Utilisation des entrées adverbiales du DELA issues des tables du Lexique-Grammaire du français

Elsa Tolone, Stavroula Voyatzi, Claude Martineau

▶ **To cite this version:**

Elsa Tolone, Stavroula Voyatzi, Claude Martineau. Utilisation des entrées adverbiales du DELA issues des tables du Lexique-Grammaire du français. Laporte, Éric ; Smarsaro, Aucione ; Vale, Oto. Dialogar é preciso: Linguística para processamento de línguas, PPGEL/UFES, PPGEL/UFES, pp.243-257, 2013, 978-85-8087-104-3. hal-01443984

HAL Id: hal-01443984

https://hal.archives-ouvertes.fr/hal-01443984

Submitted on 23 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Elsa TOLONE
FaMAF, Universidad Nacional de Córdoba, Argentine
elsa.tolone@univ-paris-est.fr

Stavroula VOYATZI
VIAVOO, France
voyatzi@univ-mlv.fr

Claude MARTINEAU
LIGM, Université Paris-Est, France
martinea@univ-mlv.fr

Utilisation des entrées adverbiales du DELA issues des tables du Lexique-Grammaire du français

Résumé
Dans cet article, nous présentons les nouvelles entrées adverbiales du DELA (Dictionnaire Électronique du LADL) issues des tables du Lexique-Grammaire du français. Ces adverbes ont été extraits à partir des tables d'adverbes en *-ment* (Molinier & Levrier, 2000) et des adverbes figés (Gross, 1986b ; Gross, 1986a). Ils ont été convertis d'abord au format *LGLex* (Tolone, 2011), un lexique syntaxique du français, grâce à l'outil *LGExtract* (Constant & Tolone, 2010), puis au format DELA (Courtois & Silberztein, 1990). Afin d'évaluer les données obtenues, nous avons utilisé le corpus de référence du français annoté par les expressions multi-mots avec leur fonction adverbiale (Laporte et all., 2008). Nous avons ainsi pu produire et analyser l'annotation automatique de ce corpus par les différents niveaux de dictionnaires entrant dans la suite de traitement d'enrichissement des adverbes.
Mots clés : lexiques, Lexique-Grammaire, adverbes, paraphrases

1. Introduction

Reconnaître les adverbes tels que *extrêmement* et *à long terme* dans les textes est utile pour la recherche et l'extraction d'information, de par l'information portée par certains de ces adverbes.
Les adverbes, ou plus généralement les compléments circonstanciels, ont souvent été négligés dans la compilation de ressources lexicales (Nölke, 1990 : 3). Plusieurs raisons expliquent ce manque d'intérêt. Tout d'abord, les adverbes sont généralement ressentis comme moins utiles que les noms pour la recherche et l'extraction d'information. Deuxièmement, les adverbes composés sont particulièrement difficiles à distinguer des phrases prépositionnelles ayant d'autres fonctions syntaxiques, telles que les arguments ou les modifieurs nominaux : la distinction est difficilement corrélée à des marqueurs dans les textes et implique des notions linguistiques complexes (Villavicencio, 2002 ; Merlo, 2003).
La disponibilité de lexiques à grande couverture fournissant des informations lexicales, syntaxiques et sémantiques est indispensable pour améliorer la reconnaissance et l'analyse des adverbes, y compris les problèmes duaux de variabilité et d'ambiguïté (Laporte et Voyatzi, 2008). Cela peut aider à résoudre l'attachement prépositionnel au cours d'une analyse syntaxique profonde ou superficielle (Agirre et al., 2008).
Dans cet article, nous présentons les nouvelles entrées adverbiales du DELA (Dictionnaire Électronique du LADL) issues des tables du Lexique-Grammaire du français. Ces adverbes ont été extraits à partir des tables d'adverbes en *-ment* (Molinier & Levrier, 2000) et des adverbes figés (Gross, 1986b ; Gross, 1986a). Ils ont été convertis d'abord au format *LGLex* (Tolone, 2011), un lexique syntaxique du français, grâce à l'outil *LGExtract* (Constant & Tolone, 2010), puis au format DELA (Courtois & Silberztein, 1990). Afin d'évaluer les données obtenues, nous avons utilisé le corpus de référence du français annoté par les expressions multi-mots avec leur fonction adverbiale (Laporte et all., 2008). Nous avons ainsi pu produire et analyser l'annotation automatique de ce corpus par les différents niveaux de dictionnaires entrant dans la suite de traitement d'enrichissement des adverbes.

