure est :

N0 Vsup Prép un Ng de Dnum Unité

Tout d'abord, les séquences *Dnum Unité* et *Ng* peuvent être permutées, le déterminant *un* étant effacé. Cette permutation ne fonctionne pas pour tous les *Ng* :

L'immeuble fait une hauteur de 100 m
= l'immeuble fait 100 m de hauteur

*Le courant (a + *fait) une fréquence de 500 Hz*
**Le courant (a + fait) 500 Hz de fréquence*

Cette propriété est un autre moyen de distinguer les deux emplois de *tension* car ils n'ont pas le même comportement :

Max a une tension de 12
= Max a 12 de tension

La ligne (fait + a) une tension de 220V
*= ?*La ligne (fait + a) 220 V de tension*

Notons que l'utilisation de certains *Vsup* est plus naturelle que pour d'autres ; c'est le cas de *faire* :

Le bateau (a + fait) une longueur de 110 m
Le bateau (?a + fait) 100 m de longueur

Ces permutations sont parfois accompagnées d'accidents morphologiques. Certains *Ng* sont parfois remplacés par des adjectifs morphologiquement associés (*Ng-a*). Ce n'est d'ailleurs le cas que pour les *Ng* sélectionnant une unité métrique :

La corde fait 10 m de long
Le gratte-ciel fait 200 m de haut
La piscine fait 20 m de large

Ce phénomène ne fonctionne pas pour *épaisseur* :

Le mur a une épaisseur de 30 cm
Le mur a 30 cm d'épaisseur
**Le mur a 30 cm d'épais*[40]

Le nom *profondeur* (*Ng-a* =: *profond*) subit un accident plus étrange encore car il peut être remplacé par le nom *fond* (noté *Ng'*) :

Le bassin a 3 m de profondeur

[40] Cette dernière phrase est acceptée en français du Québec.

**Le bassin a 3 m de profond*
Le bassin a 3 m de fond

Notons que le nom *fond* ne rentre pas dans la phrase de base équivalente :

? Le bassin a un fond de 3 m*

Cette propriété permet également de distinguer les deux emplois de *longueur* car ils n'ont pas le même comportement :

Le bateau fait 100 m de long
** Le spectacle fait 2 heures de long*

Notons qu'il est possible de substituer la séquence *en Ng* à la séquence *de Ng*, lorsque l'on a le verbe support *faire*. Cette séquence a alors le comportement d'un adverbe car elle peut s'insérer n'importe où dans la phrase :

La piscine fait 50 mètres (de + en) longueur
*(En + ?*de) longueur, la piscine fait 50 mètres*

3.3.3.4 Nominalisation et adjectivation

Notre phrase de base est également sujette à une adjectivation :

*N*0 *Vsup (Prep) un Ng de Dnum Unité*
*= N*0 *être Ng-a de Dnum Unité*[41]

La cour est large de 50 m
? Le bassin est volumineux de 40 litres*

Tous les *Ng* ne possèdent pas de *Ng-a* associé : *tension*, *température*, etc.

Notre phrase de base peut aussi être transformée en une phrase à prédicat verbal *Ng-v* où *Ng-v* est le verbe morphologiquement associé à *Ng* :

*N*0 *Vsup (Prep) un Ng de Dnum Unité*
*= N*0 *Ng-v Dnum Unité*

Max a un poids de 30 kg
= Max pèse 30 kg

la chaise a un coût de trente euros
= la chaise coûte trente euros

Le nom *population* a un comportement différent car le sujet de *Ng-v* est *Dnum Unité* et son objet est *N*0 :

*N*0 *Vsup (Prep) un Ng de Dnum Unité*
*= Dnum Unité Ng-v N*0

[41] *Ng- a* est l'adjectif morphologiquement lié à *N*.

Le village a une population de 300 habitants
= 300 habitants peuplent le village

Nous remarquons après une analyse quasi-exhaustive que l'intersection entre l'ensemble de nos *Ng* qui entrent dans une structure adjectivale et l'ensemble de nos *Ng* qui entrent dans une structure verbale est vide.

3.3.3.5 Effacement du nom prédicatif

Dans notre structure de base, le nom *Ng* peut être effacé, mais cela dépend d'abord du *Ng* et du *Vsup* :

N0 Vsup (Prep) un Ng de Dnum Unité
= N0 Vsup (Prep) Dnum Unité

Cette corde fait une longueur de trente mètres
Cette corde fait trente mètres

Cette ligne (a + ?fait) une tension de 220V
Cette ligne fait 220V

Cette transformation est difficilement réalisable avec le verbe support *avoir*, excepté pour quelques noms comme *âge*.

Cette corde a une longueur de trente mètres
**Cette corde a trente mètres*

Max a un âge de 10 ans
Max a 10 ans

Le verbe support *faire* est souvent le plus naturel, mais le verbe *être à* est aussi possible :

La salle de classe (fait + est à) une température de 20°C
La salle de classe est à 20°C
la salle de classe fait 20°C

Certains *Ng* ne s'effacent pas (ou très difficilement) comme *périmètre*.

La piscine a un périmètre de 30 m
** La piscine fait 30 m*

On s'aperçoit que la propriété dépend aussi de la nature de *N*0. On ne comprend la phrase *la corde fait 30 m* que parce qu'une corde a la propriété d'être longiligne et on en déduit que la caractéristique mesurée est la longueur. Pour la vitesse, on observe un phénomène particulier : l'effacement de vitesse dans la phrase de base est interdit sauf dans le cas exceptionnel où *N*0 =: *vent* (ou *tornade*, *courant*, etc…). Comme l'explique J. Giry-Schneider (1991), ceci semble être dû à la nature dynamique du vent.

**Cette voiture (fait + être de) 10 km/h*
Ce vent (fait + être de) 20 km/h

Certains noms comme *coût* n'entrent pas dans le schéma de phrase de base en *faire*, mais sont acceptés dans une forme réduite (*Ng* effacé).

*Cet achat (a + *fait) un coût de 30 euros*
*Cet achat (*a + fait) 30 euros*

Certains noms comme *vitesse* ne sont effaçables que si l'on rajoute le déterminant partitif *du* :

*Le bus fait (*E + du) 35 km/h*

Les deux structures (avec et sans déterminant partitif) sont acceptables avec les noms *intensité électrique* et *tension électrique* :

La ligne fait (E + du) 220V
La ligne fait (E + du) 2A

Par contre, l'ajout du déterminant partitif *du* n'est pas valable pour tous les noms :

** La corde fait du 10 m.*

Le verbe *mesurer* peut aussi être utilisé à la place du verbe support ; l'acceptabilité de la structure engendrée dépend aussi du *Ng* effacé.

La corde (fait + mesure) 10 m
*La ligne (fait + *mesure) 220 V*

Il est parfois possible d'ajouter un adjectif non prédicatif entre *Dnum* et *Unité* pour nuancer une mesure objective. Cet adjectif a un peu le même rôle sémantique qu'un prédéterminant.

La corde fait cinq (*petits mètres + *mètres qui sont petits*)

3.3.3.6 Autres

Revenons maintenant à notre phrase de base *N*0 *avoir un Ng de Dnum Unité*. Comme nous l'avons dit au début de la section, elle peut s'analyser à partir de deux phrases élémentaires : *N*0 *avoir un Ng* et *Ce Ng être de Dnum Unité*. En réduisant la première phrase au groupe nominal *le Ng de N0* (=: *la longueur de la piscine*) que nous substituons à *ce Ng* dans la deuxième phrase, nous obtenons une autre structure équivalente à notre structure de base :

(N0 avoir un Ng ; ce Ng être de Dnum Unité)
= Le Ng de N0 être de Dnum Unité

La longueur du chemin est de 100 m
= Sa longueur est de 100 m

Pour finir, si nous utilisons la notion d'opérateur à lien[42] de M. Gross (1981), nous avons une équivalence entre les phrases suivantes

La longueur du chemin être de 100 m

[42] Exemple : *La sœur de Léa est malade = Léa a sa sœur qui est malade = Léa a sa sœur malade*

? Le chemin a sa longueur qui est de 100 m

Par ailleurs, on observe :

Le chemin a pour longueur 100 m

Nous synthétisons toutes les structures équivalentes dans lesquelles peuvent rentrer les noms *Ng* dans le graphe théorique **N0AvoirUnNDeDnumUnite** ci dessous[43].

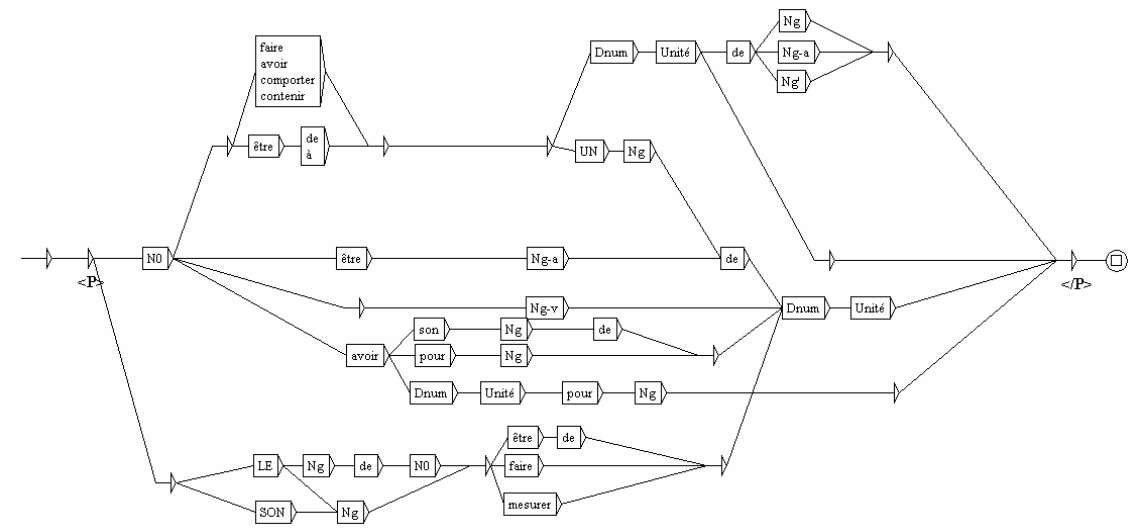

Figure 51 : N0AvoirUnNDeDnumUnite

3.3.4 Codage des propriétés

Notre analyse linguistique a montré l'existence de nombreuses contraintes dépendant des noms *Ng*. Nous souhaitons maintenant coder ces contraintes dans une table syntaxique. Nous nous inspirons de J. Giry-Schneider (1991).

Dans notre étude, les entrées lexicales correspondent aux prédicats nominaux *Ng* rentrant dans le schéma de phrase *N0 avoir un Ng de Dnum Unité*. Les propriétés que nous voulons coder sont les suivantes :

- la distribution du sujet
- la valeur lexicale des noms (*longueur, tension artérielle*)
- la contrainte entre *Ng* (ex : *largeur*) et *Unité* (ex : *mètre*)
- la variation lexicale du verbe support : *avoir, être (de + à), comporter,* etc.
- la permutation entre la séquence *Dnum Unité* et le nom *Ng*, plus les accidents morphologiques
- les transformations de nominalisation et d'adjectivation
- l'effacement du nom *N*

La table syntaxique codée est donnée plus bas et a le nom **ANMesure**. Les structures codées sont explicitement données dans l'intitulé des colonnes.

[43] Ce graphe n'est pas directement applicable à un texte (cf. section application à des textes)

Distribution du sujet

Les trois premières colonnes contiennent le codage de la distribution du sujet : *Nhum*, *Nconc* et *Npred*.

Valeur des noms

Nous indiquons la forme de base de nos prédicats nominaux *Ng*. Dans le cas des noms composés, le nom et l'adjectif apparaissent dans deux colonnes séparées.

Remarque : nous n'avons codé que partiellement le comportement de *taille de mémoire vive* pour éviter d'ajouter trop de colonnes à notre table. Nous n'avons entré que la variante réduite *mémoire vive* de type *NA*.

Contrainte Ng - Unité

Pour chaque entrée nous associons un ensemble d'unités appropriées sous la forme d'un nom de graphe (précédé de *:*). Si l'unité est vide, nous insérons le mot vide <*E*>.

Variation lexicale du verbe support

Les colonnes G, H et I correspondent respectivement aux emplois de *avoir*, *faire* et *comporter*, comme verbes supports. Nous codons aussi la propriété *il faire un Ng de Dnum Unité Loc N*0 dans la colonne J.

Remarque : nous n'avons codé qu'un certain nombre de verbes supports, seulement pour montrer que leur variation dépendait de *Ng*, comme l'avait fait J. Giry-Schneider sur un nombre de *Ng* plus réduit. Un codage complet n'est pas forcément intéressant pour l'instant car nous sommes dans un cadre assez théorique (cf. partie *réduction de la structure de base*).

La permutation de la séquence Dnum Unité et le nom Ng

L'adjectif dérivé de *Ng* est donné dans la colonne K. Le nom *Ng'* (accident morphologique du *Ng* lors de la permutation) est mis dans la colonne N. Les trois structures engendrées par les permutations sont dans les colonnes L, M et O.

Nominalisation et adjectivation

Le verbe morphologiquement et sémantiquement lié à *Ng* est donné dans la colonne Q. La possibilité d'avoir des structures à prédicat adjectival et verbal est codée dans la colonne P, R et S.

Effacement du prédicat nominal N

Nous avons codé la possibilité d'effacer *Ng* en faisant varier le verbe support et sa préposition associée : *être de* (T), *être à* (U), *faire* (V et W), *contenir* (X et Y), *compter* (Z). Pour les verbes *faire* et *contenir*, nous avons regarder la possibilité d'avoir le déterminant partitif *du* devant la séquence *Dnum Unité*.

