Microsyntax of Measurement Phrases in French: Construction and Evaluation of a local grammar

Matthieu Constant

Université Paris-Est, France

Abstract. This paper focuses on the description of simple measure phrases by means of a local grammar. We show that these linguistic objects can be decomposed in several elementary components and that they are subject to complex internal constraints. From these linguistic observations, we build a global grammar that is then evaluated on two corpora.

1 Introduction

Finite-state technology is now widely used in Named Entity Recognition (NER) systems. Many studies showed that finite-state representation was linguistically relevant for such entities, in particular numerical expressions (NUMEX). In this paper, we focus on specific numerical expressions related to 'quantities': measurement expressions, that are less studied because of their low frequency relatively to other types of numerical expressions.

A Measurement expression involves a value measuring a property of an item (e.g. *length*, *weight*) or a property relating two items (e.g. *distance*) as it is shown in the examples below:

(1) On this road, the **speed** of the **car** is limited to **50 mph**.

(2) **The villa** has a **distance** of **100 feet** to **the sea**.

In the construction-grammar framework, they are formalized by a relation between three elements: item, dimension and value. An item (e.g. *desk*) has a dimension (e.g. *width*, *weight*) which can be measured with a value (*2 m*, *2 kg*)[Hasegawa et al. 2008]. In addition, each dimension selects a set of appropriate units. Dimensions can be regrouped into families the dimensions of which select the same units: e.g. *width*, *height*, *radius* dimensions belong to the 1D spatial dimension family.

In this paper, we focus on the microsyntax and the recognition of the value in French measurement expressions, which, roughly speaking, consists of a simple phrase composed of a cardinal numerical determiner and a unit. We note them *Dnum Unit*. This is an obligatory step before a deeper automatic syntactic and semantic analysis. At first sight, this appears simple but we will show that this type of expressions is subject to complex lexical and syntactic contraints that can be well described and identified by a finite-state system. Our study is based on our previous work in [Constant 2002]. We use the formalism of local grammars introduced by [Silberztein 2000] and [Gross 1997]. They are made of finite-state graphs representing and structuring complex linguistic

forms. They can recursively refer to other graphs and are based on large-scaled lexicons. The grammars were encoded, compiled and applied with different programs of the platform Unitex [Paumier 2008][1].

The first part of the paper is devoted to a linguistic description of the elementary components that are numerical determiners and scientific units. Then, we detail the linguistic constraints on the combination of these two components. Finally, we show how to construct the global local grammar recognizing a regular language and we evaluate it on three distinct corpora.

2 Cardinal numerical determiners

Numerical determiners are well studied linguistic components in the field of NLP. Many researchers showed that Finite state models are best suited for their representation (in particular [Chrobot 2000,Silberztein 2003,Karttunen 2006,Laporte 2007]). They can be divided into three categories: cardinal numbers written in words, cardinal numbers constiting of digits, nominal numerical determiners.

2.1 Cardinal numbers written in words

Cardinal numbers written in words have been studied and described in the form of local grammars by [Silberztein 2003] for French. For this work, we used the proposed graph structuration to construct a local grammar, describing natural integers in words up to one trillion. The main difficulty comes from different spellings of some words depending on the context. For instance, *cent* (hundred) is spelled differently in *deux cents* (two hundred) and in *deux cent deux* (two hundred and two). The number *quatre-vingt* has two spellings whether it is located on the right or on the left of words like *mille* (thousand) or *million* : ***quatre-vingt*** *mille deux cent* ***quatre-vingts*** (i.e. 80, 280)

2.2 Cardinal numbers consisting of digits

A finite-state graph is also very convenient to describe cardinal numbers consisting of digits. There exits a writing convention, and so a specific syntax, for numbers with more than three digits: for example, in French, in *1 298*, there exists a white space between the third and the fourth digits; in English, the white space is replaced by a coma (1,298). From a general point, a white space (a comma in English) occurs in the sequence of digits every 3 digits starting from the right. Note that such strict conventional constraints are not always true in texts depending on the writer. Sometimes, required internal whitespaces do not even exist or are replaced by dots (.). Integer values are therefore sometimes ambiguous with some dates (ex. *2009*).

