Extending the adverbial coverage of a French morphological lexicon

Elsa Tolone[1], Stavroula Voyatzi[2,3], Claude Martineau[2], Matthieu Constant[2]

1. FaMAF, Universidad Nacional de Córdoba, Medina Allende s/n, Ciudad Universitaria, Córdoba, Argentina
2. LIGM, Université Paris-Est, 5 boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France
3. VIAVOO, 69 rue Danjou, 92100 Boulogne Billancourt, France
elsa.tolone@univ-paris-est.fr,{voyatzi,martinea,mconstan}@univ-mlv.fr

Abstract

We present an extension of the adverbial entries of the French morphological lexicon DELA (Dictionnaires Electroniques du LADL / LADL electronic dictionaries). Adverbs were extracted from *LGLex*, a NLP-oriented syntactic resource for French, which in its turn contains all adverbs extracted from the Lexicon-Grammar tables of both simple adverbs ending in *-ment* (i.e., '-ly') (Molinier and Levrier, 2000) and compound adverbs (Gross, 1986b; Gross, 1986a). This work exploits fine-grained linguistic information provided in existing resources. The resulting resource is reviewed in order to delete duplicates and is freely available under the LGPL-LR license.

Keywords: morhological lexicon, adverb, paraphrase

1. Introduction

Recognising adverbs such as *extrêmement* 'extremely' and *à long terme* 'in the long run' in texts is likely to be useful for information retrieval and extraction because of the information that some of these adverbs convey.

Adverbs, or more generally circumstantial complements, have often been overlooked in the compilation of lexical resources (Nølke, 1990) (p. 3). Several reasons explain this lack of interest. Firstly, adverbs are usually felt as less useful than nouns for information retrieval and extraction. Secondly, compound adverbs in particular are difficult to distinguish from prepositional phrases assuming other syntactic functions, such as arguments or noun modifiers: the distinction is hardly correlated to any material markers in texts and lies in complex linguistic notions (Villavicencio, 2002; Merlo, 2003).

The availability of large-coverage lexicons providing lexical, syntactic and semantic information is essential in order to gain insight on the recognition and analysis of adverbs, including the dual problems of variability and ambiguity (Laporte and Voyatzi, 2008). In addition, it is likely to help solving prepositional phrase attachment during shallow or deep parsing (Agirre et al., 2008).

In this paper, we propose a method for extending the coverage of the French morphological lexicon DELA (Dictionnaires Electroniques du LADL / LADL electronic dictionaries) with respect to adverbial entries. These adverbs were extracted from *LGLex*, a NLP-oriented syntactic resource for French, which in its turn contains all adverbs extracted from the Lexicon-Grammar tables (hereafter LG tables) of both simple adverbs ending in *-ment* '-ly' (Molinier and Levrier, 2000) and compound adverbs (Gross, 1986b; Gross, 1986a).

The paper is organized as follows. Section 2. provides an overview of the resources used in our work. Section 3. describes the method used to enhance DELA thanks to the integration of the new entries of *LGLex*, which are extracted from LG tables. Section 4. is focused on the generation of complex adverbial entries with variables. In section 5., we report and discuss the obtained results. Finally, in section 6. we point out several possible extensions and avenues for future research.

2. Resources

2.1. The LG tables of adverbs

LG tables are currently one of the major sources of lexical and syntactic information for the French language[1]. Their development was initiated as early as the 1970s by Maurice Gross, at the LADL (Gross, 1975), and then at the LIGM, University Paris-Est (Boons et al., 1976; Guillet and Leclère, 1992).

Lexical information is represented as tables. Each table puts together elements of a given lexical-grammatical category (for a given language) that share a certain number of defining features, which usually concern subcategorization information. These elements form a class.

Tables are represented as matrices: each row corresponds to a lexical item of the corresponding class; each column lists a feature that may be valid or not for the different members of the class; at the intersection of a row and a column, the symbol $+$ (resp. $-$) indicates that the feature corresponding to the column is valid (resp. not valid) for the lexical entry corresponding to the row.

The resources described in this paper correspond to the LG tables of both simple and compound adverbs, in which previously implicit features have been made explicit[2] for more convenient use in NLP. All tables are fully available[3] under

[1] LG tables are available in several languages. With regard to adverbs, LG tables exist in English (Gross, 1986b), German (Seelbach, 1990), Spanish (Blanco and Català, 1998), Italian (Gioia, 2001), Portuguese (Baptista, 2003), Korean (Jung, 2005) and Modern Greek (Voyatzi, 2006).