Le papier est organisé comme suit. La section 2 fournit une vue d'ensemble des ressources utilisées dans notre travail, incluant les tables du Lexique-Grammaire, le lexique *LGLex*, le DELA et le graphe dictionnaire. La section 3 décrit le corpus ayant servi à l'annotation, ainsi que les différents adverbes annotés et analysés. À titre de conclusion, nous discutons les résultats obtenus dans la section 4, tout en évoquant les travaux futurs.

2. Ressources

2.1. Tables du Lexique-Grammaire des adverbes

Les tables du Lexique-Grammaire sont actuellement l'une des sources majeures d'information lexicale et syntaxique pour le français. Leur développement a été initié dès les années 1970 par M. Gross, au LADL (Gross, 1975), puis au LIGM, Université Paris-Est (Boons et al, 1976 ; Guillet et Leclère, 1992).

Les ressources décrites dans cet article correspondent aux tables du Lexique-Grammaire des adverbes simples et composés, dans lesquelles les propriétés implicites ont été auparavant explicitées afin de les rendre exploitables en TAL (Tolone, 2009). Elles sont entièrement et librement disponibles[1] sous licence LGPL-LR.

Pour le français, il y a deux ressources d'adverbes qui suivent des principes différents à la fois dans la classification et dans la représentation du Lexique-Grammaire (Tolone et al., 2010). Il s'agit, en premier lieu, des tables d'adverbes simples se terminant en *-ment* (Molinier et Levrier, 2000), qui sont principalement dérivés à partir d'adjectifs et, d'autre part, des tables d'adverbes composés (Gross, 1986b ; Gross, 1986a). Dans l'ensemble des tables, sont codées 3.203 entrées simples et 7.285 entrées composés ayant une fonction adverbiale dans le discours.

2.2. Lexique syntaxique *LGLex*

La version actuelle des tables du Lexique-Grammaire du français a permis d'envisager une utilisation de ces données lexicales dans des outils de TAL (Tolone, 2009). À cette fin, les tables ont été converties en un format d'échange, basé sur les mêmes concepts linguistiques que ceux qui sont à l'oeuvre dans les tables. Cette conversion est basée sur *LGExtract,* un outil générique pour générer un lexique syntaxique pour le TAL à partir des tables du Lexique-Grammaire (Constant & Tolone, 2010). Il est relié tout d'abord à une table globale dans laquelle les propriétés manquantes ont été ajoutées, et à un (unique) script d'extraction incluant toutes les opérations liées à chaque propriété devant être effectuées pour l'ensemble des tables.

Grâce à *LGExtract*, un lexique pour le TAL du français a été généré à partir de l'ensemble des tables du Lexique-Grammaire, qui englobent une grande partie des catégories lexico-grammaticales : les verbes, les noms prédicatifs, les expressions figées et les adverbes. Ce lexique syntaxique est appelé *LGLex* (Constant & Tolone, 2010; Tolone, 2011). Il est librement disponible[1] sous licence LGPL-LR au format texte et XML.

Chaque entrée du lexique contient trois sections :

- la section **Lexical-information** indique les informations lexicales liées à l'entrée et, pour les entrées composées, la catégorie de chaque mot de l'entrée. Par exemple, l'entrée semi-figée *à échéance :Adj* (incluant l'adverbe *à échéance courte*), qui est codée dans la table PCA, contient les catégories *Prep*, *C* et *Adj* (*Det* et *Modif pre-adj* étant vides, la structure complète étant *Prep Det C Modif pre-adj Adj*). Les informations de **paraphrases**, d'**autres structures** et d'**autres entrées avec intensification** ont été ajoutées ;
- la section **Arguments** décrit les distributions des différents arguments du prédicat. Par exemple, l'argument sujet *N0* assigné au prédicat *à échéance :Adj* est un nom humain ou un nom non humain, représenté par *N0 =: Nhum* et *N0 =: N-hum* ;
- la section **Constructions** liste différentes constructions dans lesquelles l'entrée peut prendre part (par exemple, *N0 V Adv W* ou *Adv parlant, P*) et toutes les structures morphosyntaxiques internes : *Adv* pour tous les adverbes simples, ou *Prep Det C Modif pre-adj Adj* pour les adverbes composés tels que *à échéance :Adj,* mais aussi *Prep Det Modif pre-adj Adj C* pour sa

[1] http://infolingu.univ-mlv.fr/ > Données Linguistiques > Lexique-Grammaire > Téléchargement.

variante par permutation du nom *C*, par exemple, *à :Adj échéance* (incluant l'adverbe *à courte échéance*) et *Prep Det C* pour sa variante sans modifieur, par exemple, *à échéance*.