Decimal numbers can also be simply represented. In French, the integer and decimal parts are separated by a comma (12,7 ou 3,896). In English, it's a dot (12.7 ou 3.896). Numbers in scientific notation also have a specific syntax: e.g. *1,23.10E5*; *1,23x10+52*

[1] This free open-source system has the advantage of allowing users to add external lexicons and to refer to their lexical entries in the implemented grammars.

or *4,8x10-6*. They include a coefficient and a base. The coefficient is a decimal number between 1 and 10 not included. The integer part solely contains one digit. The decimal part is more free in size: it depends on the precision required. The number is then adjusted by means of the base which is either a positive or negative integer.

2.3 Nominal numerical determiners

Nominal numerical determiners were systematically described with a finite state representation in [Silberztein 2003,Laporte 2007]. They have the form *Det Nnum de* (Det Nnum of) where

- *Nnum* are numerical nouns representing exact multiples of 10 (*million*, *milliard*, *billion*) or sub-multiples of 10 (*dixième*, *centième*, *milliardième*). There are also numerical nouns used for approximation such as *milliers* (thousands), *centaine* (hundreds), *cinquantaine*, *douzaine* (dozen), *dizaine*.
- *Det* are indefinite determiners (ex. *plusieurs* (several)) or numbers (e.g. *dix, 10*)

(3) **(Dix+10+Plusieurs) millions de** personnes sont allergiques la poussière.
(Ten+10+Several) million people are dust-allergic

(4) Marie attend **une douzaine d'**amis cette semaine
Mary is expecting a dozen friends this week

These deteminers can include a limited set of modifiers that bring little semantic nuances such as in:

(5) Paul a perdu **une petite douzaine de** kilos.
Paul lost a few dozen kilograms.

There also exist semi-fixed forms of these determiners: *des Nnum et des Nnum de*.

(6) Marie a gagné **des centaines et des centaines de** dollars.
Mary earned hundreds and hundreds of dollars

Sequences of nominal numerical determiners are also accepted:

(7) **1,67 milliardième de milliardième de milliardième de** kg.
1.67 billion billion billionth of kg

3 Units

The use of a type of unit depends on the dimension to be measured. Dimensions can be gathered in families the members of which share the same class of appropriate units. For each family, it is therefore necessary to define the associated class, that is the list of all types of appropriate units. Each class is usually divided into two sub-classes: unit symbols - e.g. *m* - and unit names - e.g. *mètre* (meter)-. In practice, each dimension family *X* (e.g. *Energy*) is associated with two hand-drafted graphs: *XUnit* (e.g. *EnergyUnit*) for unit names and *XUnitSymb* (e.g. *EnergyUnitSymb*) for unit symbols. For instance, graph *EnergyUnit* would look like in figure 1. It refers to the graphs representing the different types of units belonging to the class: *Calory*, *Joule*, *ElectronVolt* and *WattHour*. In the two following subsections, we show how these graphs were constructed (either automatically -prefix *genbase-*, either manually -prefix *base-*).

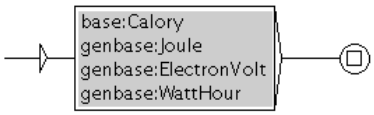


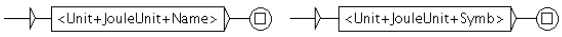
Fig. 1. EnergyUnit

3.1 Basic units

Firstly, we enumerated measurement units by using the scientific classification. We generated a specific dictionary where lexical entries correspond to unit forms (ex. *millimètres*, *mm*). Each entry is assigned a lemma (ex. *millimètre*, *mm*), a category (Unit), morphological information, its scientific unit class (ex. MeterUnit) and information weither the form is a unit name (+Name) or a unit symbol (+Symb). Given a base unit (meter, m), we automatically list all multiples and sub-multiples (e.g. kilometer, km ; millimeter, mm) by adding relevant prefixes and suffixes when required. All morphological information were encoded in a specific configuration file for each base unit, in order to be used at the dictionary generation. For instance, the configuration below for byte units (*ByteUnit*) indicate that the base name and symbol are respectively *octet* and *o*. All derived forms would be in masculine (code *m*) and names accept the plural form by adding suffix *-s* (code *morpho*). All standard prefixes are not accepted and are limited to those encoded *SUP* (*kilo-octet*, **milli-octet*). Names also accept a hyphen between the prefix and the base form (code *hyphen*).