[2] In order to make previous implicit features explicit, a table of classes has been created (Tolone, 2009; Tolone, 2011). Its role is to assign features when their value is constant over a class, e.g. class definition features. Each row stands for a class and each column stands for a feature. Each cell corresponds to the validity of a feature in a class. In particular, the table of French adverb classes is composed of 32 different classes and 164 features.

[3] http://infolingu.univ-mlv.fr/english > Language Resources > Lexicon-Grammar > Download.

a free license (LGPL-LR).
In French, there are two resources of adverbs that follow different principles both in classification and in representation within the Lexicon-Grammar framework (Tolone et al., 2010). That is, first, tables of simple adverbs ending in *-ment* '-ly' (Moliner, 1984; Molinier and Levrier, 2000), which are mainly derived from adjectives and, secondly, tables of compound adverbs[4] (Gross, 1986b; Gross, 1986a). In both tables, there are encoded 3,203 simple and 7,284 compound entries assuming an adverbial function in discourse.

Figure 1 displays a sample of the table PCA which is defined by the morphosyntactic structure Preposition, Determiner, Constrained noun, Pre-adjectival Modifier, Adjective.
In this table, each row corresponds to a lexical item with adverbial function, and each column corresponds to:

- one of the components in the morphosyntactic structure of the items, i.e. features with identifiers Prép, Det, C, Modif pré-adj, and Adj;
- a syntactic feature holding binary values, for example: Prép Det Modif pré-adj Adj C describes the possible permutation (without loss of information) of the adjectival phrase represented in this table as Modif pré-adj Adj; moreover, Neg obl encodes the constraint that the adverbial occurs obligatorily in a negative clause;
- a semantic feature holding binary values, for instance, Conjonction points out whether the compound adverb has a connector function in discourse, i.e. it links the clause in which it occurs with the previous clause as, for example, *dans le cas contraire* 'otherwise';
- an item of information provided as an aid to help human readers find examples of sentences containing the compound adverb: features with identifiers Ppv and Prédicat type give an example of a verbal predicate that combines commonly with the adverb.

Some entries in the table are flexible, like *à échéance :Adj*, which contains the variable :Adj (adjective) in order to recognize adverbs like *à échéance courte* 'short-term', *à échéance longue* 'long-term', etc.

2.2. The syntactic lexicon LGLex

The current version of French LG tables enables the use of their lexical data in NLP tools (Tolone, 2009). To this end, the tables have been converted into an interchange format, based on the same linguistic concepts as those handled in the tables. This conversion is based on *LGExtract*:

[4]According to (Laporte and Voyatzi, 2008) (p. 31) "a phrase composed of several words is considered to be a multiword expression if some or all of its components are tied together, that is, if their combination does not obey productive rules of syntactic and semantic compositionality". This criterion ensures a complementarity between lexicon and grammar. In other words, it tends to ensure that any combination of linguistic elements which is correct in the language, but is not represented in common syntactic-semantic grammars, should be stored in lexicons.

<ID>	N0 =: Nhum	N0 =: N-hum	Neg obl	Ppv	Prédicat type	<ENT>Prép	<ENT>Det	<ENT>C	<ENT>Modif pré-adj	<ENT>Adj	Prép Det C	Prép Det Modif pré-adj Adj C	Prép Modif pré-adj Adj C	Conjonction
19	+	-	-	<E>	faire N	<E>	:TOUT	:N	<E>	confondus	-	-	-	-
5	+	-	-	<E>	faire N	à	<E>	:LUI-0	<E>	:Dnum	-	-	-	-
29	+	+	-	<E>	peser Dnum Npoids	à	:DNUM	:Npoids	<E>	près	-	-	-	-
30	-	+	-	<E>	valoir Dnum Nsomme	à	:DNUM	:Nsomme	<E>	près	-	-	-	-
273	+	+	-	<E>	valoir Dnum Nsomme	à	:DNUM	franc	<E>	près	-	-	-	-
180	+	+	-	:se	produire	à	<E>	échéance	<E>	:Adj	+	+	-	-
716	-	+	-	:se	produire	par	<E>	temps	<E>	:Adj	-	-	-	-
88	-	+	-	:se	produire	dans	le	cas	<E>	contraire	-	-	-	+
89	-	+	-	:se	produire	<E>	le	cas	<E>	échéant	-	-	-	-
1	-	+	-	:se	produire	<E>	le plus	:Adj-ment	<E>	possible	-	-	-	-
91	-	+	-	:se	produire	dans	le	cas	qui	préoccupe :Nhum	+	-	-	-