Remarque : pour *vitesse*, comme l'effacement du Ng ne fonctionne que pour une classe sémantique de sujets extrêmement réduite (*vent, tornade*, etc.), nous mettons un signe '-' dans la case correspondante à cette propriété et à cette entrée. Nous construisons par ailleurs une grammaire locale pour cette classe à l'aide notamment du dictionnaire en-ligne des synonymes de CRISCO (http://elsap1.unicaen.fr/cherches.html). Nous reviendrons sur cette grammaire ultérieurement.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
N0 =: Nhum	N0 =: Nconc	N0 =: Npred	Ng	Modif Appr	classe d'unités	N0 avoir un Ng de Dnum Unité	N0 faire un Ng de Dnum Unité	N0 comporter un Ng de Dnum Unité	Il faire un Ng de Dnum Unité (Loc N0+E)	Ng-a	Ng permuté	Ng-a permuté (accident)	Ng'	Ng' permuté	N0 être Ng-a de Dnum Unité	Ng-v	N0 Ng-v Dnum Unité	Dnum Unité Ng-v N0	N0 être de Dnum Unité	N0 être à Dnum Unité	N0 faire Dnum Unité	N0 faire du Dnum Unité	N0 contenir Dnum Unité	N0 contenir du Dnum Unité	N0 compter Dnum Unité	N0 avoir Dnum Unité
+	+	+	âge	<E>	:GNmesure-temps	+	-	-	-	âgé	+	-	-	-	+	-	-	-	-	-	-	-	-	-	+	+
+	+	-	aire	<E>	:GNmesure-surface	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-
+	+	-	allure	<E>	:GNmesure-vitesse	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-
+	+	-	capacité	<E>	:GNmesure-volume	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+	+	+	coût	<E>	:GNmesure-monetaire	+	-	-	-	-	-	-	-	-	-	coûter	+	-	-	+	+	-	-	-	-	-
-	+	-	densité	démographique	:GNmesure-densite-pop	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+
+	+	-	diamètre	<E>	:GNmesure-longueur	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	+	durée	<E>	:GNmesure-temps	+	-	-	-	-	-	-	-	-	-	durer	+	-	+	-	+	-	-	-	-	-
+	+	-	énergie	<E>	:GNmesure-energie	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+
+	+	-	épaisseur	<E>	:GNmesure-longueur	+	+	-	-	épais	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
+	+	-	force	<E>	:GNmesure-force	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+	+	-	fréquence	<E>	:GNmesure-frequence	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-	-	-
+	+	-	hauteur	<E>	:GNmesure-longueur	+	+	-	-	haut	+	+	-	-	+	-	-	-	+	-	+	-	-	-	-	-
+	+	-	intensité	électrique	:GNmesure-intensite-elec	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	+	-	-
+	+	-	largeur	<E>	:GNmesure-longueur	+	+	-	-	large	+	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-
+	+	-	longueur	<E>	:GNmesure-longueur	+	+	-	-	long	+	+	-	-	+	-	-	-	+	-	+	-	-	-	-	-
-	-	+	longueur	<E>	:GNmesure-temps	+	+	-	-	long	-	-	-	-	+	-	-	-	+	-	+	-	-	-	-	-
+	+	-	masse	<E>	:GNmesure-masse	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-
-	+	-	mémoire	vive+virtuelle+cache	:GNmesure-informatique	+	-	+	-	-	+	-	-	-	-	-	-	-	+	-	+	-	+	-	-	+
+	+	-	périmètre	<E>	:GNmesure-longueur	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+	+	-	périmètre	<E>	:GNmesure-surface	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+	+	-	poids	<E>	:GNmesure-masse	+	+	-	-	-	-	-	-	-	-	peser	+	-	+	-	+	-	-	-	-	-
+	+	-	poids	<E>	:GNmesure-force	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
+	-	-	pointure	de <pied>+de <chaussure>	<E>	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
-	+	-	population	<E>	:GNmesure-population	+	-	+	-	-	-	-	-	-	-	peupler	-	+	-	-	-	-	+	-	+	+
-	+	-	pression	atmosphérique	:GNmesure-pression	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	+	-	profondeur	<E>	:GNmesure-longueur	+	+	-	-	profond	+	-	fond	+	+	-	-	-	-	-	+	-	-	-	-	-
+	+	-	puissance	énergétique	:GN-mesure-puissance	+	+	-	-	puissant	+	-	-	-	-	-	-	-	+	-	+	+	-	-	-	+
+	+	-	rayon	<E>	:GNmesure-longueur	+	+	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+	+	-	superficie	<E>	:GNmesure-surface	+	+	+	-	-	+	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-
+	+	-	surface	<E>	:GNmesure-surface	+	+	+	-	-	+	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-
+	+	-	taille	<E>	:GNmesure-longueur	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
+	-	-	taille	de <pied>+de <chaussure>	<E>	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
+	+	-	température	<E>	:GNmesure-temperature	+	+	-	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-
+	+	-	tension	électrique	:GNmesure-tension	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	+	-	-
+	-	-	tension	artérielle	<E>	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+	+	-	vitesse	<E>	:GNmesure-vitesse	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-
+	+	-	volume	<E>	:GNmesure-volume	+	+	-	-	volumineux	-	-	-	-	+	-	-	-	-	-	+	-	+	-	-	-

Table 3 : ANMesure

3.3.5 Les prédéterminants

Nous regardons maintenant le comportement de certains prédéterminants dans notre phrase de base. Nous reprenons la liste des prédéterminants fournie dans M. Gross (1977) et étudions systématiquement ceux qui peuvent apparaître dans les structures étudiées. A travers cette liste de taille modeste, nous montrons la difficulté de traiter ces composants localement. Notre liste de prédéterminants est la suivante :

> *comme, d'abord, encore, ensuite, environ, jusqu'à, même, ne…que, plutôt, presque, quelque, seul, à demi, à peine, au mieux, approximativement* et *pas tout à fait*

Nous avons travaillé sur les structures *N*0 *avoir un Ng de Dnum Unité* et *N*0 *faire Dnum Unité* et nous avons essayé d'insérer systématiquement ces prédéterminants à différents endroits. Nous avons constaté une grande variation de comportement selon les prédéterminants. Nous avons testé pour chaque prédéterminant *Préd* l'acceptabilité des sept structures suivantes :

(1) N0 avoir un Ng de Préd Dnum Unité
=: Ce bateau a une longueur d'environ 100 m

(2) N0 faire Préd Dnum Unité
=: Ses appartements font jusqu'à 50 mètres carrés

(3) N0 avoir un Ng de Dnum Unité Préd
=: Luc a un poids de 60kg à peine

(4) N0 faire Dnum Unité Préd
=: Cette ligne fait 110 V approximativement

(5) Préd N0 avoir un Ng de Dnum Unité
=: Au mieux mon moteur a une fréquence de dix tours par minute

(6) N0 avoir Préd un Ng de Dnum Unité
=: Cette ville maudite a encore une population de 100 habitants.

(7) N0 avoir un Ng Préd de Dnum Unité
=: Son spectacle n'a une durée que de dix minutes

Après examen exhaustif, nous constatons que seuls trois prédéterminants peuvent s'insérer n'importe où dans la phrase : *approximativement, au mieux* et *plutôt.* Ils jouent clairement des rôles d'adverbes. D'autres ont quelques restrictions comme les prédéterminants *à peine* ou *environ* dans :

**(A peine + Environ) Luc a une taille de 1,50 m.*
*Luc a une taille (*à peine + ?environ) de 1,50 m*

La structure non-connexe *ne … que* s'emploie sans difficulté sauf dans les cas clairs suivants :

** Que son spectacle n'a une durée de dix minutes*
** Son spectacle n'a une durée de que dix minutes*
** Son spectacle n'a une durée de dix minutes que*

D'autres ne s'emploient jamais tels que *seul* et *à demi*.

Nous construisons une table syntaxique représentant les comportements des prédéterminants. Chaque ligne correspond à un prédéterminant de notre lexique. La première colonne contient les prédéterminants. Les sept autres colonnes correspondent aux sept structures ci-dessus. Le signe '+' dans une case signifie que l'entrée correspondant à la ligne de cette case rentre dans la structure associée à sa colonne. Le signe '-' signifie le contraire.

Préd	N0 avoir un Ng de Préd Dnum Unité	N0 faire Préd Dnum Unité	N0 avoir un Ng de Dnum Unité Préd	N0 faire Dnum Unité Préd	Préd N0 avoir un Ng de Dnum Unité	N0 avoir Préd un Ng de Dnum Unité	N0 avoir un Ng Préd de Dnum Unité
comme	-	-	-	-	-	-	-
d'abord	-	-	-	-	+	+	-
encore	-	+	+	+	+	+	-
ensuite	-	-	-	-	+	+	-
environ	+	+	+	+	-	+	+
jusqu'à	-	+	-	-	-	+	-
même	-	+	-	-	-	+	-
ne...que	-	+	-	-	-	+	+
plutôt	+	+	+	+	+	+	+
presque	+	+	-	-	-	+	-
quelque	+	+	-	-	-	-	-
seul	-	-	-	-	-	-	-
à demi	-	-	-	-	-	-	-
à peine	+	+	+	+	-	+	-
au mieux	+	+	+	+	+	+	+
approximativement	+	+	+	+	+	+	+
pas tout à fait	+	+	-	-	-	+	-

Table 4 : Pred

Ainsi, pour décrire précisément les phrases de mesure en tenant compte des prédéterminants, il faut construire un graphe de prédéterminants pour chaque point potentiel d'insertion dans la structure, à l'aide de notre table syntaxique. Notre liste ne contenant qu'une petite partie des prédéterminants, notre travail est incomplet et ne permet pas de décrire précisément le comportement de tous les prédéterminants dans les phrases de mesure. Il confirme que l'on ne peut pas tenir compte des prédéterminants dans les phrases de mesure si l'on ne regarde pas la phrase complète (cf. M. Gross, 1977).

3.3.6 Réduction de la phrase élémentaire

Les phrases que nous avons étudiées dans la partie précédente sont très théoriques car elles apparaissent très peu telles quelles dans les textes. En fait, elles se retrouvent sous la forme de groupes nominaux qui sont des réductions de ces phrases. Dans cette partie, nous décrivons les processus linguistiques permettant de passer des phrases de base à ces séquences.
Nous partons de quatre schémas de phrases équivalents à *N0 avoir un Ng de Dnum Unité* :

*(a) N*0 *avoir un Ng de Dnum Unité*
*(b) N*0 *être Prep un Ng de Dnum Unité*
*(c) N*0 *être Ng-a de Dnum Unité*
*(d) Le Ng de N*0 *être de Dnum Unité*

Tout d'abord, on peut voir la structure (a) comme équivalente à *N*0 *avoir N*1 (=: *Max a un ballon*) qui se réduit en :

N1 (qu'avoir + de) N0 =: le ballon (qu'a + de) Max

Ainsi, on a :

La corde a une longueur de 10 cm
la longueur de 10 cm de la corde (réduction)

La salle a une température de 10°C
La température de 10°C de la salle (réduction)

Cette propriété a un diamètre de 14 km
Le diamètre de 14 km de la propriété (réduction)

Le spectacle a une durée de 10 min
La durée de 10 min du spectacle (réduction)

Lorsque le nom *Ng* est effacé (cf. précédemment), le déterminant est toujours *les.*

*(les + *E)10 cm de la corde*
*(les + *E) 10°C de la salle*
*(les + *E) 14 km de la propriété*
*(les + *E) 10 min du spectacle*

Pour certains *Ng*, les deux déterminants peuvent être indéfinis :

une longueur de 10 cm de (E + la) corde*
**une température de 10°C de (E + la) salle*

Le nom *Ng* peut souvent s'effacer ; dans ce cas, le premier déterminant peut être défini :

*(les + E) 10 cm de (E + *la) corde*

De plus, dans cette construction, *Det Ng de Dnum Unité de* a le statut des déterminants nominaux étudiés par P.A. Buvet (1993,1994). Le nombre de classes de noms prédicatifs *Ng* rentrant dans cette structure est limité :

- *volume, capacité* (*GNmesure-volume*)
- *longueur* (*GNmesure-longueur*)
- *durée, longueur* (*GNmesure-temps*)
- *coût, prix* (*GNmesure-monnaie*)
- *surface, superficie, aire* (*GNmesure-surface*)
- *poids, masse* (*GNmesure-masse*)

Du fait de différences de sens qui peuvent être importantes, nous représentons ces séquences dans une autre grammaire que celle décrivant les réductions nominales de nos phrases de base. Nous tenons compte uniquement des formes dans lesquelles le *Ng* est effacé du type *(E + les) 10 cm de corde*. Nous donnons ci-dessous le graphe (**DnumUniteDe**) répertoriant l'ensemble de ce type de déterminants nominaux :

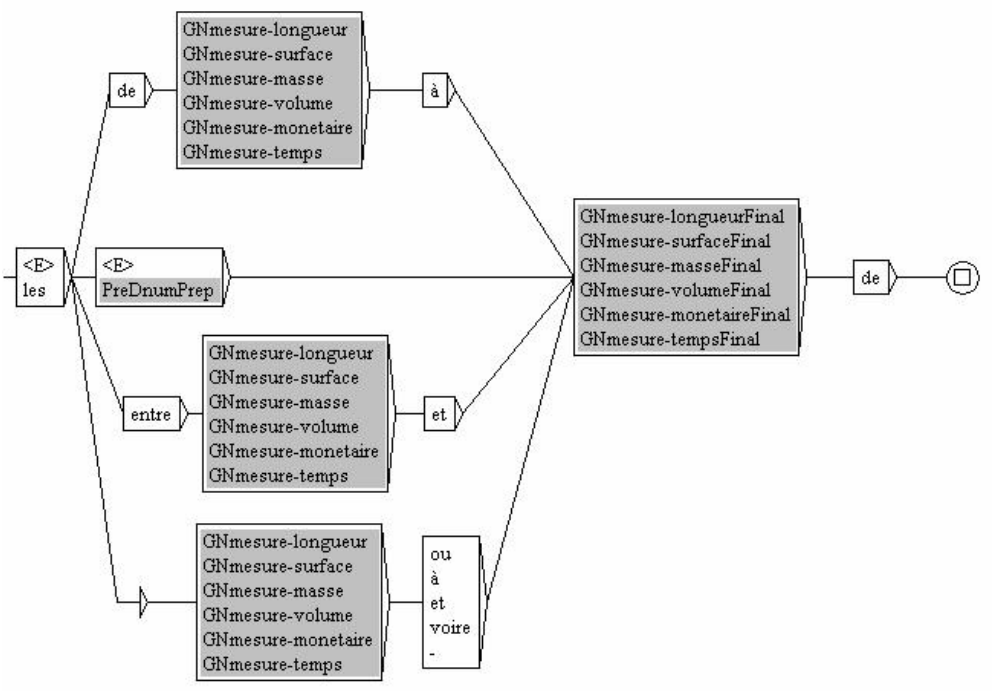

Figure 52 : DnumUniteDe

La structure (b) qui correspond au schéma de phrase *N*0 *être Prep N*1 donne lieu à une relative sujette à la réduction suivante :