LGLex est actuellement composé de 13.872 entrées verbales (provenant de 67 tables), de 14.271 entrées nominales (provenant de 81 tables), de 39.628 expressions figées principalement verbales et adjectivales (provenant de 69 tables) et de 10.488 entrées adverbiales (provenant de 32 tables), parmi lesquelles 3.203 sont des adverbes simples (provenant de 16 tables) et 7.285 sont des adverbes (semi-)figés (provenant de 16 tables).

Les entrées adverbiales de *LGLex* ont été étendues en codant leurs variantes grâce aux différentes propriétés codées dans les tables d'adverbes simples et (semi-)figés (Tolone & Voyatzi, 2011). Cela a permis d'enrichir le lexique de 11.351 entrées (+108%), il est donc composé de 21.839 entrées adverbiales au total.

2.3. Lexique morphologique DELA

Le DELA (Dictionnaire Électronique du LADL) (Courtois & Silberztein, 1990) est un lexique morphologique du français qui décrit les entrées lexicales simples et composées du français en leur associant des informations grammaticales, sémantiques et flexionnelles. Il est librement disponible[2] et est actuellement composé de 683 824 entrées simples et 108 436 entrées composées.

Chaque entrée est représentée par sa forme canonique et contient un code flexionnel qui permet de générer automatiquement toutes les formes fléchies de l'entrée à l'aide d'un graphe. Lors du traitement d'un corpus avec le logiciel Unitex[3], on peut appliquer directement le lexique DELA. Unitex génère le lexique des formes fléchies (appelé DELAF) présentes dans les textes, puis il les étiquette.

Une entrée d'un DELAF est une ligne de texte terminée par un retour à la ligne qui respecte le schéma suivant :

paresseuse,paresseux.A+d+z1:fs

Les différents éléments qui forment cette ligne sont les suivants :

- *paresseuse* est la forme fléchie de l'entrée, elle est obligatoire ; *paresseux* est la forme canonique (lemme) de l'entrée. Pour les noms et les adjectifs, il s'agit en général de la forme au masculin singulier ; pour les verbes, la forme canonique est l'infinitif. Cette information peut être omise comme dans l'exemple suivant :
  paresseux,.A+d+z1:ms
  Cela signifie alors que la forme canonique est identique à la forme fléchie. La forme canonique est séparée de la forme fléchie par une virgule.
- *A+d+z1* est la séquence d'informations grammaticales et sémantiques. Dans notre exemple, *A* désigne un adjectif et *d* indique que l'adjectif se place à droite du nom.
  Les codes *+z1*, *+z2* et *+z3* indiquent le registre du langage (cette information est optionnelle) : *+z1* indique qu'il s'agit d'un mot du langage courant (par exemple, *blague*), *+z2* d'un langage spécialisé (par exemple, *disquette*) et *+z3* d'un langage très spécialisé ou technique (par exemple, *sérialisation*).
  Toute entrée doit comporter au moins un code grammatical ou sémantique, séparé de la forme canonique par un point. S'il y a plusieurs codes, ceux-ci doivent être séparés par le caractère +.
- *:fs* est le code flexionnel qui indique que le nom est au féminin pluriel. Les codes flexionnels décrivent le genre, le nombre, les temps et modes de conjugaisons, les déclinaisons pour les langues à cas, etc. Cette information est optionnelle. Un code flexionnel est composé d'un ou plusieurs caractères codant chacun une information. Les codes flexionnels doivent être séparés par le caractère *:*, comme par exemple dans l'entrée suivante :
  adverses,adverse.A+d+z1:mp:fp

[2] http://infolingu.univ-mlv.fr/ > Données Linguistiques> Dictionnaires > Téléchargement.

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

Le caractère *:* s'interprète comme un OU logique. *:mp:fp* signifie donc "masculin pluriel" ou "feminin pluriel".