(8) octet:o::Byte:morpho+SUP+m+hyphen

We also generated two graphs for each scientific type: one for unit symbols, another one for unit names. Both make reference to unit forms encoded in the dictionary. For example, graphs in figure 2 represent units derived from *Joule*: *<Unit+JouleUnit+Symb>* and *<Unit+JouleUnit+Name>* respectively indicate all unit symbols (*+Symb*) and names (*+Name*) of type *JouleUnit*. In total, we constructed graphs for 21 unit types.



(a) Joule (b) JouleSymb

Fig. 2. Joule type

In addition, we constructed manually graphs of compound unit names with some variants. For instance, *mille nautique* (nautical mile) have several variants : e.g. *mille (E+marin+nautique)(E+international+britannique)*[2]. Finite state graphs are therefore of great interest to represent them. It is even clearer with currency units that contain many lexical variants. Graph 3 represents the different types of dollars.

[2] Symbol E is the empty word

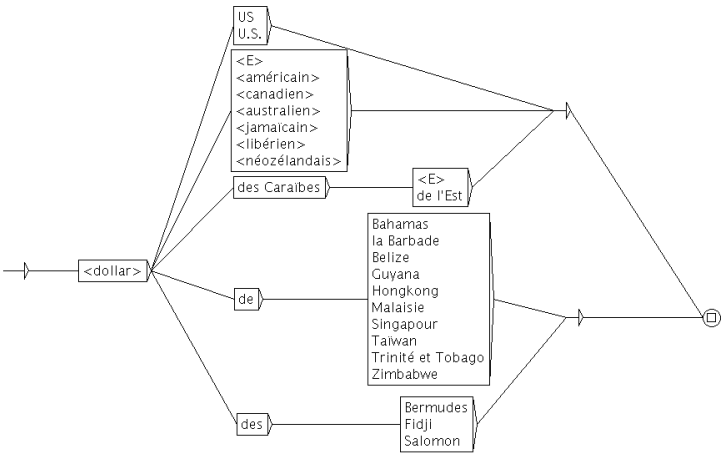


Fig. 3. Dollar

3.2 Complex units

Some units are defined by the combination of basic units. For instance, in physics, a speed unit is the ”division” of a 1D spatial unit and a time unit: *centimètres par heure* (centimeters an hour), *km/s*. In everyday language, there exists some variants such as *kilomètres à l’heure* ou *kilomètres-heure*. Exemple 9 shows that the first expression can also be reduced in the sequence *à l’heure*, but this is only available with the unit name *heure*:

(9) Max roule à une vitesse de 80 (kilomètres+E) à l’heure
Max has a speed of 80 (kilometers+E) an hour

(10) Ce météorite a une vitesse de 2 (kilomètres + *E) à la seconde
This meteor has a speed of 2 (kilometers+*E) a second

Below is given the graph *SpeedUnit* representing this type of units. *LengthUnit* is the union of graphs of units measuring length. *TimeUnit* gathers units measuring time. Note that it contains a simple unit name *noeud* measuring the speed of boats.

The area dimension lexicalized by the words *aire*, *surface*, *superficie* also selects complex units:

– the combination of a length unit followed by either the digit 2 if it’s a symbol or the modifier *carré* (square) if it’s a unit name: e.g. *m2*, *mètre carré*;
– simple surface units like *are*, *hectare* symbolized by *a* and *ha*.

In a similar way, units measuring volume are formed of two types of units: a length unit followed by the digit 3 or the modifier *cube* (cubic); units derived from *litre* (liter)

The scientific unit traditionally associated with frequency is *hertz*. Nevertheless, in everyday language, the unit is more free. It can be any countable noun phrase with no determiner followed by a preposition (or symbol ’/’), an optional determiner and a time unit:

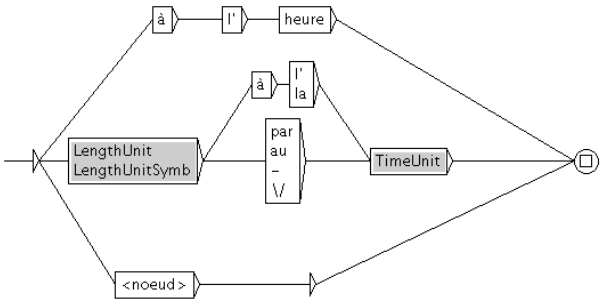


Fig. 4. SpeedUnit

(11) L'usine produit trente poulets (à la+/) minute
The factory produces thirty chicken a minute