*N*0 *(qui être+ E) Prep N*1
=: *Ce projet (qui est + E) de grande envergure*
=: *Cet homme (qui est + E) à la rue*

*N*0 *(qui être + E) Prep un Ng de Dnum Unité*
=: *la corde (qui est + E)d'une longueur de 10 mètres*
=: *L'eau (qui est + E) à une température de 100°C*

L'effacement du nom prédicatif *Ng* dans la forme réduite est possible lorsqu'il est possible dans la phrase de base :

Son tuyau est d'une longueur de 10 m
Son tuyau est de 10 m
Son tuyau de 10 m

L'eau est à une température de 50 degrés
L'eau est à 50 degrés
L'eau à 50 degrés

Comme on l'a déjà noté, le nom *vitesse* a un comportement particulier. Il n'accepte pas d'effacement sauf pour les *N*0 de la classe des vents :

** un bus de 40 km/h*
un (vent + courant) de 30 nœuds

Les phrases en *être à* génèrent des groupes prépositionnels adverbiaux : *à 10 km/h*. Ce verbe support a pour variantes certains verbes qui dépendent du sujet utilisé :

*(La voiture + Max + *Le vent) roule à (une vitesse de + E) 17 km/h*
*(*La voiture[44] + Max +* Le vent) court à (une vitesse de + E) 17 km/h*
*(*La voiture + ?*Max + Le vent) souffle à (une vitesse de + E) 60 noeuds*

Comme on l'indiqué précédemment, nous construisons une grammaire spécifique à la classe des vents. Nous ajoutons aussi la possibilité de reconnaître les expressions :

Un vent de force 8
Un vent de force 8 à 9

Le graphe représentant ces expressions est donné ci-dessous (**UnVentDeDnumNmesure-vitesse**). Les expressions reconnues par ce graphe sont semi-figées car la variation lexicale est faible pour *N*0 et la structure est figée.

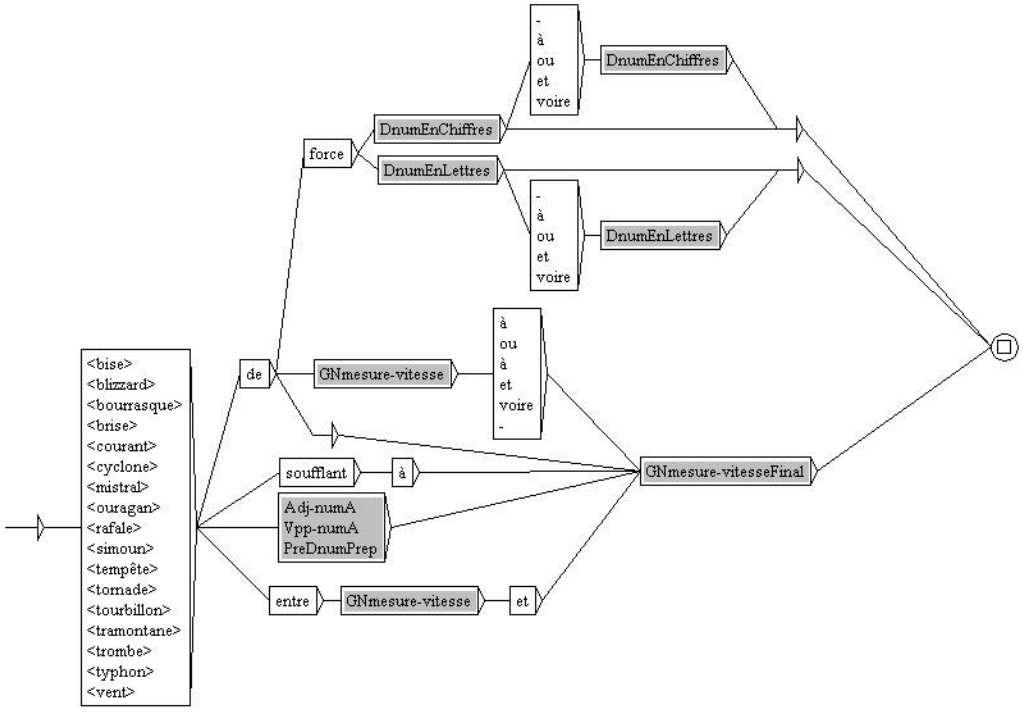

Figure 53 : VentDeDnumNmesure-vitesse

La troisième structure (c) peut également être sujette à une réduction à un groupe nominal. La partie à droite du verbe support *être* devient alors un modifieur adjectival de *N*0.

Max veut escalader une falaise (qui est + E) haute de 120 m

La dernière structure (d) peut aussi être réduite, même si cela paraît peu naturel :

La hauteur de la falaise est de 100 m
?La hauteur de la falaise de 100 m

[44] Ce nom peut être accepté si on imagine que l'on a un homme déguisé en voiture.

La pronominalisation du *N*0 la rend plus naturelle :

Sa hauteur de 100 m m'effraie

Nous proposons ci-dessous le graphe paramétré décrivant l'ensemble des formes réduites de notre phrase de base. Ces formes réduites sont toutes des groupes nominaux. Pour l'entrée *largeur*, nous générons le graphe associé. La variable *@D* est remplacée par l'entrée lexicale *largeur*. La variable *@L* correspondant à la propriété de permutation entre *Dnum Unité* et *Ng* est remplacée par le mot vide (<*E*>) car cette propriété est autorisée pour cette entrée (+). Par contre, la boîte contenant *@U* est supprimée car la structure correspondante (*N*0 *être à Dnum Unité*) est interdite (symbole -), etc.

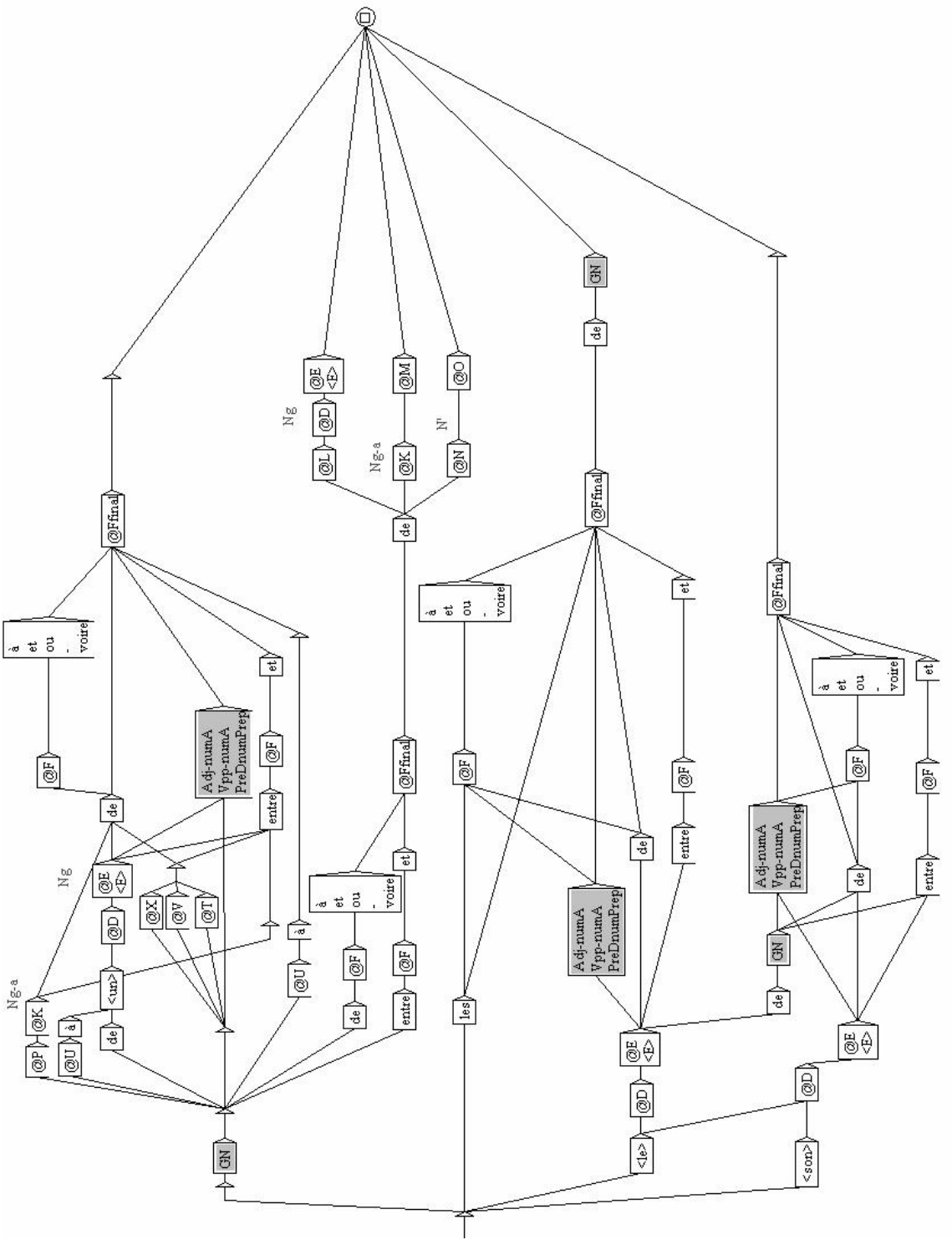

Figure 54 : graphe patron décrivant les réductions de N0 avoir un Ng de Dnum Unite

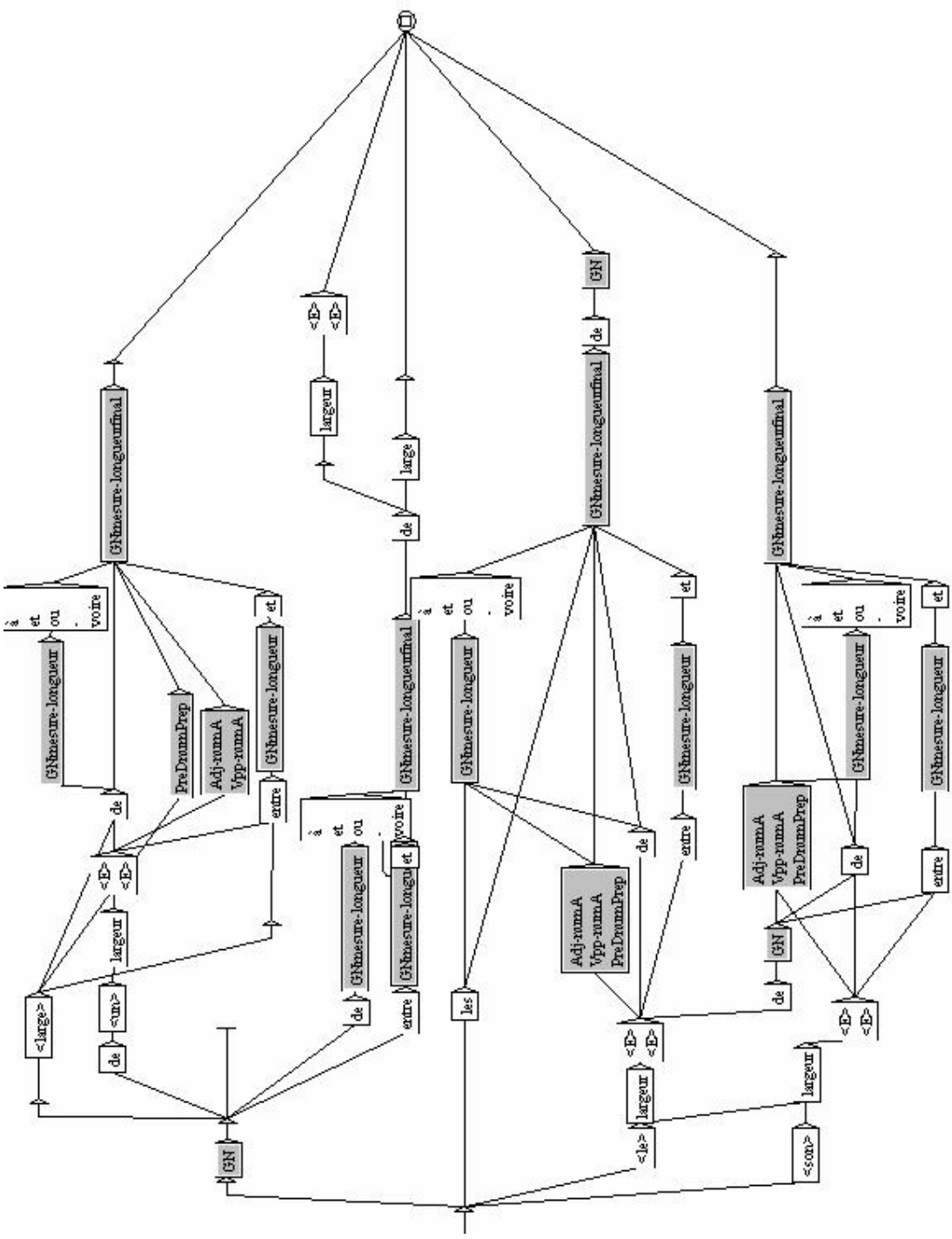

Figure 55 : graphe généré pour l'entrée *largeur*

3.4 Représentation des mesures relatives

3.4.1 Généralités

Dans cette section, nous regardons des structures mesurant une caractéristique ou propriété (*Ng*) d'un élément *N*0 par rapport à un autre (*N*1). Les schémas de phrase étudiés sont de la forme :

*N*0 *Vsup Prép un Ng de Dnum Unité Prép N*1
=: *Paris est à une distance de quelques milliers de kilomètres de New York*
=: Max est dans un rayon de 50 km autour de Reims
=: *Le stylo forme un angle de 90° avec la règle*

Nous regardons d'abord le comportement syntaxique des prédicats nominaux *Ng* rentrant dans cette structure. Puis, nous montrons que les expressions de pourcentage peuvent aussi être classées dans cette catégorie et nous examinons leurs propriétés distributionnelles. Enfin, nous étudions certaines expressions exprimant une comparaison relative de mesure.