Toutes les variantes des entrées adverbiales ajoutées dans *LGLex* ont été converties au format du DELA pour pouvoir les intégrer dans le DELA existant (appelé par la suite "DELA initial"), composé de 9.036 entrées (Tolone *et. all*, 2012). Cela a permis de créer 20.761 entrées directement exploitables et 830 entrées contenant des variables, telles que l'entrée vue en 2.2 contenant la variable *:Adj* (*à échéance :Adj)*. Sans tenir compte des entrées avec variables (voir 2.4), et après fusion et suppression des doublons, le DELA du français (appelé par la suite "DELA étendu") a été enrichi de 13.445 entrées (+149%), il comporte donc 22.481 entrées au total.

2.4. Graphe dictionnaire

Reste donc à interpréter les 830 entrées adverbiales comportant des variables, qui viennent des 16 tables du Lexique-Grammaire des adverbes (semi-)figés. Le but est de pouvoir les reconnaître si elles apparaissent dans les textes. Nous utilisons pour cela ce qu'on appelle un graphe dictionnaire. C'est une sorte de transducteur qui appelle des sous-graphes et est capable de produire dynamiquement de nouvelles entrées du dictionnaire au format du DELA.

Il y a 178 noms de variables au total, chacune doit donc être explicitée.

Certaines variables, telles que *:Adj* (adjectif), *:N* (nom) et *:N-hum* (nom non humain), sont représentées par des masques lexicaux, respectivement, <*A*>, <*N*> et <*N~hum*>. Ils font référence à des unités lexicales précédemment reconnues par d'autres ressources, telles que les dictionnaires ou les graphes dictionnaires.

Toutes les autres variables sont transformées en un appel à sous-graphe du même nom. Nous avons donc créé 175 sous-graphes, vides par défaut. Par exemple, la variable *:DNUM* est transformée par un appel au sous-graphe DNUM, qui reconnaît les déterminants numériques, qu'ils soient exprimés en nombres ou en lettres.

Un exemple de graphe dictionnaire est donné à la Figure 1. La notation compacte ",." (cf. 2.3) indique que la forme fléchie et la forme canonique sont identiques. La notation <*franc*> se réfère à toutes les formes fléchies du nom. Les boîtes "*A(*" et "*)A*" permettent de capturer la valeur de l'adjectif pour pouvoir le reproduire dans la forme canonique donnée en sortie par "*A*". La boîte grise représente un appel à un sous-graphe.

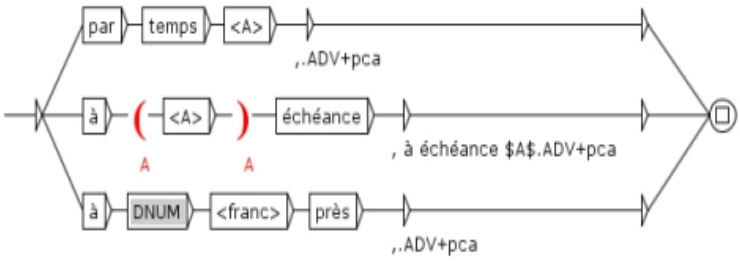

Figure 1 : Exemple de graphe dictionnaire

Certains sous-graphes existent déjà dans d'autres bases de données, comme par exemple GraalWeb[4]. C'est le cas par exemple des sous-graphes DNUM et Dnum-ieme (déterminants numériques ordinaux), qui ont été repris de GraalWeb. De plus, le sous-graphe DNUM peut être réutilisé pour différentes variables (*:Dnum*, *:Dnum-0* et *:DNUM*).

Cela réduit donc le nombre total de sous-graphes à créer. Par exemple, les cinq variables :*Poss-0* (au singulier en position sujet), *:Poss-1* (au singulier en position objet), *:POSS* (au singulier ou au pluriel), *:POSS-0* (au singulier ou au pluriel en position sujet) et *:POSS-1* (au singulier ou au pluriel en position objet)[5] font appel à un même sous-graphe POSS-all, qui contient l'ensemble des pronoms possessifs.