Thus, it is necessary to describe a complete noun phrase which in theory is not finite-state. Nevertheless, in practice, this phrase is short because it is followed by a sequence expressing time. We therefore limited our noun phrase graph to the following maximal sequence *Adj N Adj de Det N*. Note that frequency expressions can be expressed in a non-connex manner. They may require the analysis of a whole sentence, which makes finite-state description more difficult even if pragmatic solutions based on finite-state techniques exist. Examples 12 and 13 are equivalent in meaning, which shows that uterrance *par an* (a year) can be inserted in different locations of the sentence (like standard adverbials).

(12) **Par an**, l'usine produit dix mille voitures

(13) L'usine produit dix mille voitures **par an**
The factory produces ten thousand cars **a year**

4 Measure expressions *Dnum Unit*

4.1 Numerical Predeterminers

Simple measurement expressions often contain numerical predeterminers that can be located either before or after *Dnum Unit* like *presque* (almost), *environ* (around), *exactement* (precisely), *à peu près* (around):

(14) Marie a (**environ+exactement**) 10 ans

(15) Marie a 10 ans (**environ+exactement**)
Mary is (**around+precisely**) 10-year old

These words modify the interpretation of the value of the numerical determiner. The distributions of the predeterminers situated before *Dnum Unit* (*PreDnum*) and after (*PreDnumPost*) are different like in:

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

(17) Luc possède 30 voiliers (**?à peu près+environ +ou presque+*presque**)
Luc owns **(around+almost)** 30 sailboats

Predeterminers *PreDnum* et *PreDnumPost* can occur in the same sequence and are easy to represent in the form of graphs in order to integrate them in the basic structure *Dnum Unit*.

(18) Paul a couché avec **environ** mille femmes **au total**
Paul slept with **around** one thousand women **in total**

4.2 Internal constraints

This section focuses on the description of the linguistic constraints existing between the numerical determiner (*Dnum*) and the measurement unit (*Unit*).

Stylistic constraints The two components *Dnum* and *Unit* are firstly subject to stylistic contraints. For instance, the combination of a numerical determiner written in words followed by a unit symbol is not natural although the other combinations are allowed:

(19) * dix m

(20) 10 (m+mètres)

(21) (dix + quelques dizaines de) mètres

Determiner connexity Numerical determiners are not always connex and can be divided in two parts separated by the unit as shown in the examples below.

(22) Max a un retard de **huit** minutes (**trente** + **et demi**)
Max is eight minutes and 30 seconds late

(23) Marie a sauté **5** m **60**
Mary jumped 5.60 meters

There exists a constraint of homogeneity between the left and right parts of the non-connex determiner: they are either both written in words or both written with digits. In addition, the left part of the determiner must be an integer.

Precision The structure *Dnum Unit* can sometimes be commuted to a sequence of several *Dnum Unit*. It has a specific syntax for each unit class. This is mainly used to add precision to a measurement. The following examples show constraints on the sequences of simple phrase *Dnum Meter*:

(24) Jean a couru 10 kilomètres et (30 mètres+*3 000 mètre+*30 secondes)
John ran 10 kilometers and (30 meters+*3,000 meters+*30 seconds)

(25) Ce champ a une surface de 100x200m
This field has an area of 100m by 200m

4.3 Approximations

In this section, we focus on more complex combinations representing approximations of values.

Conjunction *ou* (or) The conjunction *ou* (or) is used to approximate the numerical value of the determiner in the form of a choice between several values.

(26) Max a **cinq ou six** ans
Max is five or six year-old.

The sequence *cinq ou six* (five or six) could be interpretated as a compound determiner, but this involves issues because the sentence 26 is equivalent to the sentence :

(27) Max a cinq ans ou six ans
Max is five year-old or six year-old

Moreover, it is possible to have a sentence like:

(28) La table est à une distance de 90 cm (ou+voire)[3] 1 m du mur
The table is 90 cm or 1 m from the wall

Range There exist three types of complex structures that are used to express ranges of values: (1) entre *entre* $Dnum_1$ $Unit_1$ *et* $Dnum_2$ $Unit_2$ (between ... and ...), (2) *de* $Dnum_1$ $Unit_1$ *à* $Dnum_2$ $Unit_2$ (from ... to ...), (3) $Dnum_1$ $Unit_1$ - $Dnum_2$ $Unit_2$. In all structures, $Unit_1$ and $Unit_2$ can be factorized if $Unit_1$ equals $Unit_2$.