3.4.2 Etude de la structure N0 Vsup Prép un Ng de Dnum Unité Prép N1

L'ensemble des prédicats *Ng* rentrant dans cette structure est restreint :

altitude, angle, distance, hauteur, périmètre, profondeur, rayon

Nous constatons qu'ils désignent tous des caractéristiques « géométriques ». Le nom *altitude* semble particulier car la phrase ci-dessous est un peu bancale :

?L'avion est à une altitude de 10 000 m au dessus du niveau de la mer

En effet, on ressent une certaine redondance avec l'utilisation de *au dessus de la mer* car cette séquence se trouve implicitement dans le nom *altitude* qui est la distance verticale entre le niveau de la mer et l'élément dont on veut mesurer l'altitude. Cependant, l'effacement de *altitude* force la présence de cette séquence.

*L'avion est à 10 000 m (?*E + au dessus du niveau de la mer)*

Comme dans la structure précédente exprimant une mesure « absolue », chaque *Ng* sélectionne un ensemble d'unités comme montré dans le tableau ci dessous.

altitude	GNmesure-longueur	*5 mètres*
angle	GNmesure-angle	*360 °*
distance	GNmesure-longueur	*trois kilomètres*
	GNmesure-temps	*douze secondes*
hauteur	GNmesure-longueur	*13 cm*
périmètre	GNmesure-longueur	*1 centimètre*
	GNmesure-surface	*78 m2*
profondeur	GNmesure-longueur	*dix mètres*
rayon	GNmesure-longueur	*5 km*

Table 5 : contrainte entre *Ng* et *Unité*

On constate l'apparition d'une nouvelle classe d'unités *GNmesure-angle* qui comporte les unités mesurant un angle tels que *radian, degré* (*rad*, °). Notons le cas particulier de *distance* qui sélectionne à la fois des unités de mesure de longueur (*Nmesure-longueur*) et des unités de mesure de temps (*Nmesure-temps*) :

Paul est à une distance de (?10 min + 10 km) de la maison

Cette forme est peu naturelle avec les unités de temps, mais l'emploi de ces unités le devient tout à fait lorsque l'on efface le prédicat *distance* :

Paul est à (10 min + 10 km) de la maison

Les noms *hauteur* et *profondeur* entrent aussi dans le schéma de phrase exprimant une mesure « absolue ». Cependant, l'emploi absolu et l'emploi relatif sont bien distincts :

Paul est à une hauteur de 10 m (E + au dessus du sol)
** Paul a une hauteur de 10 m*

*L'immeuble a une hauteur de 100 m (E + *au dessus du sol)*
**L'immeuble est à une hauteur de 100 m*

Dans le premier ensemble de phrases, le nom *hauteur* désigne la distance verticale entre Paul et le sol. Dans le deuxième ensemble de phrases, il désigne une caractéristique intrinsèque de l'immeuble du même type que *largeur* ou *longueur*.

Nous pouvons diviser cet ensemble de noms en trois. En effet, ils entrent dans trois structures bien distinctes :

(a) *N0 avoir un Ng de Dnum Unité avec N1*
=: *le crayon a un angle de 45° avec le livre*

(b) *N0 être à un Ng de Dnum Unité (de + Loc) N1*
=: *Le plongeur est à une profondeur de 10 m sous l'eau*
=: *Marie est à une distance de trois kilomètres de Paris*

(c) *N0 être dans un Ng de Dnum Unité autour de N1*
=: Les soldats sont dans un rayon de 100 km autour de la ville

Le seul nom prédicatif *Ng* rentrant dans le schéma de phrase (a) est *angle*. Les phrases en *être à* sont interdites et les verbes supports les plus naturels sont *faire* et *former*.

*Le livre (a + fait + forme + *est à) un angle de 45° avec le crayon*

Dans le cas général, la structure *N*0 *avoir un Ng avec N*1 est symétrique et se réduit à la forme nominale en *entre …et …*, comme dans l'exemple ci-dessous :

La France a une frontière avec la Belgique
La Belgique a une frontière avec la France
La France et la Belgique ont une frontière (?E + commune)
=> la frontière de la France avec la Belgique (réduction)

la frontière entre la France et la Belgique (réduction)

Dans notre cas, *angle* se comporte de la même manière, même si mathématiquement la symétrie n'est pas vérifiée car un angle est signé. En effet, on peut analyser (a) comme deux phrases :

Le livre a un angle de 45° avec le crayon
= Le livre a un certain angle avec le crayon ; cet angle est de 45°

Ainsi, on retrouve le cas général dans la première phrase que l'on peut réduire à *l'angle entre le livre et le crayon.* On obtient les phrases équivalentes suivantes :

Le livre a un angle de 45° avec le crayon
= L'angle du crayon avec le livre est de 45°
= L'angle entre le crayon et le livre est de 45°

L'emploi du verbe support *avoir* est également possible avec le nom *altitude* :

?L'avion a une altitude (E + de croisière) de 10 000 m (E + au dessus de la mer)

Cependant, *altitude* ne rentre pas dans la même structure de base que *angle* :

**Ce pic a une altitude de 3 290 m avec le niveau de la mer*

Nous examinons maintenant la structure (b) en *être à*. Les noms entrant dans ce schéma de phrase sont : *altitude*, *distance*, *hauteur*, *profondeur*. Ces quatre entrées ont toutes un comportement propre. Tout d'abord, seuls les noms *distance* et *hauteur* ont un comportement symétrique, bien qu'ils n'entrent pas dans le schéma de phrase (a). Cela n'est pas étonnant pour *distance* du fait de sa définition mathématique. Pour le nom *hauteur*, c'est moins net comme le montrent les phrases un peu bancales ci-dessous.

*Max (*a + est à) une distance de 10 m (de + *avec) Marie*
Max et Marie sont à une distance de 10 m (l'un de l'autre + E)
La distance entre Max et Marie est de 10 m

La hauteur entre l'avion et le toit de la maison est de 15 m
? l'avion et le toit de la maison sont à une hauteur de 15 m (l'un de l'autre + E)

Seul le nom *distance* peut être sujet à une transformation d'adjectivation. L'adjectif morphologiquement lié (*Ng- a*) à *distance* est *distant* :

Paul est distant de 15 m de Max
Paul et Max sont distants de 15m (E + l'un de l'autre)

Même des noms comme *profondeur* et *hauteur* possédant un *Ng- a* ne sont pas sujets à cette transformation :

** Paul est haut de 15 m du sol*
** Le plongeur est profond de 50 m sous l'eau*

Pour tous les noms rentrant dans (b), la séquence *Dnum Unité* peut être permutée avec *Ng* ; de même, la séquence *de N*1 est effaçable selon le contexte :

Max est à deux mètres de distance (E + du mur)
Marie est à 2 m de hauteur (E + du sol)
L'explorateur est à 200 mètres de profondeur (E + du niveau du sol)
L'avion est à 10 000 m d'altitude (E + ? au-dessus du niveau de la mer)

On observe couramment dans des textes la présence de formes réduites de la structure très théorique *N0 être à un Ng de Dnum Metre Loc N1* :

Marie est à 20 m sous l'eau
= Marie est à une profondeur de 20 m sous l'eau

Luc est (suspendu) à 2 m au-dessus du sol
= Luc est (suspendu) à une hauteur de 2 m au-dessus du sol

L'avion est à 10 000 m au-dessus du niveau de la mer
= ? L'avion est à une altitude de 10 000 m au-dessus du niveau de la mer

On peut retrouver le *Ng* effacé à partir de la préposition locative qui indique une direction. Par exemple, *sous* indique une direction verticale vers le bas et ainsi on déduit que l'on a une *profondeur*. Le *N*1 revêt une importance certaine. En effet, la préposition locative *au-dessus de* indique une direction verticale vers le haut, ce qui peut correspondre soit à une *hauteur*, soit à une *altitude*. Le nom *altitude* sélectionnant clairement un ensemble restreint d'expressions quasi-figées de la forme *Loc N1* comme *au-dessus du niveau de la mer*, il est facile de choisir entre les deux possibilités.
Cette analyse est faisable mais ne nous paraît pas très convaincante car elle est restreinte à un petit ensemble de prépositions. En effet, comment analyser la phrase suivante ?

La piste est à 10 km en aval de Val d'Isère

Nous décidons d'analyser la structure *N*0 *être à Dnum Metre Loc N*1 à l'aide des deux schémas de phrase suivants :

*N*0 *être à une distance de Dnum Metre de N*1
*N*0 *être Loc N*1 (Si *Loc* ≠ *de*)

Par ce moyen, on distingue clairement distance et direction : la première phrase indique la distance entre *N*0 et *N*1. La deuxième phrase donne la direction (optionnellement, le sens) ou une autre information géométrique pour retrouver (mathématiquement) la position de *N*0 par rapport à *N*1. Soit la phrase :

Bordeaux est à 550 km au sud-ouest de Paris

Elle s'analyse par :

Bordeaux est à une distance de 550 km de Paris ⇔ d(Bordeaux, Paris) = 550 km[45]

[45] *d*(*x*,*y*) désigne la distance entre le point *x* et le point *y*.

Bordeaux est au sud-ouest de Paris ⇔ direction = *sud-ouest*

Dans les phrases en *être à*, l'effacement de la préposition *à* est possible lorsque le *Ng* a déjà été effacé et que la préposition est autre que *de* :

Max est (E + à) 100 m sous terre
L'avion est (E + à) 10 000 m au dessus du niveau de la mer
Marie est (E + à) 100 km au nord de Marseille

*Marie est (*E + à) 100 km de Marseille*
*Max est (*E + à) 100 m du sol*
*Max est (*E + à) une distance de 200 m de la maison*

En anglais, on observe une structure de phrase équivalente à la différence près que l'on n'a pas de préposition après le verbe support :

N0 be Dnum Unit Loc N1
=: John is 30 miles (in the north of + from) London

On constate que la préposition locative *in the north of* peut être réduite à *north of* dans la phrase, ce qui rend l'expression plus compacte qu'en français :

John is 30 miles north of London
Jean est 50 km nord de Paris

Le nom *distance* accepte des modifieurs adverbiaux dans sa structure en *être à* tels que *à vol d'oiseau, à la ronde*. Son insertion dans la phrase conserve la symétrie, ce qui n'est pas le cas lorsque l'on a une préposition locative indiquant un sens et/ou une direction :

Paris est à 220 km à vol d'oiseau de Lille
Lille est à 220 km à vol d'oiseau de Paris

Lille est à 220 km au nord de Paris
** Paris est à 220 km au nord de Paris*
Paris est à 220 km au sud de Lille (équivalence sémantique et non syntaxique)

Nous constatons également une autre différence de comportement entre ces deux types de séquences si l'on analyse la phrase de base comme deux phrases élémentaires :

Paris est à une distance (de 200 km à vol d'oiseau + à vol d'oiseau de 10 km) de Lille
*Paris est à une distance de 200 km de Lille ; Cette distance est à vol d'oiseau (E + *de Lille)*

Paris est à une distance de 220 km au sud de Lille
Paris est à une distance de 220 km de Lille ; Paris est au sud de Lille

Le nom *distance*, lorsqu'il sélectionne des unités de temps, autorise l'insertion de modifieurs adverbiaux appropriés comme dans :

Le centre-ville est à dix minutes (en voiture + à pied + de marche) de Paris

Ils s'analysent à peu près de la même manière que *à vol d'oiseau*, même si l'on constate que la phrase en *être* est difficile. Pour rendre cette dernière phrase plus naturelle, il faut ajouter la forme passive du verbe *parcourir*.

Le centre-ville est à une distance de 10 minutes de Paris
Cette distance est (?E + parcourue) (en voiture+à pied)

Par ailleurs, l'analyse de *de marche* est impossible par ce moyen, ce qui n'est pas étonnant du fait que *10 minutes de marche* est équivalent à *une marche de 10 minutes*. \

**Cette distance est de marche*
Cette distance correspond à une marche d'une durée de 10 minutes

Les graphes **N0EtreADnumMetreLocN1** et **N0EtreADnumNtempsLocN1** décrivent ces structures avec le *Ng* effacé. Dans le premier, les unités sélectionnées appartiennent à la classe *GNmesure-longueur* ; dans le deuxième, elles appartiennent à *GNmesure-temps*. Les adverbes appropriés au nom *distance* sont représentés dans les graphes **Adv-app-distance1** et **Adv-app-distance2** : le premier concerne les unités de mesure de longueur et le deuxième concerne les unités de mesure de temps.