[4] http://igm.univ-mlv.fr/~mconstan/library/index_graalweb.html

On a donc au total 155 sous-graphes ayant un réel contenu et ne faisant pas seulement appel à un autre sous-graphe.
L'ensemble des autres sous-graphes ont donc été complétés manuellement. Par exemple, 1e sous-graphe AB faisant référence à la variable *:AB* correspont aux différentes formes d'écrire *assez-bien*. Il est utilisé par exemple dans l'entrée suivante :
avec mention :AB,avec mention :AB.ADV+pca

3. Évaluation

3.1. Corpus et résultats de l'annotation

Afin d'évaluer les données obtenues au format DELA (aussi bien les entrées présentées en 2.3 que les graphes présentés en 2.4), nous avons utilisé le corpus de référence du français annoté par les expressions multi-mots avec leur fonction adverbiale (Laporte et all., 2008)[6].
Le but est de produire l'annotation automatique (multi-mot) de ce corpus par les différents niveaux de dictionnaires entrant dans la suite de traitement d'enrichissement des adverbes. Il est important de souligner que la référence n'avait pour but que l'étiquetage des adverbes polylexicaux (ou complexes) figés, alors que le DELA contient également les adverbes simples. De plus, des formes temporelles comme *à minuit* avaient plutôt fait l'objet d'un étiquetage en tant qu'entités nommées.
Le corpus annoté contient 8.794 phrases ou 168.846 mots et est composé de :

- la transcription des sessions de l'Assemblé Nationale Française du 3 et 4 octobre 2006 ;
- le roman de Jules Vernes *Le Tour du monde en quatre-vingts jours*, écrit en 1873.

Le tableau 1 donne le nombre d'adverbes annotés dans le corpus pour la référence, l'application du DELA initial, du DELA étendu, du graphe dictionnaire, ainsi que simultanément le DELA étendu et le graphe dictionnaire. Remarquons que le graphe dictionnaire a permis de reconnaître 503 entrées supplémentaires, 224 entrées étant déjà incluses dans le DELA étendu.

Données appliquées au corpus	Nombre d'adverbes annotés
Référence	3.239 (adverbes figés uniquement)
DELA initial (9 036 entrées)	3.690
DELA étendu (22 481 entrées)	4.706
Graphe dictionnaire (830 entrées avec variables)	727
DELA étendu + Graphe dictionnaire	5.209

Tableau 1 : Nombre d'adverbes annotés dans le corpus en fonction des données appliquées.

3.2. Analyse des annotations

Quelques remarques peuvent être faites après avoir analysé les différentes annotations produites.
Tout d'abord, certains adverbes sont reconnus grâce au DELA étendu et non au DELA initial :

(Réference) <ADV fs='PCDN'>sur votre compte</ADV>
(DELA étendu) <ADV fs='PCDN'>sur votre compte</ADV>

Certains adverbes ne sont pas reconnus dans leur intégralité si on les compare à la référence :
(Référence) <ADV fs='PCDN'>sur la recommandation de MM. Baring frères, chez lesquels il avait un crédit ouvert</ADV>
(DELA initial) <ADV fs='PCDN'>sur la recommandation</ADV>
(DELA étendu)<ADV fs='PCDN'>sur la recommandation ***de***</ADV>
Ici, le DELA étendu reconnaît bien la préposition *de*, contrairement au DELA initial.

[5]Les minuscules et les majuscules indiquent le genre de la variable, mais cette distinction n'est pas gardée dans les graphes.

[6]http://infolingu.univ-mlv.fr/corpus/fr-MW-Adv/

(Référence) <ADV fs='PCDN'>au sujet des voyageurs perdus ou égarés</ADV>
(DELA étendu) <ADV fs='PCDN'>au sujet</ADV>
On pourraît reconnaître toutes les combinaisons qui sont prises en compte dans un graphe servant à reconnaître, selon les différents cas de figure, un "groupe nominal humain" (ici *MM. Baring* et *voyageurs*) ou un "groupe nominal non humain". Pour cela, on peut utiliser le graphe de reconnaissance des Entités Nommées du type PERSONNE. Mais, la reconnaissance ne sera jamais complète au risque aussi de générer beaucoup de bruit. Ce serait plus faciles sur un corpus étiqueté en POS ou étiqueté en CHUNKS.