(29) La longueur est comprise **entre 90 cm et 1,10m**
The length is comprised between 90 cm and 1.10 m

(30) La température atteint **(de + E) 13 à 15 degrés**
The temperature reaches 13 to 15 degrees

(31) L'intensité du courant sur cette ligne est de **150 ampères-200 ampères**
The electric intensity on this line is 150 amperes-200 amperes

In construction *de ... à ...*, the preposition *de* is not always obligatory like it is shown in example 30. In many cases, it is soly the introducer of a prepositional measurement phrase and should not be included: for instance, *un bateau de **cinq à six tonnes*** (a boat of five to six tons). Moreover, the construction with *de* is naturally ambiguous. It can also express an evolution in time. Its interpretation depends on the main verb of the elementary sentence.

(32) Le prix du pain est passé de 65 à 70 centimes
The price of the bread went from 65 to 70 cents

In this sentence, the initial price of the bread is 65 cents and reaches 70 cents at the final state. In many cases, this ambiguity is impossible to remove locally.

[3] *voire* is a variant of *ou*

5 Construction of a Local Grammar

All linguistic constraints described in the previous sections were gathered in a local grammar making use of morphological finite-state resources. These resources consisted of the general language dictionary DELA [Courtois and Silberztein 1990] and the dictionary of 2,007 simple unit forms, generated in section 3.1. The elementary linguistic components (numerical determiners and predeterminers, plus units) are described in various graphs:

- Numerical predeterminers can be either predeterminers located before sequence *Dnum Unit* (graph *PreDnum*) or ones located after (graph *PreDnumPost*)
- Numerical determiners can be of different types (graphs *NumberWithLetters*, *NominalDnum*, *NumberWithDigits*) and subtypes (graphs *IntegerNumber*, *DecimalNumber*, ...).
- Each dimension family has its own graphs of appropriate units.

For each dimension family, we automatically produced a graph *DnumUnit* recognizing: (1) simple expressions *Dnum Unit* including predeterminers and internal constraints combining numerical determiners and units; (2) complex structures formed of simple expressions (ex. *entre 10 mm et 15 cm*). We used a so-called parameterized graph[4] (cf. figure 5) that contains all possible constructions which include parameters (@...@) the values of which depend on the dimension family processed. Parameter *@dim_family@* is the name of the dimension family (e.g. *Energy*). The graph refers to two other graphs: *SimpleDnum@dim_family@Unit* (e.g. *SimpleDnumEnergyUnit*) representing simple measurement expressions and *SimpleDnum@dim_family@UnitZ* (e.g. *SimpleDnumEnergyUnit*) representing simple measurement expressions where the unit can be absent. All generated graphs are gathered in one single graph *DnumUnit*, equivalent to a finite state transducer.

The two types of simple expression graphs were also built using a parameterized graph (cf. figure 6). Parameter *@precision@* indicates whether there exists a graph descibing precised expressions (cf. subsection 4.2). Parameter *@zeroing@* indicates whether the absence of a unit is allowed. For each dimension family *@dim_family@*, two graphs (*SimpleDnum@dim_family@Unit* and *SimpleDnum@dim_family@UnitZ*) were produced by resolving the parameters. According to their values, there are two cases: (1) the parameter is replaced by the empty word or the lexical value corresponding to the parameter; (2) the parameter transition is removed.

6 Evaluation

The evaluation process requires a corpus where measure expressions are annotated. Unfortunately, such a corpus does not exist for French. We therefore constituted three corpora: the first one (namely SCIENCE) made of articles of two French scientific periodicals (*Science et Vie*, *Science et Avenir*); the second and third ones (namely NEWS1 and

[4] This concept has been first introduced in the software INTEX [Silberztein 2000]. Note that the system NOOJ [Silberztein 2005] has also an interesting simpler mechanism to integrate such lexical and syntactic constraints into grammars. For this, it uses non strictly finite-state methods.