Figure 1: Compound adverbs of table PCA

a generic tool for generating a syntactic lexicon for NLP from the LG tables (Constant and Tolone, 2010). It relies, first off, on a global table of classes in which we added the missing features and, second, on a single extraction script including all operations related to each feature to be performed for all tables.
Thanks to *LGExtract*, a French lexicon for NLP has been generated from all LG tables and for most lexical-grammatical categories: verbs, predicative nouns, idioms and adverbs. This syntactic lexicon is named *LGLex* (Constant and Tolone, 2010; Tolone, 2011). It is manually evaluated and freely available[5] under the LGPL-LR license in both plain text format and XML.
Each entry of the lexicon includes three sections:

1. section **Lexical-information** identifies the lexical entry and gives the category of each lexical component. For instance, the flexible entry *à échéance :Adj* (including the adverb *à échéance courte* 'short-term'), which is encoded in table PCA (see 2.1.), contains the categories Prép, C and Adj (Det and Modif pré-adj are empty). We added the information of **paraphrases**, **other structures** and **other entries with intensification**;
2. section **Arguments** gives information about the arguments of the predicate: for instance, the subject argument N0, assigned to the predicate that may be modified by the entry *à échéance :Adj* is a human or a non human noun phrase, represented by N0 =: Nhum and N0 =: N-hum;
3. section **Constructions** enumerates the identifiers of all constructions of the lexical entry (e.g. N0 V Adv W or Adv parlant, P)[6] and of all internal morphosyntactic structures, that is Adv for all simple adverbs or Prép Det C Modif pré-adj Adj for compound adverbs like *à échéance :Adj*, but also Prép Det Modif pré-adj Adj C for its variant with

[5]`http://infolingu.univ-mlv.fr/english` > Language Resources > Lexicon-Grammar > Download.

[6]Symbols with obvious interpretation are used such as: Prép (preposition), Det (determiner), Adj (adjective), Modif pré-adj (pre-adjectival modifier), N (noun), V (verb), Conj (conjunction), W (a range of verbal complements), and C (noun tied with the rest of the adverbial structure).

the permutation of the noun C, e.g. *à :Adj échéance* (including the adverb *à courte échéance* 'short-term'), and Prép Det C for its variant without prepositional noun phrase modifier, e.g. *à échéance* 'at expiry date'.

LGLex is currently composed of 13,872 verbal entries (from 67 tables), 14,271 nominal entries (from 81 tables), 39,628 idioms (from 69 tables) and 10,492 adverbial entries (from 32 tables) of which 3,207 are simple adverbs (from 16 tables) and 7,285 are compound adverbs (from 16 tables).
In order to enrich the French DELA, we first extended *LGLex* with respect to adverbial entries by using various types of features that are encoded in the tables of both simple and compound adverbs (Tolone and Voyatzi, 2011). We added 11,351 entries (+108%), so the lexicon is now composed of 21,843 adverbial entries in total.

2.3. The morphological lexicon DELA

The French morphological lexicon DELA (Dictionnaires Électroniques du LADL / LADL electronic dictionaries) describes the simple and compound lexical units of French and provides the corresponding grammatical, semantic and inflectional information. It is freely available[7], and is currently composed of 683,824 simple entries and 108,436 compound entries.
Each entry is represented in its canonical form and contains an inflectional code which allows to automatically generate all inflected forms of the entry using a graph. When we process a corpus with the Unitex software[8] we can apply directly the DELA lexicon. Unitex generates the lexicon of all inflected forms (called DELAF) present in texts; then, it tags them.

An entry of a DELAF is a line of text that ends in a newline and conforms to the following syntax:

paresseuse,paresseux.A+d+z1:fs
'lazy'

The different elements of this line are:

- *paresseuse* is the inflected form of the entry; it is mandatory; *paresseux* is the canonical form (lemma) of the entry. For nouns and adjectives (in French), it is usually the masculine singular form; for verbs, it is the infinitive. This information may be left out as in the following example:

  paresseux,.A+d+z1:ms

  This means that the canonical form is the same as the inflected form. The canonical form is separated from the inflected form by a comma.