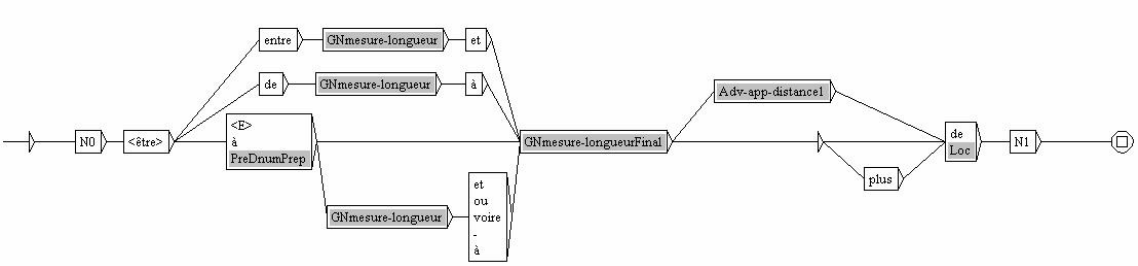

Figure 56 : N0EtreADnumMetreLocN1

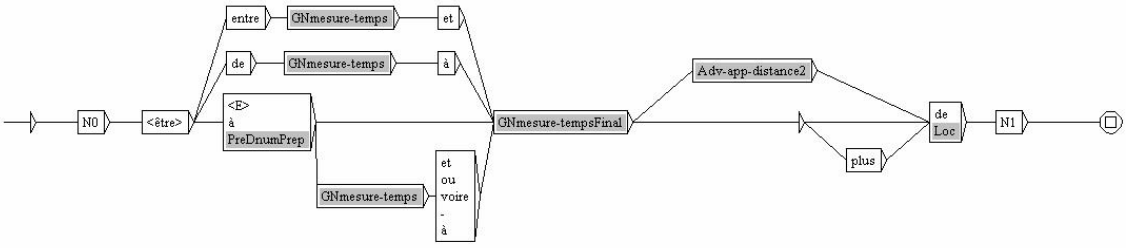

Figure 57 : N0EtreADnumNtempsLocN1

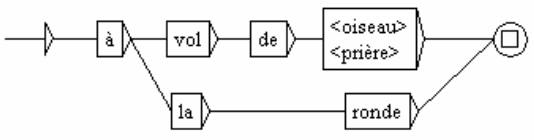

Figure 58 : Adv-app-distance1

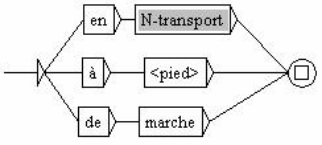

Figure 59 : Adv-app-distance2

La structure (c) diffère des deux autres par les prépositions (*Prep* =: *dans* et *prep1* =: *autour de*). Il existe deux noms rentrant dans cette structure : *périmètre* et *rayon*. Ils ont deux

comportements différents. Tout d'abord, *périmètre* sélectionne deux types d'unités : les unités de mesure de longueur (conformément à l'emploi mathématique) et les unités de mesure de surface (emploi courant). D'autre part, la phrase de base pour *rayon* est la réduction d'une phrase plus longue, ce qui n'est pas le cas pour *périmètre* :

Max est dans un rayon de 10 km autour de Lille
= Max est dans un cercle d'un rayon de 10 km autour de Lille

3.4.3 Codage des propriétés dans une table syntaxique

Bien que le nombre d'entrées lexicales soit peu élevé, le nombre de phrases à représenter dans les graphes est très important. Ainsi, nous décidons de coder les contraintes décrites précédemment dans une table syntaxique. Chaque ligne correspond à une entrée lexicale. La première colonne contient l'information indiquant si le sujet peut être un élément prédicatif. La deuxième colonne indique le verbe support employé alors que la troisième colonne donne la préposition suivant *Vsup*. La quatrième colonne comporte les entrées lexicales. La colonne 5 indique les classes d'unités sélectionnées par les entrées. Les colonnes 6 à 10 concernent la variation lexicale des prépositions *Prep1* : soit *avec*, soit *de*, soit *au-dessus de*, soit *autour de*, soit *Loc* qui désigne n'importe quelle préposition locative (simple ou composé, cf. Chapitre suivant). La onzième colonne concerne l'effacement de *Prep1 N1* dans la phrase de base. Le figement de la séquence *Prep1 N1* est codé dans la colonne 12. Le graphe **Prep1N1-altitude** représente l'ensemble des séquences figées *Prep1 N1* du nom *altitude*. Les propriétés de symétrie et de permutation sont respectivement données dans les colonnes 13 et 14. Les adverbes appropriés sont indiqués dans la colonne 15. Les colonnes 16 et 17 concernent les transformations d'adjectivation.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
N0Pred	Vsup	Prep	Ng	GNmesure	Prep1=: avec	Prep1=:de	Prep1=:à le-dessus de	Prep1=:autour de	Prep1=:Loc	N0 Vsup Prep un Ng de Dnum Unite	PrepN1-figé	permutation	symétrie	Adv-app	Ng-a	N0 être Ng-a de Dnum Unite Prep N1	N0 être dans un cercle de un Ng de Dnum Unité
-	<faire>	<E>	angle	:GNmesure-angle	+	-	-	-	-	-	-	-	+	-	-	-	-
+	<être>	à	distance	:GNmesure-longueur	-	+	-	-	-	+	-	+	+	:Adv-app-distance1	<distant>	+	-
+	<être>	à	distance	:GNmesure-temps	-	+	-	-	-	+	-	+	+	:Adv-app-distance2	<distant>	+	-
+	<être>	à	hauteur	:GNmesure-longueur	-	+	+	-	-	+	-	+	+	-	-	-	-
+	<être>	à	profondeur	:GNmesure-longueur	-	+	-	-	+	+	-	+	-	-	-	-	-
+	<être>	à	altitude	:GNmesure-longueur	-	-	-	-	-	+	:Prep1N1-altitude	+	-	-	-	-	-
+	<être>	dans	périmètre	:GNmesure-longueur	-	-	-	+	+	+	-	-	-	-	-	-	-
+	<être>	dans	périmètre	:GNmesure-surface	-	-	-	+	+	+	-	-	-	-	-	-	-
+	<être>	dans	rayon	:GNmesure-longueur	-	-	-	+	-	+	-	-	-	-	-	-	+

Table 6 : N0 Vsup Prep un Ng de Dnum Unité Prep1 N1

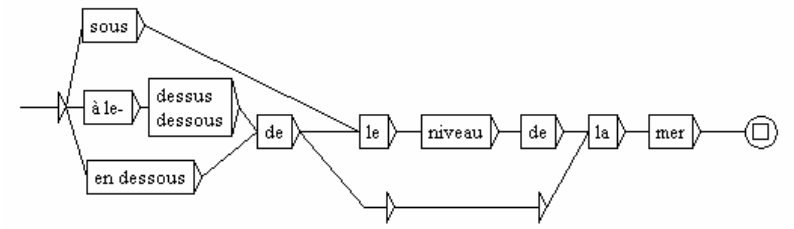

Figure 60 : Prep1 N1 figé pour le nom *altitude* (Prep1N1-altitude)

Nous présentons ci-dessous le graphe paramétré associé à la table ci-dessus, représentant l'ensemble des structures adverbiales dérivées de notre structure (ex : *à 10 m de profondeur*, *dans un rayon de trente kilomètres autour de Paris* ou *à une distance inférieure à 2m de la maison*). Juste en dessous nous donnons le graphe généré pour l'entrée *distance*.[46]

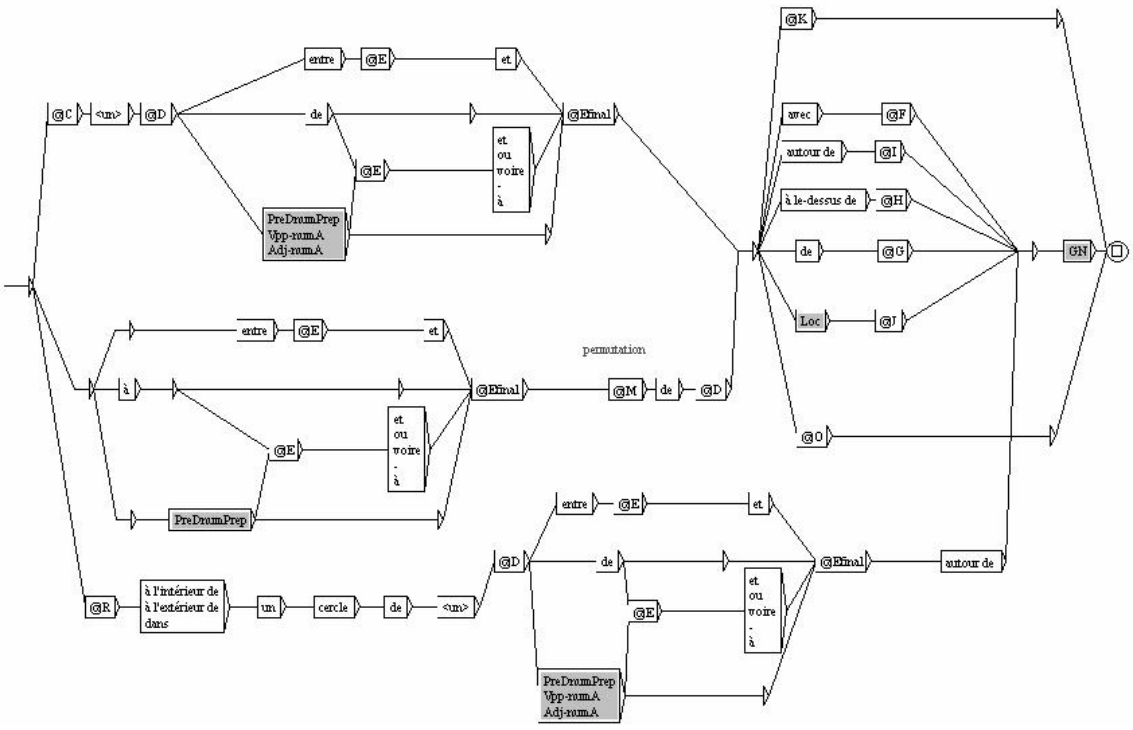

Figure 61 : graphe patron des structures adverbiales dérivées de *N0 Vsup Prep un Ng de Dnum Unite Prep1 N1*

[46] Dans le cas de l'adverbe, le graphe généré pour *angle* n'a pas lieu d'être. Nous le supprimons simplement à la main.

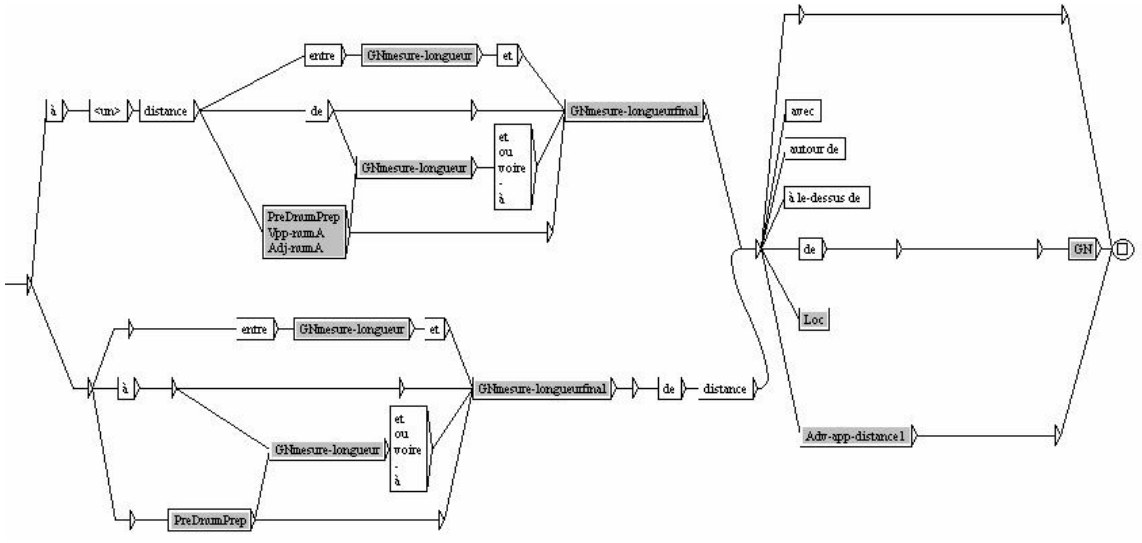

Figure 62 : graphe généré pour *distance*

3.4.4 Les expressions de pourcentage

Dans cette partie, nous nous concentrons sur les pourcentages. Nous montrons qu'ils entrent eux aussi dans le schéma de phrase *N0 Vsup Prép un Ng de Dnum Unité de N1* exprimant une mesure relative et étudions leur comportement syntaxique. Nous faisons une brève synthèse de notre étude réalisée en collaboration avec T. Nakamura (T. Nakamura et M. Constant, 2001). Cette étude montre que les expressions de pourcentages rentrent dans les structures suivantes :

N0 représenter Dnum % de N1
=: Les étudiants de Jussieu représentent 19% des étudiants parisiens

N0 comporter Dnum % de N1
=: Les étudiants parisiens comportent 19% d'étudiants de Jussieu

Ces structures sont en quelque sorte une forme réduite des structures théoriques suivantes contenant le prédicat *pourcentage*[47] :

N0 représenter un pourcentage de Dnum % de N1
=: Les étudiants de Jussieu représentent un pourcentage de 19% des étudiants parisiens

N0 comporter un pourcentage de Dnum % de N1
=: Les étudiants parisiens comportent un pourcentage de 19% d'étudiants de Jussieu

Ainsi, nous retombons bien sur notre structure de base représentant une mesure relative. Le prédicat *pourcentage* admet deux arguments (*N*0 et *N*1) et utilise les verbes supports (de pourcentage) *représenter* et *comporter*. L'unité sélectionnée est *%* (ou *pour cent* en toutes

[47] Le nom *proportion* semble également bien marcher :

Les étudiants de Jussieu représentent une proportion de 19% des étudiants parisiens

lettres). Etant donné que ces phrases sont théoriques, nous travaillons dorénavant sur les structures réduites. Ces phrases apparaissent comme les phrases élémentaires permettant d'analyser un pourcentage. Soit la phrase :

40% de (E + les) Français regardent la télé chaque soir

Cette phrase s'analyse comme suit à l'aide de deux phrases :

Des Français regardent la télévision chaque soir
(Les Français qui regardent la télévision chaque soir + Ces Français) représentent 40% des Français

Une analyse à l'aide du verbe *comporter* est également possible. La deuxième phrase deviendrait alors :

Les français comportent 40 % de personnes qui regardent la télévision chaque soir

Mais cette analyse n'est pas toujours valable. Cela dépend essentiellement de la nature sémantique (notamment le trait humain collectif) du nom tête du groupe nominal suivant la séquence *Dnum % de* :

40 % de la population regarde la télé chaque soir
**De la population regarde la télévision chaque soir*
**Cette population représente 40% de la population*