Parfois, les adverbes trouvés sont des adverbes en fonction du contexte, information non codée dans les tables :

- *(DELA étendu, 71 fois) <ADV fs='PC'>pour avis</ADV>* dans *rapporteur pour avis de la commission des affaires économiques.*
  Ici, *rapporteur pour avis* est un nom composé et *rapporteur pour avis de la commission de (la justice + la pêche + les affaires étrangères + des affaires économiques + etc.)* est une collocation. Il faudrait, dans un premier temps, ajouter ces deux noms au dictionnaire, et dans un deuxième temps, créer des règles de priorité selon lesquelles "lorsqu'une séquence présente dans le dictionnaire des adverbes fait partie d'une structure nominale complexe (qui est présente dans le dictionnaire des noms composés) alors la séquence est reconnue en tant que nom".
- *(DELA étendu) <ADV fs='PCDN'>du Droit</ADV>* dans *Institution du Droit* n'est pas un adverbe car il l'est uniquement s'il est suivi de la préposition *de* (prédicat type *agir*, table PCDN) ou si *droit* appartient au domaine sportif (et non du domaine juridique comme ici) comme dans *frapper du droit* (table PDETC).
- *(DELA étendu) <ADV fs='PDETC'>une seconde</ADV>* dans *comme inspirées par une seconde vue* n'est pas un adverbe mais un nom, car il l'est uniquement si *seconde* a le sens de durée (et non de répétition introduit par une préposition non temporelle *par* comme ici), comme dans *durer une seconde* (table PDETC), pouvant être précédé par les prépositions de durée telles que *durant* ou *pendant*. Ici, *seconde* est un adjectif qui modifie (ou détermine) le substantif *vue*.
- *(DELA étendu) <ADV fs='PC'>sur ce</ADV>* dans *si ce n'est sur ce chemin direct* n'est pas un adverbe mais une préposition suivi d'un déterminant, car il l'est uniquement lorsqu'il a le sens de "après ce qui a été dit, après ce qui s'est passé", où il s'agit d'un adverbe temporel déictique qui cooccurre le plus souvent avec des "verbes de mouvement" (table PC, prédicat type *partir*).

L'analyse a permis dans certains cas de détecter des erreurs dans les tables :

- *(DELA étendu) <ADV fs='PJC'>et des</ADV>* n'est jamais un adverbe. Il figurait par erreur dans la table PJC avec d'autres adverbes tels que *et des broquilles*, *et des brouettes*, *et des poussières* (prédicat type *payer N*), qui ont des fonctions conjonctives et ne permettent pas la permutation.
- *(DELA étendu) <ADV fs='PCDN'>sa vie</ADV>* dans *Cependant sa vie était à jour*. On a les entrées suivantes grâce à la colonne *Prép1 Det1 C1 de N2 = Prép1 Poss2 C1* :
  sa vie,toute la vie de.ADV+pcdn
  pendant sa vie,pendant toute la vie de.ADV+pcdn
  Le problème vient de la façon dont l'adverbe figé est découpé en constituants élémentaires et est par conséquent représenté dans la table. Car dans les deux entrées précédentes, *Det1 =: toute la* (forme complexe et dissociable selon la représentation dans la table) mais en réalité il aurait fallu ajouter des colonnes supplémentaires pour représenter, par exemple, *PrepDet1 =: toute* et *Det1=: la*.
  Dans ce cas, les variantes générées auraient été bonnes avec la propriété *Prép1 PrepDet1 Det1 C1 de N2 = Prép1 PrepDet1 Poss2 C1 =:*
  toute la vie de = toute sa vie
  pendant toute la vie de = pendant toute sa vie

Il manque encore des adverbes dans les tables :

- *(Référence) <ADV fs='PDETC'>tout au moins</ADV>*
  (DELA étendu) <ADV fs='PDETC'>au moins</ADV>
  L'adverbe *tout* peut modifier assez librement n'importe quel adverbe/adverbial. Idéalement, il aurait fallu ajouter des colonnes pour décrire la combinatoire des adverbes. Mais cela peut très bien se faire aussi avec des règles. En effet, c'est un phénomène régulier et productif de la langue.
- *(Référence) <ADV fs='PF'>si ce n'est</ADV>*
  (Référence) <ADV fs='PC Conj'>D'ailleurs</ADV>
  Aucun de ces adverbes n'a été annoté avec le DELA étendu, c'est le cas de certains adverbes très fréquents tels que *d'ailleurs*, *de fait*, etc.
- *(Graphe dictionnaire) <ADV fs='PDETC'>à huit heures</ADV>* dans *A huit heures vingt, le cab s'arrêta*. L'adverbe *à :DNUM heures* de la table PDETC (prédicat type *se passer*) est reconnu grâce au graphe DNUM (cf. 2.4), qui identifie le déterminant numérique *huit* . Il peut être complété par un autre déterminant numérique après le nom *heures* : il manque l'entrée *à :DNUM heures :DNUM.*