- A+d+z1 is the sequence of grammatical and semantic information. In our example, A designates an adjective and d indicates that the adjective occurs after the noun.

  The codes +z1, +z2 and +z3 indicate the language register (this information is optional): +z1 is used for general language (for example, *blague* 'joke'), +z2 for specialized language (for example, *disquette* 'floppy disk') and +z3 for very specialized (or technical) language (for example, *sérialisation* 'serialization').

  Each entry must have at least one grammatical or semantic code, separated from the canonical form by a period. If there are more codes, these are separated by the + character.

- :fs is an inflectional code which indicates that the noun is feminine singular. Inflectional codes are used to describe gender, number, declination, and conjugation. This information is optional. An inflectional code is made up of one or more characters that represent one information each. Inflectional codes have to be separated by the : character, for instance in an entry like the following:

  adverses,adverse.A+d+z1:mp:fp
  'opposite'

  The : character is interpreted as a logical OR. Thus, :mp:fp means "masculine plural" or "feminine plural".

3. Extending DELA

We completed DELA with the new adverbial entries added in *LGLex* in order to evaluate their accuracy by means of both a corpus-annotation practice and a detailed comparison with related work by (Laporte et al., 2008). Using this method, we managed to increase the number of the adverbial entries in the morphological lexicon.

3.1. Adverbial variants in LGLex

The first step of our method consisted of representing various types of features, already encoded in LG tables, as variants of the adverbial entries present in *LGLex*.
These features describe paraphrases (à Adv parler, P or N0 V W de (façon+manière) Adj), substructures (Prép1 Det1 C1 derived from the basic structure Prép1 Det1 C1 Prép2 C2), or intensified structures (plus Adv)[9].

[7] http://infolingu.univ-mlv.fr/english > Language Resources > Dictionaries > Download.

[8] http://igm.univ-mlv.fr/~unitex/

[9] Entries with intensification are included in the lexicon when their combination does not seem to obey regular rules of syntactic and semantic compositionality. For instance, *particulièrement* 'particularly' is a quantifier in the following example and, thus, refuses intensification: *Elle est (*très+*plus) particulièrement grande pour son âge* 'She is (*very+*more) particularly tall for her age'. On the contrary, it can be intensified when it is used as a focus adverb: *Les adolescents, (tout+plus) particulièrement les filles, sont grands pour leur âge* 'Teens, (more) particularly girls, are tall for their age'. However, this is not the case of the focus adverbs *essentiellement* 'essentially', *principalement* 'basically', *notamment* 'notably'. Thus, *plus particulièrement* 'more particularly' is considered as an idiomatic compound, as opposed to an open series of identical forms with a different syntax. It is represented in *LGLex* as a variant of the focus adverbial entry *particulièrement* 'particularly'.

Hence we added the following fields to **lexical-info** :

- **paraphrases**, for instance, *à franchement parler* 'frankly speaking' and *de (manière+façon) franche* 'in a frank way' for the adverb *franchement* 'frankly' ;
- **other-structures**, for instance, *jusqu'à la fin* 'until the end' for the adverb *jusqu'à la fin des (=de les) temps* 'until the end of time' ;
- **other-ID**, refering to other entries with intensification, for instance, *plus particulièrement* 'more particularly' for the adverb *particulièrement* 'particularly'.

The following example is taken from the lexicon *LGLex*[10] and concerns the adverbial entry *paresseusement* 'lazily', which is encoded in table ADVMS of subject oriented manner adverbs. Four new adverbs were added in section **lexical-info** and are represented as **paraphrases**, followed by their internal morphosyntactic structure in **structureAdv** of section **all-constructions** :

```
ID=P_advms_643;status=completed
lexical-info=[cat="adv",
  exprF=[adv=[notperm=[complete=
    "paresseusement"]]],
  paraphrases=(adv="de facon <paresseux>",
    adv="d'une facon <paresseux>",
    adv="de maniere <paresseux>",
    adv="d'une maniere <paresseux>"),
  autres-ID=(),
  autres-structures=()]
args=()
all-constructions=[
  structureAdv=(construction="base::Adv",
    construction="o::de facon Adj",
    construction="o::d'une facon Adj",
    construction="o::de maniere Adj",
    construction="o::d'une maniere Adj"),
  absolute=(construction="true::N0 V Adv W",
    construction="true::Adv, N0 V W",
    construction="o::N0hum V W de (E+une)
      (facon+maniere) Adj"),
  relative=()]
```

In *LGLex*, adverbial variants do not form new entries, but are only considered and represented as variants. If one wants to find in texts the various forms of an adverb using DELA, they cannot do so unless these forms are added in the lexicon and are associated to a canonical (or standard) form.