Dans ces cas-là, l'utilisation du déterminant nominal *une partie de* est préférable :

Une partie de la population regarde la télévision chaque soir
Cette partie de la population représente 40% de la population

Mais nous n'entrons pas dans la discussion. Nous souhaitons maintenant comparer les deux phrases de base des expressions de pourcentage :

Les produits laitiers représentent 80 % de notre production
Notre production comporte 80 % de produits laitiers

Ces deux phrases constituent une classe d'équivalence sémantique. En effet, ces deux phrases, qui ont exactement le même sens, ne diffèrent que par le verbe et les positions des arguments des verbes. Nous avons les équivalences suivantes :

*N*0*(représenter) = N*1*(comporter)*
*N*1*(représenter) = N*0*(comporter)*

Notons que ce schéma peut être étendu à d'autres phrases comme la phrase en *il y a* :

Il y a 80 % de produits laitiers (dans + parmi) notre production

Nous appelons complément d'inclusion d'une phrase de pourcentage l'argument *N*1 situé juste après la séquence *Dnum % de*. Le complément d'inclusion de la phrase avec *représenter* comprend obligatoirement un article défini, alors que celui du verbe *comporter* ne doit avoir

aucun article, ce qui peut correspondre à l'article indéfini par la règle de cacophonie. Ainsi, on a les quatre phrases suivantes (les deux acceptables sont équivalentes) :

Les étudiants de Jussieu représentent 19 % des étudiants parisiens
**Les étudiants de Jussieu représentent 19 % d'étudiants parisiens*
Les étudiants parisiens comportent 19% d'étudiants de Jussieu
**Les étudiants parisiens comportent 19% des étudiants de Jussieu*

Pour résumer, nous avons le schéma d'équivalence suivant :

N0 représenter Dnum % de LE N1
= N1 comporter Dnum % de ∅ N0

Notons que chacun des deux verbes est le verbe représentatif d'un ensemble de verbes ayant le même comportement syntaxique dans notre schéma de phrase : *représenter* pour l'ensemble {*représenter*, *constituer*, etc.} et *comporter* pour {*comporter*, *contenir*, *avoir*, etc.}. Pour plus de détails, se référer à T. Nakamura et M. Constant (2001).

Nous donnons ci-dessous les graphes représentant ces phrases. Les graphes **N0def, N1def** et **N0DetZ** représentent des groupes nominaux : les deux premiers comportent un déterminant défini et le dernier a un déterminant vide. **GNmesure-pourcentage** contient des unités de pourcentage (*pour cent* en toutes lettres et le symbole *%*).

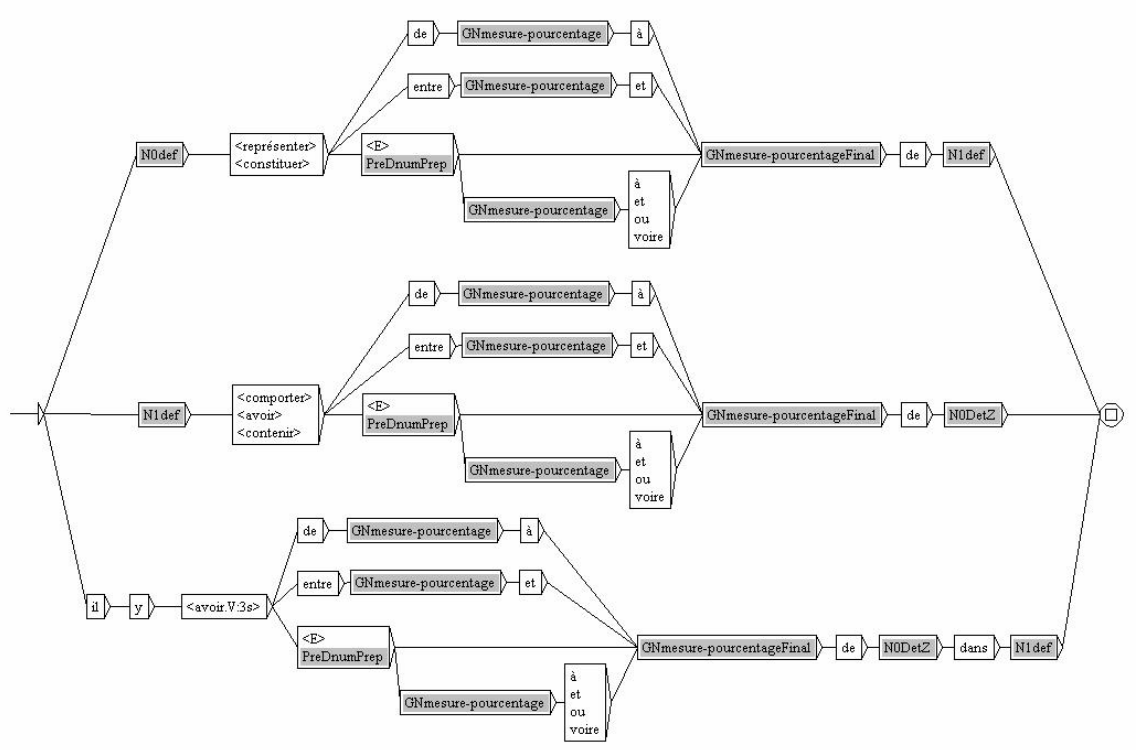

Figure 63 : N0 représenter Dnum% de N1

3.4.5 Comparaisons : quelques remarques sur les variations lexicales

Nous faisons maintenant quelques remarques sur la variation lexicale d'un nouveau type de phrases qui peuvent être considérées comme des mesures relatives. Prenons les deux phrases suivante :

Max a une taille de 178 cm
Luc a une taille de 174 cm

Il est possible d'interpréter ces phrases à l'aide d'une phrase comparative :

La taille de Max est 4 cm plus (élevée + grande) que celle de Luc
= La taille de Max dépasse de 4 cm celle de Luc

Nous nous intéressons à un autre type de structure mais qui est sémantiquement équivalente à celle-ci. La propriété mesurée n'y est plus explicite sous la forme d'un nom *Ng* (ex : *taille*) mais sous la forme implicite d'un adjectif comme dans la phrase ci-dessous équivalente à la phrase précédente :

Max est 4 cm plus grand que Luc

Ainsi, nous étudions la structure *N0 être Dnum Unité (plus+moins) Adj que N1*. Elle apparaît bien comme une mesure relative car elle met en jeu une mesure (*Dnum Unité*) et deux arguments (*N*0 et *N*1). Nous nous attachons surtout à montrer les contraintes lexicales. Nous partons des listes de noms prédicatifs *Ng* utilisées précédemment et nous regardons l'ensemble des adjectifs pouvant être utilisés dans chacun des cas. Chaque nom *Ng* sélectionne un ensemble d'adjectifs appropriés très restreint que nous pouvons facilement répertorier. Par exemple, pour le nom *poids* sélectionnant les unités de mesure de masse (*Nmesure-masse*) et pour le nom *longueur* sélectionnant les unités *Nmesure-longueur*, on a :

*Paul est 10 kg plus (léger + lourd + *grand + *chaud) que Luc*
*Ta barque est 10 m plus (petite + grande + longue + courte + *lourde + *chaude) que mon voilier*

Les noms morphologiquement dérivés de ces adjectifs (quand ils existent) ne rentrent pas, dans la plupart des cas, dans une phrase de la forme *N0 avoir un Ng de Dnum Unité* :

** Paul a une légèreté de 75 kg*
** Ta barque a une grandeur de 3 m*

Un même adjectif peut être sélectionné par plusieurs *Ng* comme *grand* sélectionné par *aire*, *hauteur*, *taille*, etc… Dans ces cas, c'est la nature sémantique des arguments *N*0 et *N*1 qui permet de lever l'ambiguïté. Nous n'entrons pas dans la discussion.

Si l'on regarde les *N*g du type *distance*, on constate le même phénomène de sélection d'adjectifs pour deux d'entre eux : *distance* et *hauteur*.

- *Ng =*: *distance*

*Paris est 300 km plus (loin + proche + *distant) de Bordeaux que de Tours*

- *Ng =*: *hauteur*

  Dans l'ascension du pic du Midi, Paul est 30 m plus (haut + bas) que Max

Dans la plupart des textes, on retrouve ces expressions sous la forme réduite d'un adverbe :

100 km plus (loin + haut), des soldats ont tiré sur des manifestants

Nous donnons dans les deux tableaux ci-dessous, les adjectifs *Adj* (lorsqu'ils existent) associés à chaque *Ng* entrant respectivement dans la structure *N0 avoir un Ng de Dnum Unité* et dans la structure *N0 Vsup Prep un Ng de Dnum Unité Prep N1* :

A	B	C
Ng	GNmesure	Adj
longueur	:GNmesure-longueur	<long>+<grand>+<petit>+<court>
profondeur	:GNmesure-longueur	<profond>
hauteur	:GNmesure-longueur	<haut>+<petit>+<bas>
largeur	:GNmesure-longueur	<large>
taille	:GNmesure-longueur	<grand>+<petit>
surface	:GNmesure-surface	<grand>+<petit>+<vaste>+<étendu>
vitesse	:GNmesure-vitesse	<vite>+<rapide>+<lent>
poids	:GNmesure-masse	<gros>+<maigre>+<lourd>+<léger>
épaisseur	:GNmesure-longueur	<épais>
température	:GNmesure-temperature	<chaud>+<froid>+<tiède>
énergie	:GNmesure-energie	<calorifique>+<énergétique>
coût	:GNmesure-monetaire	<coûteux>+<cher>+<économique>
durée	:GNmesure-temps	<rapide>+<lent>+<long>+<court>
volume	:GNmesure-volume	<volumineux>+<grand>+<petit>
puissance	:GNmesure-puissance	<puissant>+<faible>

Table 7 : entre *Ng*, *GNmesure* et *Adj* (absolu)

Ng	GNmesure	Adj
distance	:GNmesure-longueur	<loin>+<proche>+<près>
hauteur	:GNmesure-longueur	<haut>+<bas>

Table 8 : entre *Ng*, *GNmesure* et *Adj* (relatif)

Nous présentons ci-dessous le graphe patron représentant l'ensemble des expressions réduites dérivées de la structure comparative et le graphe généré pour *surface*.

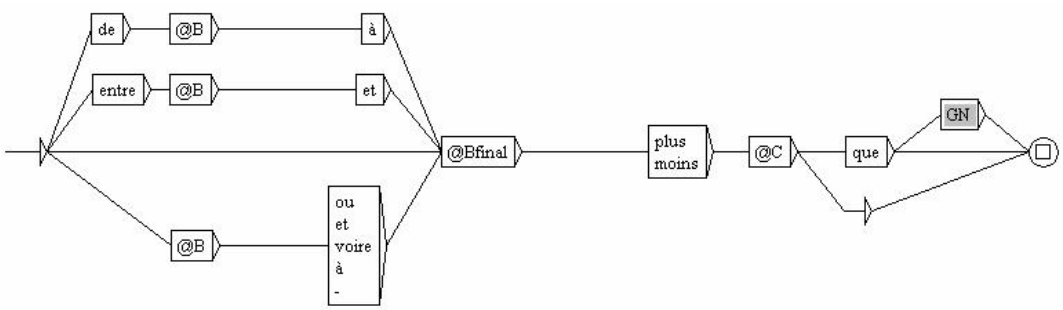

Figure 64 : graphe patron des mesures comparatives dérivées de *N0 Etre Dnum Unite plus Adj que N1*

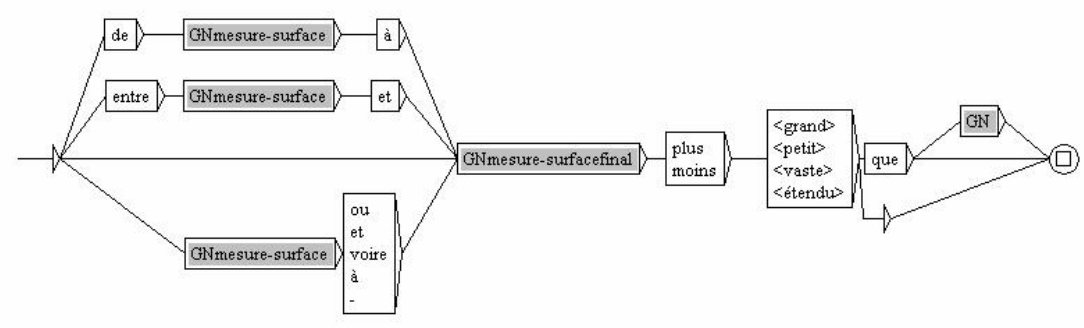

Figure 65 : graphe généré pour *surface*

3.5 Application à des textes

3.5.1 Généralités

L'étude linguistique que nous avons réalisée a pour but ultime de confronter à des textes les contraintes locales codées sous la forme de graphes. Nous disposons d'un ensemble de grammaires important. Pour chaque type de phrase étudiée, nous associons un ensemble de grammaires composé de :

- la grammaire décrivant la phrase de base et l'ensemble de ses transformées (souvent peu fréquentes dans les textes),
- la grammaire décrivant les réductions générées à partir de la phrase de base, c'est-à-dire des groupes nominaux,
- les grammaires représentant des expressions dérivées et partielles de la phrase (groupes nominaux, adverbes, prépositions, déterminants, modifieurs).

Ces grammaires sont construites en général à l'aide de graphes paramétrés et de la méthode d'E. Roche (1993). Nous synthétisons ci-dessous l'ensemble des grammaires construites:

- *Det Ng être de Dnum Unité*
    - phrase : *cette longueur est supérieure à 14 miles*
    - *GN* : *une longueur époustouflante de 130 m = 130 m de longueur*

- *N0 avoir un Ng de Dnum Unité*
    - phrase : *Le camion a un poids de dix tonnes*
    - *GN* : *Le camion d'un poids de dix tonnes*
    - déterminants composés : *10 tonnes de (plutonium)*
    - modifieurs : *haut de 120 m* (adjectival) ; *d'une hauteur de 120 m* (nominal)

- *N0 Vsup Prep un Ng de Dnum Unité Prep1 N1*
    - phrase : *Paul est à une distance de trente kilomètres de Paris*
    - GN : *la distance de trente kilomètres entre Paul et Paris*
    - adverbes : *à une distance de trente kilomètres de Paris*
    - prépositions : *à une distance de trente kilomètres de*

- *N0 être Dnum Unité Adj que N1*
    - phrase : *Paul est 10 kg plus lourd que Max*
    - modifieur : *10 kg plus lourd (que Max + E)*

Avant toute chose, il est nécessaire d'évaluer les grammaires obtenues et leur intérêt pour l'analyse automatique de textes. Nous illustrons ensuite comment elles peuvent être utilisées.