Une validation manuelle a été réalisée afin de valider les adverbes simples qui ne sont pas en *-ment* et issus d'entrée adverbiales plus longues. Par exemple, l'adverbe *pourboire compris*, qui est défini par la structure morpho-syntaxique *Prép Det C Modif pré-adj Adj*, acceptait la suppression du modifieur *compris*, ce qui produisait une entrée non adverbiale *pourboire* (table PCA). De même pour *chacun* (provenant de l'adverbe *chacun de son côté*, table PCDC), *ironie* (*ironie du sort*, PCDC), *avis* (avis aux amateurs, PCPC), etc. Le codage a été modifié directement dans les tables. En revanche, certains de ces adverbes simples sont bien des adverbes, tels que *jamais*, *point*, *rien* (table PCDC), *partout* (PCPC), *nonobstant* (PCPN), *ici*, *là*, *non*, *oui* (PCA), etc. Mais ils sont ambigus. Cela ne pose donc pas de problème qu'ils figurent dans le lexique *LGLex* comme sous-structures d'une autre entrée. Cependant, ils ne peuvent pas figurer comme de nouvelles entrées dans un lexique comme le DELA. Il faut donc leur ajouter une contrainte supplémentaire, par exemple de négation (*jamais[+Neg]*), après avoir codé dans la table une colonne *Neg obl*.

La référence n'a pas non plus détectée tous les adverbes, donc il difficile de savoir si les adverbes supplémentaires avec le DELA étendu et le graphe dictionnaire sont corrects ou non :

- *(DELA étendu) <ADV fs='PCONJ'>de moins</ADV>* dans *rien de moins communicatif que ce gentleman*. En fait, l'adverbe est *de moins que* (codé dans la table PCPN, prédicat type *faire Dnum N*), c'est une forme discontinue à valeur comparative (table PCPN).
- *(DELA étendu) <ADV fs='PDETC'>des plus</ADV>* dans *l'un des plus beaux gentlemen* n'est pas un adverbe car étant précédé du déterminant *un*, l'adverbe est uniquement *plus*, qui est un adverbe simple régulier (non figés et donc pas compris dans la référence).
- *(DELA étendu) <ADV fs='PDETC'>chaque jour</ADV>* dans *qu'il parcourait chaque jour pour venir*. Il s'agit bien de l'adverbe *chaque jour* de la table PDETC (prédicat type *se passer*).
- *(DELA étendu) <ADV fs='PAC'>A ce jeu</ADV>* dans *A ce jeu du silence, si bien approprié à sa nature, il gagnait souvent*. Il s'agit bien de l'adverbe *à ce jeu* de la table PAC (prédicat type *fatiguer,* sous-structure *Prép Det C* de l'entrée complète *à ce petit jeu*), qui devrait même être l'entrée plus complète *à ce jeu du silence*.
- *(DELA étendu) <ADV fs='PC'>à minuit</ADV>* dans *il ne rentrait chez lui que pour se coucher, à minuit précis*. Il s'agit bien de l'adverbe *à minuit* de la table PC (prédicat type *se passer*), qui devrait même être l'entrée plus complète *à minuit précis*. Comme nous l'avons signalé au départ, des formes temporelles comme *à minuit* ont plutôt fait l'objet d'un étiquetage en tant qu'entités nommées dans la référence.
- *(Graphe dictionnaire) <ADV fs='PDETC'>à huit heures</ADV>* dans *A huit heures, Passepartout avait préparé le modeste sac*. Il s'agit bien de l'adverbe *à :DNUM heures* de la table PDETC, dont le déterminant numérique *huit* est reconnu par le graphe DNUM.