3.2. Conversion into DELA format

In order to produce adverbial entries in DELA format from the adverbial variants that have been added in *LGLex*[11], we followed the following steps:

- First, a few specific treatments had to be performed only once. For example, the paraphrasal feature du point de vue de Ddef Ndomaine 'from the point of view of Ndomain', present in table ADVPM of viewpoint adverbs, requires to encode explicitely the definite determiner (Ddef) associated to the domain noun (Ndomaine), which is encoded in the table. To do so, we added and encoded a new column for Ddef;
- Second, we retrieve the forms of the lexical-grammatical categories that are described inside angle brackets in various constructions from DELAF. For instance, all verbal manner adverbs ending in *-ment* '-ly' (e.g. *paresseusement* 'lazily') have the paraphrase de (E+une) (façon+manière) Adj 'in a Adjective way'. In order to reconstruct the adverbial entry, it is necessary to associate to each simple verbal manner adverb the correspondent adjective (Adj) in the feminine singular form (here, *paresseuse*, cf. 2.3.);
- Third, we extract the list of adverbial variants from the three following fields present in *LGLex*: **paraphrases**, **other-structures** and **other-ID**;
- Then, we must do a few language-dependent substitutions which are of common use in French. For instance, in *jusqu'à la fin de les temps* 'until the end of time', the preposition *de* 'of' and the definite determiner *les* have to be merged into a single form, and so *de les* becomes *des*;
- Finally, we convert this list into DELA format, linking all variants of a given adverb to its canonical (or standard) form (**complete**), and specifying the grammatical category **ADV**, followed by the name of the morphosyntactic or syntactico-semantic class (found in **ID**) in which it is encoded. All this information is provided in *LGLex* (cf. 3.1.).

For instance, the four variants associated in *LGLex* to the adverbial entry *paresseusement* 'lazily' (see 3.1.), produce in DELA format:

de façon paresseuse, paresseusement.ADV+advms
d'une façon paresseuse, paresseusement.ADV+advms
de manière paresseuse, paresseusement.ADV+advms
d'une manière paresseuse, paresseusement.ADV+advms
'in a lazy way'

Moreover, we added the 24 missing adjectives in DELA[12] followed by their inflectional code. This allowed us to construct the corresponding variants with the feminine form of the adjectives, such as:

[10] This is an extract of the version 3.4 that takes into account the paraphrases which are encoded in LG tables, notably, the associated lexical features.

[11] Another possibility is to construct graphs used in Unitex to generate these variants in DELA format. But graphs must be constructed manually. So, it is better to generate new entries with these paraphrases and other structures that have been already added in *LGLex* rather than create graphs to generate them.

[12] These adjectives are: *abrégé*, *affaireux*, *amiteux*, *amitieux*, *aucun*, *bijournalier*, *componctueux*, *considéré*, *drôlatique*, *exaspéré*, *flâneux*, *gent*, *goujat*, *indévot*, *intra-utérin*, *irrévérent*, *marmiteux*, *maupiteux*, *médiatif*, *peineux*, *plaignard*, *recta*, *révérent*, *salaud*.

de façon abrégée,abrégément.ADV+advmv
'concisely'

Then, we filtered the entries in order to remove duplicates. In fact, some entries are exactly the same. For example, *ces temps derniers* 'recently', defined by the morphosyntactic structure Prép Det C Modif pré-adj Adj and encoded in table PCA, can also take the form *ces derniers temps* due to the permutation of the adjective *derniers* 'recent'. This latter is already encoded in table PAC:

ces temps derniers,.ADV+pca
ces derniers temps,ces temps derniers.ADV+pca
ces derniers temps,.ADV+pac