3.5.2 Evaluation des grammaires

Nous souhaitons évaluer nos grammaires en répondant à deux questions :

- La construction et la mise à jour des grammaires sont-elles faciles à mettre en oeuvre ? (production et maintenance)
- Les grammaires sont-elles à la fois complètes et précises ? (rappel et précision)

Ses questions sont légitimes pour pouvoir prétendre utiliser ces données pour l'analyse automatique des textes.

3.5.2.1 Production et maintenance

Nous avons montré dans ce chapitre le processus complet de construction de grammaires d'expressions de mesure. A première vue, la production de telles grammaires n'est pas difficile. En effet, les formalismes utilisés sont extrêmement simples et visuels (tables et automates) et donc facilement compréhensibles pour les linguistes. Malgré cette simplicité apparente, notre méthode basée sur M. Gross (1975) permet de décrire systématiquement des phénomènes très précis. La grosse difficulté réside dans la quantité astronomique de données à accumuler. Ainsi, notre processus requiert une extrême rigueur dans l'organisation des données qui peuvent rapidement devenir illisibles et donc incompréhensibles.

Nous avons vu que nous disposions de deux méthodes de production des grammaires à partir d'une analyse linguistique détaillée :

- par construction manuelle (ex : les graphes **DnumUniteDe** et **ADnumMetreDeN1**)
- au moyen d'une représentation intermédiaire (tables syntaxiques) et d'un mécanisme semi-automatique de conversion en graphes

Ces deux méthodes se mélangent : les graphes patrons utilisent des sous-graphes faits entièrement à la main comme les déterminants numériques et les unités. Il n'existe pas de critères clairs pour choisir l'une ou l'autre. Il faut simplement prendre la méthode la moins contraignante et la plus flexible. La méthode par tables syntaxiques est extrêmement intéressante si la description des expressions nécessite de nombreuses duplications de morceaux de graphes. Les séquences relativement simples (type mots composés) sont très naturellement décrites à la main dans des graphes : par exemple, les déterminants numériques. La construction des graphes *GNmesure*, bien que simple, nécessite la duplication systématique de morceaux de graphes pour chaque classe d'unités. Il est donc préférable d'automatiser ces opérations à l'aide de la méthode d'E. Roche. Pour l'ensemble des unités de mesure, nous aurions pu automatiser une grande partie des opérations de construction car chaque unité a le même ensemble de préfixes (*milli-*, *centi-*, *déci-*, *déca-*, *hecto-*, *kilo-*, etc.). Cependant, les duplications, dans ce cas-là, ne sont pas très contraignantes. Pour les structures de phrase qui sélectionnent un ensemble de prédicats, il est très souvent préférable de construire des tables syntaxiques. En effet, chaque prédicat rentre dans un certain nombre de structures appartenant à un sur-ensemble commun. Cependant, leur comportement diffère dans le détail, ce qui est très difficile à coder manuellement sous la forme de graphes. Il est plus facile de coder les informations syntaxiques dans une table à l'aide de valeurs booléennes. Par contre, si le nombre d'entrées est très réduit, il est peut-être préférable de le faire directement sous la forme de graphes car la gestion des variables dans les graphes patrons n'est pas toujours facile. Notons que les structures *N*0 *Vsup Prep un Ng de Dnum Unité Prep*1 *N*1 sont codées dans une table bien que le nombre d'entrées soit faible, du fait de la répétitivité des duplications.

Le gros désavantage de ces deux méthodes est que le codage des comportements syntaxiques est manuel, donc très coûteux en temps. Cependant, cette approche est nécessaire pour décrire

des phénomènes très précis car chaque entrée lexicale a un comportement propre qui ne peut être prédit automatiquement. Cela n'empêche pas l'utilisation de certains processus automatiques qui atténuent cet inconvénient. Ils permettent d'extraire rapidement des informations linguistiques de grands corpus. Par exemple, il est intéressant de chercher tous les noms prédicatifs *Ng* dans les textes afin d'examiner les contextes droits et gauches et, ainsi, de compléter nos grammaires par les séquences trouvées dans le texte mais manquantes dans les grammaires (exemple : *5 mètres en largeur = une largeur de 5 mètres*). Une fois les unités simples décrites à l'aide de dictionnaires de manière quasi-exhaustive, il est possible de les chercher dans les textes et ainsi trouver des améliorations à la description des déterminants et prédéterminants numériques[48]. Au cours de telles opérations de maintenance, l'utilisation de tables syntaxiques comme représentation intermédiaire permet d'éviter certaines duplications de graphes à la main comme nous l'avons montré ci-dessus. L'application des graphes *GNmesure* permet de compléter la liste des *Ng* (exemple : *âge*, etc.).
D'autre part, notre formalisme n'impose pas de coût supplémentaire en temps lors de la mise à jour des données. En effet, dans les graphes, l'insertion d'une nouvelle séquence linguistique est une opération très simple : l'ajout de nouvelles transitions dans le graphe. La modification des tables est également très simple et très économique. Par exemple, l'ajout d'une nouvelle propriété ne requiert que l'ajout d'une colonne dans la table, la modification du graphe patron et la génération automatique des graphes associés à chaque entrée lexicale.

3.5.2.2 Bruit et silence

Traditionnellement, pour évaluer l'efficacité des grammaires, on effectue des évaluations quantitatives en les appliquant[49] sur un corpus de taille moyenne que l'on décortique ensuite manuellement. Deux critères d'évaluation sont alors calculés : le silence et le bruit. Le silence est la quantité d'expressions pertinentes non trouvées par la grammaire. Le bruit est la quantité d'occurences reconnues par la grammaire mais qui sont incorrectes. En général, ces deux critères sont donnés sous la forme de pourcentages de ces quantités par rapport au nombre d'occurences correctes trouvées manuellement.
Avant de réaliser ce type d'évaluation, nous faisons quelques remarques. Tout d'abord, d'un point de vue général, notre approche purement linguistique se distingue de la plupart des méthodes utilisées en TAL qui, dans la majorité des cas, utilisent des approches statistiques passant par des phases d'apprentissage sur des corpus. Par définition, les résultats sont des approximations et donnent invariablement des erreurs. Le but est d'évaluer quantitativement le modèle statistique utilisé. Par notre méthode, nous décrivons tous les cas possibles et non pas seulement ceux qui apparaissent dans les corpus. Ainsi, les expressions retrouvées dans les corpus ne représentent qu'une infime proportion des expressions réellement représentées dans les grammaires. Ainsi, une évaluation ne porte que sur une toute petite partie de la grammaire. Cependant, ces évaluations permettent de compléter nos descriptions au fur et à mesure au moyen des expressions non trouvées car une grammaire est toujours en évolution. Une évaluation n'est donc valable qu'à un instant *t*. Dans notre étude spécifique, nous constatons que les expressions de mesure dans les corpus journalistiques généraux ou dans les corpus scientifiques de vulgarisation, n'apparaissent que très rarement et leur fréquence selon les types d'unités est très hétérogène. En effet, alors que le mot *kilomètre(s)* apparaît 2 645 fois dans une année du Monde (1994) (environ 100 millions de mots), il n'existe que 19 séquences appartenant aux classes **Volt** et **Volt_abr**. Les unités *Newton* et *électron-volt* n'apparaissent même pas ; les unités mesurant la température n'apparaissent quant à elles pas

[48] L'ambiguïté des symboles des unités nécessite un filtrage manuel important (ex : *a* pour *are* ou *s* pour *seconde*).
[49] Nous appliquons nos grammaires sur un texte prétraité à l'aide du logiciel Unitex avec la règle du 'longest-match'.

plus d'une centaine de fois. Les unités que l'on retrouvent les plus fréquemment sont les unités de mesure de temps (ex : *mois*), de longueur (ex : *mètre*) et de masse (ex : *kilogramme*) et les monnaies (ex : *dollar*). Une évaluation globale est donc difficile. Il faut regarder chaque unité séparément. Cependant, il est clair que les classes d'unités n'apparaissant que quelques fois ne peuvent être évaluées de manière représentative comme pour la famille de *volt*. Nous décidons de nous consacrer aux expressions contenant des unités de longueur et de masse. Les expressions temporelles sont trop ambiguës et très complexes pour pouvoir être correctement traitées par nos grammaires et elles méritent des études approfondies (D. Maurel, 1990 ; M. Gross, 2002). Par ailleurs, nous avons utilisé des groupes nominaux libres extrêmement simples se résumant à l'expression rationnelle *<DET> (<E>+<A>) <N> (<E>+<A>)* car autrement cela amène trop d'erreurs du fait de l'ambiguïté naturelle de la langue. La description de ce type d'expressions étant fondamentale et difficile, nous décidons de ne pas la faire par manque de temps matériel. Ainsi, nous avons des expressions de mesure souvent « tronquées » mais contenant beaucoup d'information : des groupes nominaux lexicalisés (*une altitude de 10 000 m = 10 000 m d'altitude*) ; des modifieurs (*d'une longueur de dix mètres* ; *supérieurs à 30 kg*) ; des prépositions locatives (*à 1 km au nord de*) et des déterminants composés (*trente litres de*). Les unités de mesure que nous évaluons ne sont pas extrêmement fréquentes : il existe seulement 1 240 unités de mesure de longueur sur les vingt premiers millions de mots du corpus. Pour les unités massiques, c'est encore plus difficile à évaluer du fait de l'ambiguïté de certains symboles avec des mots extrêmement fréquents dans la langue, comme *t* ambigu avec le *t* de *Max a-t-il mangé ?; livre* est aussi ambigu avec la monnaie *livre sterling* et l'objet que l'on lit. Si l'on applique la grammaire décrivant les unités massiques, seules 142 occurrences sur 2 000 appartiennent à des expressions de mesure de masse (7%), ce qui n'est pas suffisant pour faire un calcul représentatif. Pour les expressions mettant en jeu des unités de mesure de longueur, notre démarche a été la suivante : nous avons appliqué en même temps nos grammaires représentant des expressions de mesure de longueur avec celles des unités de mesure de longueur. Nous avons arrêté l'application après 2 000 occurrences trouvées. Après examen des occurrences, nous n'avons gardé que 1 240 d'entre elles car certaines unités sont ambiguës comme *m* ambigu avec le *m* de *Max m'a dit que ça allait* et d'autres appartiennent à d'autres types de mesure telles que les mesures de surface ou de volume (*un volume de 10 m3*). Après examen des 1 240 occurrences pertinentes, nous obtenons les résultats suivants :

Silence pur	*Reconnaissance partielle*	*Bruit pur*
7	56	9

La colonne « silence pur » indique le nombre d'expressions de mesure de longueur qui n'ont pas été trouvées automatiquement (même partiellement). La colonne « reconnaissance partielle » donne le nombre de séquences qui n'ont été reconnues que partiellement. La colonne « bruit pur » indique le nombre d'occurrences qui n'auraient pas dû être reconnue du tout. Nous calculons le taux de silence de deux manières : soit sans tenir compte des séquences reconnues partiellement ; soit en en tenant compte. On agit de la même manière pour le taux de bruit.

Taux de silence = 0.6% (5,1%)[50]
Taux de bruit = 0.7% (5,3%)

[50] Le premier résultat est calculé en considérant les occurrences partielles comme correctes alors que le résultat entre parenthèses est calculé en les considérant comme incorrectes.

Pour la plupart des séquences qui n'apparaissent pas, cela est dû à de simples oublis dans la construction des graphes, comme *soixante* qui, malencontreusement, a été oublié dans le graphe des déterminants numériques écrits en lettres.

e à Istrana, une base située à soixante kilomètres à le nord-est de Vicence (Italie), de

La présence d'un adjectif entre le déterminant numérique et l'unité est également une source de silence comme :

toute sa longueur actuelle un petit kilomètre_, que vingt-huit ans après sa naissance

Les reconnaissances partielles sont d'abord dues à des expressions auxquelles nous n'avions pas pensé dans un premier temps comme

J'enfonce dans le terrain deux pieux à vingt et un pieds environ l'un de l'autre.{S} Je

Certaines expressions numériques peu courantes nécessitent une conversion en une mesure avec une unité plus courante pour qu'elle puisse être compréhensible pour le lecteur. Cette conversion peut être insérée en plein milieu de l'expression :

à 5 milles nautiques (environ 9 kilomètres) à le sud de Brest

Plusieurs occurrences partiellement reconnues sont les conséquences d'oublis dans la description des prépositions locatives *Loc* (cf. les deux premières phrases ci-dessous) ou d'erreurs de codage dans les tables (cf. la dernière phrase) :

heures et son rayon de action à 40 km autour de l'hôtel de Manhattan où se tenait la
enfin l'aménagement, à 15 mètres sous terre, de la station Tolbiac-Masséna de
le grand circuit, ce est-à-dire 40 kilomètres de périmètre de temples entretenus.{S}

Notons que nous n'avons pas essayé de reconnaître les adverbes lorsque la séquence *Loc N1* a été transformée en adverbe. Il est prévu d'ajouter ce type de données ultérieurement.

passe la piste qui mène à Tadjourah, 12 kilomètres plus à l'ouest.{S} Les

Les erreurs proviennent parfois du corpus lui-même qui contient des fautes d'orthographes :

aurait coulé dans le lac par 160 m de fonds le 24 janvier dernier après un amerrissa

Dans une revue scientifique telle que *Science et Vie*, la distribution entre les unités est plus homogène que dans *Le Monde*, même si les expressions temporelles sont toujours largement majoritaires. Nous décidons de réaliser une évaluation quantitative globale sur ce type de corpus. Nous avons assemblé un ensemble d'articles de *Science et Vie* datant de l'année 1992 pour former notre corpus (environ 100 000 mots). Nous avons dû faire quelques modifications dans nos grammaires : par exemple, nous avons supprimé la règle du blanc intercalé tous les trois chiffres dans les *DnumEnChiffres* car elle n'était respectée que partiellement par les journalistes.