Pour finir, la référence reconnaît également des adverbes erronés (non reconnus par le DELA étendu et le graphe dictionnaire), ce qui encore une fois ne permet pas d'évaluer correctement nos données :

(Référence) <ADV fs='PCDN'>au lieu de quatre-vingt-six</ADV>

4. Conclusion

L'annotation avec le DELA étendu et le graphe dictionnaire permet de reconnaître des adverbes valides, y compris les adverbes simples qui ne sont pas reconnus dans la référence. Cependant, cette nouvelle annotation augmente également le bruit. Il faudrait donc filtrer les nouvelles entrées ajoutées dans le DELA afin d'en retirer ou modifier certaines (en y ajoutant des contraintes supplémentaires), notamment celles qui sont ambiguës.

Nous envisageons ensuite de convertir les nouvelles entrées adverbiales au format Le*fff* (Sagot, 2010), dans le but de les intégrer dans un parseur, comme cela a été fait dans les travaux de (Tolone & Sagot, 2011) et (Tolone, 2011). De plus, nous pouvons également améliorer le Wordnet du français avec ces entrées adverbiales (Sagot et al., 2009).

Références

Agirre E., Baldwin T., and Martinez D. (2008), Improving parsing and PP attachment performance with sense information, in : *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, Columbus, Ohio.

Boons, J.-P., Guillet, A. & Leclère, C. (1976), *La structure des phrases simples en français : Constructions intransitives,* Genève, Librairie Droz.

Constant M. & Tolone E. (2010), A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables, in : De Gioia M. (ed.), *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*, Aracne, p. 79-93.

Courtois B., Silberztein M., (1990), *Dictionnaires électroniques du français*, Langue Française 87, Larousse, Paris.

Gross M. (1975), *Méthodes en syntaxe*, Paris, Hermann.

Gross M. (1986a), *Grammaire transformationnelle du francais : Syntaxe de l'adverbe*, Volume 3, ASSTRIL, Paris, France.

Gross M. (1986b), Lexicon-grammar: The representation of compound words, in : *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn, Allemagne.

Guillet A. & Leclère G. (1992), *La structure des phrases simples en francais : Les constructions transitives locatives*, Droz, Genève, Suisse.

Laporte E., Nakamura, T. & Voyatzi, S. (2008), A French Corpus Annotated for Multiword Expressions with Adverbial Function, in : *Proceedings of the 6th Language Ressources and Evaluation Conference (LREC'08), Workshop on Linguistic Annotation Conference (LAW II)*, Marrakecj, Morocco, p. 48-51.

Merlo P. (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 Linguitics (EACL'03)*, Budapest, Hongrie.

Molinier C. & Levrier F. (2000), *Grammaire des adverbes : description des formes en -ment*, Droz, Genève, Suisse.

Nølke H. (1990), *Classification des adverbes*. Volume 88, p. 3-127.

Sagot B., Fort K. & Venant F. (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.

Sagot B. (2010), The Le*fff*, 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.

Tolone E. (2009), Les tables du Lexique-Grammaire au format TAL, in : *Actes de la 7ème Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'Information et de la Communication (MAJESCTIC'09)*, Avignon, France.

Tolone E., Voyatzi S. & Leclère C. (2010), Constructions définitoires des tables du Lexique-Grammaire, in : *Actes du 29ème Colloque International sur le Lexique et la Grammaire (LGC'10),* Belgrade, Serbie, p. 321-331.

Tolone E. (2011), *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français,* Thèse de doctorat, Université Paris-Est, 340 pp.

Tolone E. & Sagot B. (2011), Using Lexicon-Grammar tables for French verbs in a large-coverage parser, in : Vetulani Z. (ed.), *Human Language Technology, Challenges for Computer Science and Linguistics, 4th Language and Technology Conference, LTC 2009, Poznán, Poland, November 6-8, 2009, Revised Selected Papers*, Volume 6562 of *Lecture Notes in Artificial Intelligence (LNAI)*, Springer Verlag, p. 183-191.

Tolone E. & Voyatzi S (2011), Extending the adverbial coverage of a NLP oriented resource for French, in : *Proceedings of poster session of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, Chiang Mai, Thailande, p. 1225-1233,

Tolone E., Voyatzi S. Martineau C. & Constant M. (2012), Extending the adverbial coverage of a French moprhological lexicon, in : *Proceedings of the 8th Language Ressources and Evaluation Conference (poster LREC'12)*, Istanbul, Turquie.

Villavicencio A. (2002), Learning to distinguish PP arguments from adjuncts, in : *Proceedings of the 6th Conference on Natural Language Learning (CoNLL'02)*, Taipei, Taiwan.