We can also evoke the case of the deleting relation that associates different adverbs of table PCDC. For example, the adverbs *dans l'état actuel des choses* 'in the current state of things' and *dans l'état actuel des connaissances* 'in the current state of knowledge' are both encoded in table PCDC. They both accept the substructure *dans l'état actuel* 'in the current state', which is obtained after deletion of the prepositional noun phrase modifiers *des choses* 'of things' and *des connaissances* 'of knowledge', and without loss of information:

dans l'état actuel des choses,.ADV+pcdc
dans l'état actuel,dans l'état actuel des choses.ADV+pcdc
dans l'état actuel des connaissances,.ADV+pcdc
dans l'état actuel,dans l'état actuel des connaissances.ADV+pcdc

In fact, each substructure provides information about the corresponding entry in *LGLex*, and so the generated substructures in DELA format are filtered automatically in order to delete duplicates. We obtained 21,587 new entries (22,850 new entries before removing duplicates) in DELA format.

Last, some errors in the new entries are due to the way initial adverbial entries are encoded in tables. Considering the adverb *à cette heure-ci* 'at the present time' represented in table PCA: the noun component *heure* 'time' is encoded together with the hyphen, and thus form an amalgam that is automatically reproduced in the substructure *à cette heure-* 'at this time'. So, we deleted hyphens:

à cette heure- ci,.ADV+pca
à cette heure-,à cette heure- ci.ADV+pca

In addition, some spaces had to be deleted. We mention the previous example, written *à cette heure- ci*, or the entry *y compris* 'including' ending by a space:

y compris ,.ADV+pcpn

These corrections are necessary for two reasons: to improve the quality of entries and to compare with the current version of DELA.

Some entries contain variables like :Adj (adjective) or :DNUM (numerical determiner) (see 2.1.):

par temps :Adj,.ADV+pca
à :Adj échéance,à échéance :Adj.ADV+pca
à :DNUM franc près,.ADV+pca

We have 20,757 entries without variables and 830 entries with variables. The first step consists in integrating all entries without variables in the existing DELA. See the section 4. for the treatment of entries with variables.

3.3. Integration in the existing DELA

All these variants need to be added to the existing DELA, which currently contains 9,036 entries, after merging the two lists and verifying that no variant is already present in DELA. In order to produce the new DELA of adverbs, we followed the following steps:

- First, we removed from the existing DELA the adverbial entries of the morpho-syntactic class PECO as they describe idiomatic adjectival phrases rather than adverbs:

  lent comme une tortue,.ADV+PECO+z1
  'slow as a turtle'

- Then, we lowercased the LG table names in existing DELA in order to homogenize the notation. In addition, we deleted the quotes around the LG table name 'PADV':

  à altitude moyenne,.ADV+pca
  à altitude moyenne,.ADV+PCA+z1
  'at medium altitude'

  abondamment,.ADV+'PADV'+z1
  abondamment,.ADV+padv
  'amply'

  We can see in these examples the code +z1 (cf. 2.3.), which is used in existing DELA but not in *LGLex*. We sorted and compared all pairs of entries in order to detect any differencies between them with respect to the language register;

- Finally, we merged the entries encoded in the 16 LG tables of simple adverbs with the relevant entries present in the existing DELA in order to add the information about the language register. For instance, *abominablement* 'abominably' already exists in current DELA but is followed only by the language register code +z1; after merging the two entries, we obtained a complete new entry:

  abominablement,.ADV+advmqi (LG table)
  abominablement,.ADV+z1 (existing DELA)
  abominablement,.ADV+advmqi+z1 (new entry)

Except entries with variables, we obtained 22,564 final entries (29,793 final entries before removing duplicates) merging 9,036 initial entries and 20,757 new entries in DELA format.

4. Graph dictionary

There are 830 entries with variables[13] (cf. 3.2.) which are represented in the 16 LG tables of compound adverbs.
In order to produce dictionary entries from these data we used what we call a graph dictionary. This is a sort of transducer which calls upon subgraphs and so it is capable to dynamically produce new dictionary entries in a format similar to that of DELA.
Variables present in adverbial entries can be expressed, according to their type, in two different ways:

- Variables such as Adj (adjective), N (noun) and N-hum (non human noun) are represented in the form of lexical masks, respectively <A>, <N> and <N∼hum>. They use lexical items previously recognized by other resources (such as dictionaries or graph dictionaries);
- All other variables are transformed into a call to a subgraph of the same name. So, we created 171 subgraphs (they are empty by default but some subgraphs already exist). For example, variable :DNUM is transformed into a call to the subgraph DNUM that recognizes numerical determiners whether they are expressed in numbers or letters.