Nombre total d'occurrences	334
Silence pur	4
Bruit pur	21
Reconnaissance partielle	12

Taux de silence : 1,2% (4,8%)
Taux de bruit : 6,3 % (9,9%)

Les 4 expressions non reconnues (silence) sont dues à trois facteurs différents :

- Des fautes typographiques sont présentes dans le corpus

approchant celle de la lumière:{S} 300 0000 km/s. {S}La masse est une forme de

- Un cas de déterminant nominal n'a pas été répertorié (*le dixième de*) :

dépassant à peine le dixième de mm en longueur pour une épaisseur de 4 ðm. {S}Avec

- certaines unités dans nos grammaires ne se trouvent pas dans le dictionnaire électronique (ex : *méga-électrons-volts*)

Certaines expressions reconnues sont du bruit pur du fait de :

- l'ambiguïté naturelle de la langue et d'une analyse trop locale

un indice de cétane amélioré.{S} De 50 à 52 à l'heure actuelle, ils ne désespèrent pas
s de la peau).{S} L'activation de le VIH-1 a ainsi été reproduite chez un animal entier,
la GT 31 s 'apparente

- certaines expressions n'ont pas été répertoriées

à 10 m près
de la couche de ozone constatée entre 30ø et 64ø de latitude nord entre 1969 et 1986 ne

Les résultats obtenus sont tout à fait satisfaisants. Cependant, cette évaluation n'est valable que pour un instant *t* et un corpus bien précis. Nos grammaires sont mises à jour en permanence : les oublis et les petites erreurs de codage disparaissent au fur et à mesure. Ainsi, nos taux de silence tendent de plus en plus vers 0%. le bruit n'est dû qu'à l'ambiguïté naturelle de la langue et ne peut être supprimé qu'au prix d'une analyse plus contextuelle.

3.5.3 Opérations utilisant les grammaires

Certaines opérations que nous proposons ci-dessous sont pour l'instant théoriques car elles nécessitent une grammaire complète et précise des groupes nominaux.

3.5.3.1 Localisation de constituants syntaxiques

La seule opération directement utilisable à ce jour est la localisation automatique de constituants syntaxiques tels que :

- des groupes nominaux :

  une longueur de 10 m
  10 m de long

- des groupes adjectivaux :

  âgé de 10 ans
  distant de 10 m

- des groupes prépositionnels (modifieurs)

  (un camion) *d'un poids de dix tonnes*

- des prépositions locatives composées :

  deux cents mètres en amont de (la station)

- des adverbes :

  à 10 m de hauteur

- des déterminants nominaux :

  trois tonnes de (pétrole)

Pour réaliser une telle opération, il faut ajouter des informations de sortie aux graphes. Dans le cas où nous gardons le format du DELAF et du DELACF[51], nous obtenons le graphe ci-dessous pour les déterminants nominaux de mesure. Si l'on applique ce graphe à la phrase *Max a mangé 200 grammes de frites*, la séquence *200 grammes de ,.DET+Dnom+Mesure* est générée en mode fusion[52] et peut donc être rajoutée au dictionnaire du texte car elle est compatible avec le format.

[51] Rappel : le DELAF et le DELACF sont les dictionnaires électroniques que nous utilisons.
[52] Dans le mode fusion, les informations de sortie sont insérées dans la séquence reconnue par le graphe.

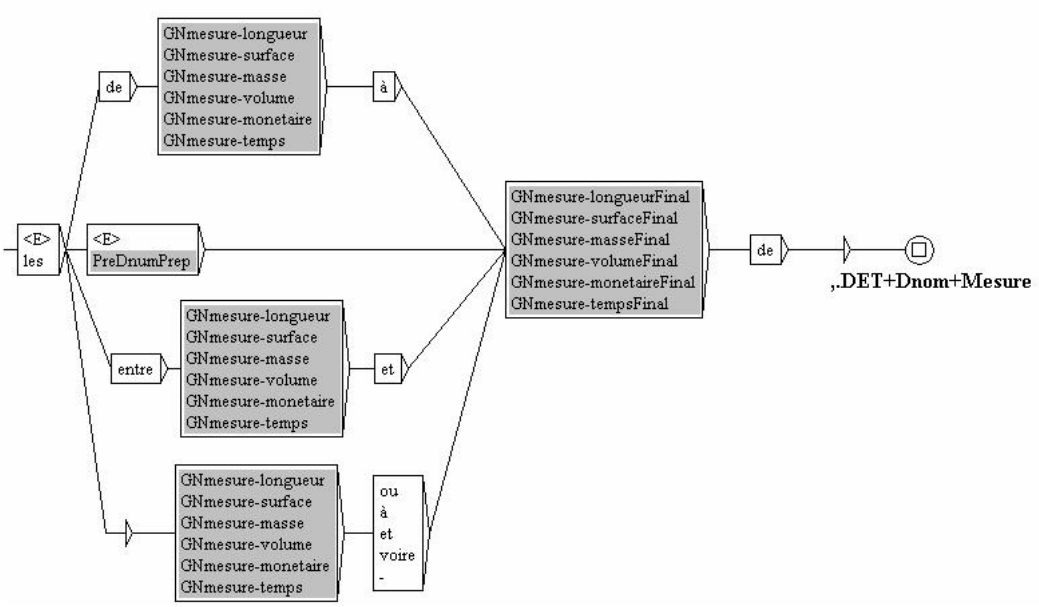

Figure 66 : localisation de déterminants nominaux de mesure

3.5.3.2 Analyse transformationnelle

A plus long terme, nous pourrions utiliser les techniques d'E. Roche pour une analyse transformationnelle d'expressions de mesure à l'aide de transducteurs à variables. En effet, le groupe nominal *une salle à 10°C* peut s'analyser par la séquence suivante :

une salle à 10°C, une salle avoir une température de 10°C.GN+mesure

Cette séquence signifie que l'expression reconnue *une salle à 10 °C* est un groupe nominal qui est le résultat de la réduction de la phrase élémentaire : *la salle a une température de 10°C*. Le graphe utilisé pourrait être le suivant :

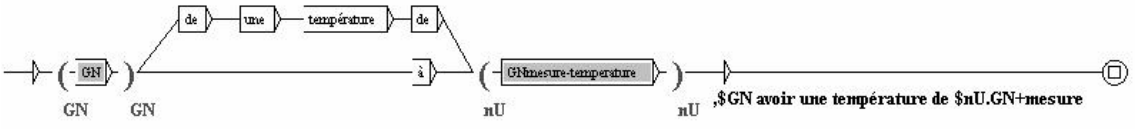

Figure 67 : analyse transformationnelle de groupes nominaux de mesure

Les séquences *$GN* et *$nU* sont des variables. Lors de l'application de ce graphe à la séquence ci-dessus : chaque partie reconnue par les chemins entre parenthèses du graphe est stockée dans la variable associée. Ainsi, *une salle* est placé dans *$GN* et *10°C* est placé dans *$nU*. En sortie de l'application de ce graphe, les variables sont remplacées par leur contenu.

3.5.3.3 Extraction et normalisation d'information

A l'aide des grammaires construites, nous pouvons réaliser des normalisations. En effet, il a été montré à plusieurs reprises (A. Chrobot, 2000 ; L. Karttunen, 2003) que les déterminants numériques cardinaux en toutes lettres pouvaient très facilement être normalisés sous la forme de nombres écrits en chiffres à l'aide de transducteurs, facilitant ainsi le travail de traduction de ces séquences :

français ↔ formel ↔ anglais

dix-sept ↔ 17 ↔ *seventeen*

Cependant, nous avons vu que la syntaxe des nombres en chiffres pouvait dépendre de la langue :

Deux mille trois cent douze ↔ 2 312
Two thousand three hundred and twelve ↔ 2,312

Mais, cela n'est pas un problème si l'on utilise une représentation numérique indépendante de la langue.

Nous pourrions étendre cette application aux mesures. En effet, il existe un symbole standard international, pour chaque unité écrite en lettres. Ainsi, on a :

français ↔ formel ↔ anglais
mètre ↔ *m* ↔ *meter*

Le formalisme du transducteur à états finis est parfaitement adapté. Il suffit donc de reprendre nos grammaires d'unités et leur ajouter une sortie comme ci-dessous avec le graphe **Metre** :

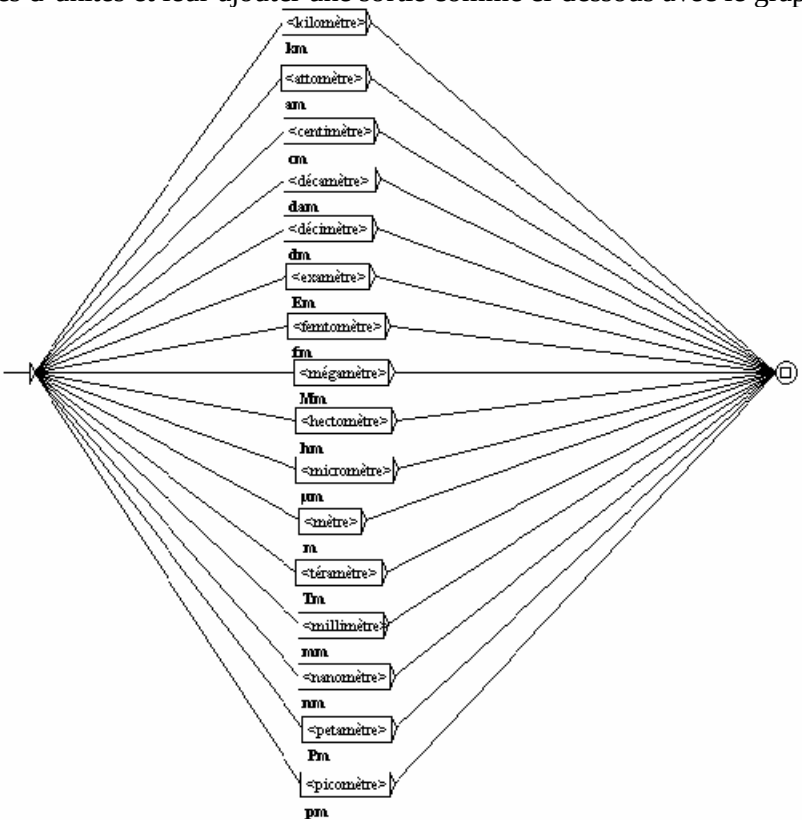

Figure 68 : normalisation du graphe Metre

Nous pourrions même aller plus loin dans la normalisation en convertissant chaque unité (ex : *kilomètre*) en son unité de base (ex : *mètre*). On aurait une conversion comme suit :

kilomètre ↔ *1 000 m*

Il est alors très facile de combiner les normalisations des déterminants numériques et des unités pour normaliser les phrases de mesure (ou leurs réductions) de la manière suivante :

cette salle à 17°C ↔ T(*cette salle*) = 17°C[53]
Paul est à 18 km de Paris ↔ d(*Paul,Paris*) = 18 km[54]

Les transducteurs à variables sont extrêmement efficaces comme le montre le graphe théorique ci-dessous. Lors de l'application de ce graphe à *Paul est à 18 km de Paris*, la séquence reconnue par *N*0 (*Paul*) est stockée dans la variable *$0*, *18km* est stockée dans *$nU* et *Paris* est stockée dans *$1*. Ainsi, en sortie, on obtient en remplaçant les variables par leur contenu : d("*Paul*", "*Paris*") = 18km.

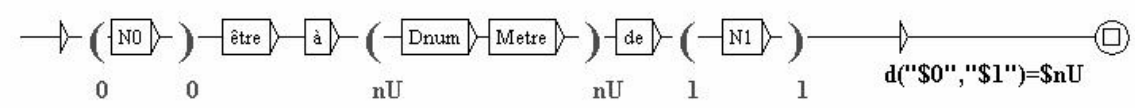

Figure 69 : normalisation

Cependant, des variantes lexicales introduisent des modifications sémantiques. C'est le cas pour :

la longueur de la corde est inférieure à dix mètres ↔ longueur(*la corde*) < 10 m
une longueur de corde d'à peu près dix mètres ↔ longueur(*la corde*) ≅ 10 m

Par ailleurs, il existe des expressions de mesure sous la forme d'intervalles comme nous l'avons montré précédemment avec les structures *entre … et …* et *de … à …*. Là aussi, il y a moyen d'utiliser une normalisation mathématique :

Le spectacle a une durée comprise entre 45 min et 1 h ↔ durée(spectacle) = [45 min, 1 h]

Nous n'entrons pas dans les détails, mais il est clair que cette application mériterait une étude complète.

3.6 conclusion

Nous avons procédé à la description systématique de certaines expressions de mesure sous la forme de grammaires locales. Par cette étude, notre but principal était d'exposer le processus complet de représentation d'un phénomène linguistique relativement simple mais peu étudié jusque-là, en mélangeant deux méthodes différentes mais complémentaires : soit une construction directe en graphes, soit une construction par l'intermédiaire de tables syntaxiques. Nous avons également montré l'intérêt d'une telle étude pour le TAL. Clairement, nous n'avons pas examiné tous les types d'expressions de mesure, notamment les phrases décrivant une évolution où existent des contraintes lexicales fortes comme dans :

Paul a augmenté son poids de 10 kg par rapport à il y a deux mois
Paul a pris 10 kg
*Paul a (grossi + maigri + *grandi) de 10 kg*

[53] T est le symbole de température.
[54] d est le symbole de distance.

Par ailleurs, nous ne sommes pas allé très loin dans l'examen des expressions figées ou semi-figées contenant une mesure. Nous nous sommes simplement contenté du cas *vent de 45 nœuds*.