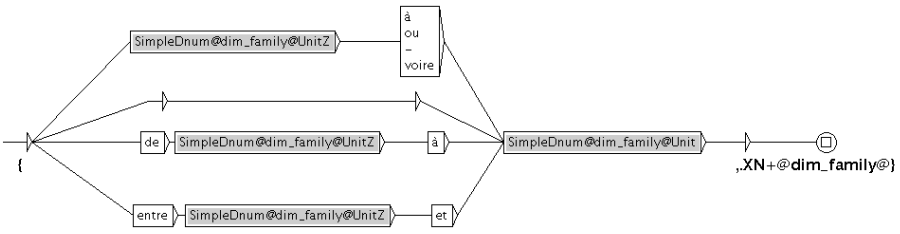


Fig. 5. parameterized DnumUnit

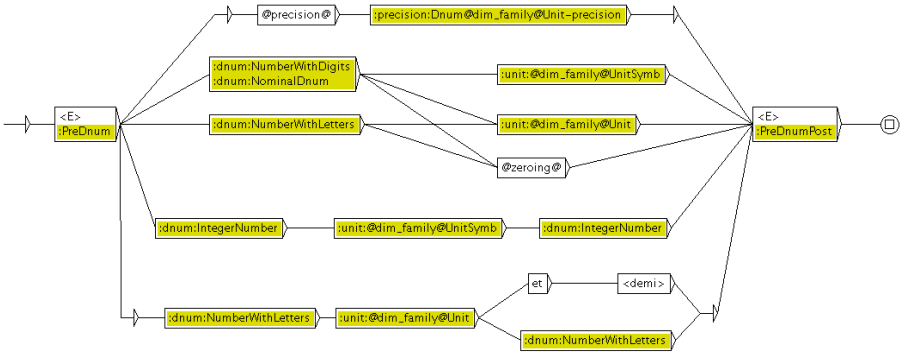


Fig. 6. parameterized SimpleDnumUnit

NEWS2) composed of journalistic news taken from yahoo web site. SCIENCE aims at testing the grammar on many different types of expressions. NEWS1 and NEWS2 aim at evaluating the grammar on expressions in a more common language. NEWS1 is the strict concatenation of journalistic articles. In table 9, we observe that NEWS1 is almost limited to the three main types of numerical expressions studied for NER (time, percentage and monetary expressions). NEWS2 is a concatenation of paragraphs containing the other types of expressions.

Each corpus was divided in two parts of the same size: the first one for the development of the grammar; the other one for its evaluation. After the construction of the grammar, we manually annotated the evaluation part of the three corpora. Different figures on the annotated evaluation corpora can be found in figure 6. Note that the corpora are rather small in terms of number of annotated expressions[5]. An example of the SCIENCE corpus in given below:

De violentes tempêtes de sable emportent {**au moins 150 000 tonnes,.XN+Weight**} de sel et de sable du lit asséché de la mer et les transportent sur {**plusieurs centaines de kilomètres,.XN+Length**}.

[5] The small size of the annotated corpora and the high number of lexical and syntactic measure variants prevent from using a statistical process.

The evaluation stage consists in applying separately three different grammars compiled into their equivalent finite-state transducers on the three corpora with the longest match rule: (1) BASELINE (composed of 135 graphs: 325 states and 9,554 transitions) recognizing very simple sequences formed of a number (with digits and with letters) followed by a unit; (2) SIMPLE (165 graphs: 5,578 states and 69,553 transitions) recognizing simple expressions (cf. section 5); (3) COMPLEX (203 graphs: 8,940 states and 131,408 transitions) being the local grammar described in section 5. The grammars were all compared by using standard evaluation measures: precision (p), recall (r) and F1-measure (f). The results are shown in figure 8.

Eventhough the small size of the corpus attenuates the significance of the results obtained, it is possible to draw some conclusions. Firstly, the COMPLEX grammar reaches relatively good scores for NEWS1 and NEWS2. The heterogeneity of measure expressions in SCIENCE makes them more difficult to annotate automatically: for instance, some infrequent complex units are missing such as *seconde d'arc* measuring an angle; some expressions are not described, e.g. *+55 45' 24"*; there are other non-connex types of expressions (ex. *100g CO2/km*). The experiments also confirm that the measure phrases studied in this paper cannot be reduced to a number followed by a unit. By looking carefully at the annotations produced, we can observe the importance of nominal determiners. Even if complex structures are not so frequent in the corpora, the grammar recognizing them is very useful, especially in the corpus SCIENCE.