For instance, the entries described in the end of section 3.2. are represented as follows:

par temps <A>,.ADV+pca
à <A> échéance,à échéance <A>.ADV+pca
à :DNUM franc près,.ADV+pca

The graph dictionary corresponding to these entries is given in the Figure 2.

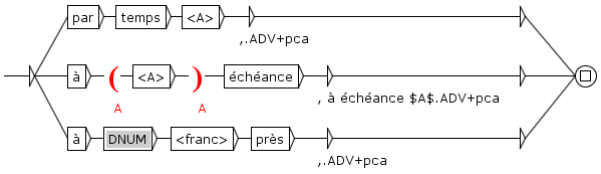

Figure 2: Example of a graph dictionary

The compact notation ",." (cf. 2.3.) indicates that the inflected form and the canonical form are identical. The notation <franc> refers to all inflected forms of the noun. The boxes "A(" and ")A" allow to capture the value of the adjective in order to reproduce it inside the canonical form given in output by "A". The grey box represents a call to a subgraph.

[13]We corrected the names of variables directly in the tables of LG in order to delete duplicates. So, we have 174 differents names of variables in total.

During a corpus processing, the analysis of the sequences *par temps sec* 'in dry weather', *par temps pluvieux* 'in rainy weather', *à brève échéance* 'in the short term' and *à cinq francs près* 'to the very last penny' produce the four following entries in DELA format:

par temps sec,.ADV+pca
par temps pluvieux,.ADV+pca
à brève échéance,à échéance brève.ADV+pca
à cinq francs près,.ADV+pca

In order to generate dictionary entries in DELA format from the *LGLex* adverbial entries with variables we need to create manually all 171 subgraphs. Some subgraphs already exist: DNUM, Ntps, POSS.

5. Results

Table 1 shows the number of the initial adverbial entries in DELA and the new entries extracted from *LGLex*:

	without variables	with variables
Initial entries in DELA	9,036	/
New entries from *LGLex*	21,898	952
New entries from *LGLex* (without duplicates entries)	20,757	830
All entries	29,793	/
All entries (without duplicates entries)	22,564	/

Table 1: Number of entries in DELA

We enhanced the DELA with 13,528 entries, and we have 830 entries with variables which enable the generation of more entries by using the graph dictionary method. The results are quite satisfactory as we obtain 150% new entries in the lexicon only by exploiting precise linguistic information of high coverage, which is freely available in existing resources.

6. Conclusion and future work

At a time when the lack of large-scale lexical syntactic resources for French impedes on NLP research, we have shown the interest of using fine-grained linguistic information, which is provided in existing resources, in order to enrich or diversify their content. This work led to an increase of 150% of the adverbial entries in DELA, and this method is being tested in other languages such as Modern Greek (Voyatzi, 2006). These encouraging results confirm it is worthwhile exploiting features such as paraphrases. Therefore, we plan to complete the LG tables in that direction, starting, for example, with the table of verbal manner adverbs:

Adj-ment = en tout Nabstrait =:
amicalement = en toute amitié
'friendly' 'in all friendship'

Adj-ment = par Nmoyen_communication =:
téléphoniquement = par téléphone
'by telephone' 'by telephone'

Furthermore, we plan to convert the new adverbial entries into the L*efff* format (Sagot, 2010), in order to integrate them into a parser, following similar work by (Tolone and Sagot, 2011) and (Tolone, 2011). Finally, we can also consider enhancing the French Wordnet with respect to adverbial entries (Sagot et al., 2009).

7. References

Eneko Agirre, Timothy Baldwin, and David Martinez. 2008. Improving parsing and pp attachment performance with sense information. In *Proceedings of ACL*, Columbus, Ohio.

Jorge Baptista. 2003. Some families of compound temporal adverbs in portuguese. In *Proceedings of the Workshop on Finite-State Methods for Natural Language Processing*, Budapest, Hungaria.

Xavier Blanco and Dolors Català. 1998. Quelques remarques sur un dictionnaire électronique d'adverbes composés en espagnol. *Lingvisticae Investigationes*, 22:213–232.

Jean-Paul Boons, Alain Guillet, and Christian Leclère. 1976. *La structure des phrases simples en français : Constructions intransitives*. Droz, Geneva, Switzerland.

Matthieu Constant and Elsa Tolone. 2010. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In Michele De Gioia, editor, *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008), Seconde partie*, volume 1 of *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, pages 79–193. Aracne.

Michele De Gioia. 2001. Avverbi idiomatici dell'italiano.

Maurice Gross. 1975. *Méthodes en syntaxe : Régimes des constructions complétives*. Hermann, Paris, France.

Maurice Gross. 1986a. *Grammaire transformationnelle du français : Syntaxe de l'adverbe*, volume 3. ASSTRIL, Paris, France.

Maurice Gross. 1986b. Lexicon-grammar: The representation of compound words. In *Proceedings of the Eleventh International Conference on Computational Linguistics*, Bonn, West Germany.

Alain Guillet and Christian Leclère. 1992. *La structure des phrases simples en français : Les constructions transitives locatives*. Droz, Geneva, Switzerland.

Eun Jin Jung. 2005. *Grammaire des adverbes de durée et de date en coréen*. Ph.D. thesis, Université Paris-Est Marne-la-Vallée, France.

Éric Laporte and Stavroula Voyatzi. 2008. An electronic dictionary of french multiword adverbs. In *Proceedings of the the LREC workshop Towards a Shared Task on Multiword Expressions*, Marrakech, Morocco.

Éric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008. A french corpus annotated for multiword expressions with adverbial function. In *Proceedings of the the LREC workshop Towards a Shared Task on Multiword Expressions*, Marrakech, Morocco.

Paola Merlo. 2003. Generalised pp-attachment disambiguation using corpus-based linguistic diagnostics. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungaria.

Christian Moliner. 1984. *Étude syntaxique et sémantique des adverbes de manière en -ment*. Ph.D. thesis, Université de Toulouse – Le Mirail, France.

Christian Molinier and Françoise Levrier. 2000. *Grammaire des adverbes : description des formes en -ment*. Droz, Geneva, Switzerland.

Henning Nølke. 1990. Classification des adverbes. 88:3–127.

Benoît Sagot, Karën Fort, and Fabienne Venant. 2009. Extending the adverbial coverage of a french wordnet. In *Proceedings of the NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources*, Odense, Danemark.

Benoît Sagot. 2010. The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th Language Resources and Evaluation Conference*, La Valette, Malte.

Dieter Seelbach. 1990. Zur entwicklung von bilingualen mehrwortlexica franzsisch-deutsch-sttzverbkon struktionen und adverbiale ausdrcke. 11:179–207.

Elsa Tolone and Benoît Sagot. 2011. Using Lexicon-Grammar tables for French verbs in a large-coverage parser. In Zygmunt Vetulani, editor, *Human Language Technology, Forth Language and Technology Conference, LTC 2009, Poznań, Poland, November 2009, Revised Selected Papers*, Lecture Notes in Artificial Intelligence. Springer Verlag. To appear.

Elsa Tolone and Stavroula Voyatzi. 2011. Extending the adverbial coverage of a NLP oriented resource for French. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1225–1234, Chiang Mai, Thailand.

Elsa Tolone, Stavroula Voyatzi, and Christian Leclère. 2010. Constructions définitoires des tables du Lexique-Grammaire. In Ljubomir Popović, Cvetana Krstev, Duško Vitas, Gordana Pavlović-Lažetić, and Ivan Obradović, editors, *Actes du 29ème Colloque Lexique et Grammaire*, pages 321–331, Belgrade, Serbie.

Elsa Tolone. 2009. Les tables du Lexique-Grammaire au format TAL. In *Actes de MajecSTIC*, Avignon, France.

Elsa Tolone. 2011. *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*. Ph.D. thesis, LIGM, Université Paris-Est, France.

Aline Villavicencio. 2002. Learning to distinguish pp arguments from adjuncts. In *Proceedings of the 6th Conference on Natural Language Learning*, Taipei, Taiwan.

Stavroula Voyatzi. 2006. *Description morpho-syntaxique et sémantique des adverbes figés en vue d'un système d'analyse automatique des textes grecs*. Ph.D. thesis, Université Paris-Est Marne-la-Vallée, France.