Corpus	**#words**	**#sentences**	**#measures**
NEWS1	26,717	785	222
NEWS2	29,611	950	427
SCIENCE	25,411	648	186

Fig. 7. Evaluation corpora figures

	BASELINE	**SIMPLE**	**COMPLEX**
NEWS1	r=0.65 p=0.77 f=0.71	r=0.88 p=0.89 f=0.88	**r=0.89** **p=0.91** **f=0.90**
NEWS2	r=0.67 p=0.73 f=0.70	r=0.87 p=0.88 f=0.87	**r=0.89** **p=0.90** **f=0.89**
SCIENCE	r=0.49 p=0.52 f=0.51	r=0.77 p=0.74 f=0.76	**r=0.84** **p=0.82** **f=0.83**

Fig. 8. Evaluation

NEWS1	Time (51)	Percent (26)	Currency (17)	1D-Space (5)	Misc (1)
NEWS2	1D-Space (38)	Frequency (22)	Weight (12)	Area (9)	Misc (19)
SCIENCE	Time (25)	1D-Space (24)	Weight (9)	Frequency (9)	Misc (33)

Fig. 9. Distribution of types of measurement (percentage in parenthesis)

7 Conclusions and Future Work

This paper focused on a description of measure phrases expressing a value by means of a finite-state approach. We showed that the resulting grammar reaches correct scores when it is applied in a context-free manner. In the short-run, we would to like to size up the evaluation corpora and to increase the coverage of our linguistic resources for improving recall. Future work would also involve integrating the grammar in a more global process taking all other linguistic phenomena into account (a finite-state chunker for instance) in order to improve precision. From this, it should be interesting to implement an automatic process that analyses complete expressions of measurement and annotates the relation between items, dimensions and values.

References

[Chrobot 2000] Agata Chrobot 2000. Description des déterminants numéraux anglais par automates et transducteurs finis. In A. Dister (ed.). *Actes des 3e Journées INTEX*, Revue Informatique et Statistique dans les Sciences humaines, Liège, Belgium, pp. 101-118

[Constant 2002] Matthieu Constant 2002. Methods for Constructing Lexicon-Grammar Resources. The Example of Measure Expressions. *Proceedings of the 3rd Language Resources and Evaluation Conference*, Las Palmas, Spain, pp. 1341-1345

[Courtois and Silberztein 1990] Blandine Courtois, Max Silberztein 1990. Dictionnaires électroniques du français. *Langue Française*, 87, Larousse:Paris

[Gross 1997] Maurice Gross 1997. The construction of local grammars. In Roche, E., Schabes, Y. (eds.). *Finite-State Language Processing*, Cambridge, Mass., The MIT Press, pp. 329–352

[Hasegawa et al. 2008] Yoko Hasegawa, Kyoko Hirose Ohara, Seiko Fujii, Russell Lee-Goldman, and Charles J. Fillmore 2008. Constructions for measurement and comparison in Japanese and English. *International Conference on Construction Grammar 5*

[Karttunen 2006] Lauri Karttunen 2006. Numbers and Finnish numerals. In A Man of Measure; Festschrift in Honour of Prof. Fred Karlsson on his 60th Birthday. special supplement to SKY Journal of Linguistics, 19, pp. 407-421

[Laporte 2007] Éric Laporte 2007. Extension of a Grammar of French Determiners. In C. Camugli, M. Constant, A. Dister (eds.). *Proceedings of the international conference on Lexicon and Grammar*, Bonifacio, pp. 65-72

[Paumier 2008] Sébastien Paumier 2008. *Unitex User Guide*, http://igm.univ-mlv.fr/ unitex

[Silberztein 2000] Max Silberztein 2000. INTEX: An FST Toolbox. Theoretical Computer Science 231(1), Elsevier Science, pp. 33–46

[Silberztein 2003] Max Silberztein 2003. Finite-State Description of the French Determiner system. *Journal of French Language Studies*, 13(2)

[Silberztein 2005] Max Silberztein 2005. NooJ's Dictionaries. *In the Proceedings of the 2nd Language and Technology Conference*, Poznan University