Cahiers du plurilinguisme européen
ISSN : 2970-989X

10 | 2018
Normes et rapports aux normes. Éléments de réflexion pluriels

Dictionnaires électroniques : normes de représentation

Amalia Todirascu

https://www.ouvroir.fr/cpe/index.php?id=1053

DOI : 10.57086/cpe.1053

Electronic reference
Amalia Todirascu, « Dictionnaires électroniques : normes de représentation », *Cahiers du plurilinguisme européen* [Online], 10 | 2018, Online since 01 janvier 2018, connection on 07 novembre 2023. URL : https://www.ouvroir.fr/cpe/index.php?id=1053

Dictionnaires électroniques : normes de représentation

Amalia Todirascu

OUTLINE

TEXT

1 Le problème de standardisation des ressources lexicales (parmi lesquelles on identifie les dictionnaires électroniques, les lexiques, les bases lexico-sémantiques) est un réel défi pour le domaine du Traitement Automatique des Langues (TAL). En effet, la construction de ressources lexicales monolingues et multilingues est une tâche difficile et coûteuse en temps et en ressources humaines. Malgré la disponibilité des méthodes d'extraction automatique des informations à partir des corpus, la construction de ressources lexicales nécessite une intervention manuelle pour la sélection des candidats, pour la structuration des informations. Les ressources lexicales peuvent contenir des informations très variées : informations morphosyntaxiques, syntaxiques, sémantiques. Les informations sont structurées en fonction de l'objectif final pour lequel le dictionnaire a été construit, pour consultation par un humain ou par un système de

traitement automatique des langues. La structure du dictionnaire permet l'accès facile aux informations spécifiques et leur extraction automatique.

2 Les différences de choix quant à la structure et les données représentées dans les dictionnaires constituent des difficultés majeures pour la réutilisation des ressources dans le cadre d'autres applications et contextes d'utilisation. Les dictionnaires traditionnels à caractère encyclopédique (*Le Petit Robert*, *Larousse*) ne permettent pas facilement le partage des informations et sont disponibles en formats « propriétaires » et, par conséquent, peu réutilisables en dehors du contexte pour lequel ils ont été développés.

3 Par ailleurs, la communauté scientifique, en particulier dans le domaine du TAL, privilégie la réutilisation des ressources lexicales existantes ou des informations contenues dans certaines de ces ressources. Afin de rendre ces ressources partageables entre plusieurs applications informatiques, il est nécessaire que les dictionnaires respectent les standards et les normes disponibles, telles que *Text Encoding Initiative* (TEI [1]) ou *Lexical Markup Framework* (LMF [2]) aussi bien pour la structure des données que pour les normes de représentation des informations morphosyntaxiques.

4 Les expressions polylexicales en particulier, comme les collocations et les expressions idiomatiques, posent des problèmes pour la recherche et la représentation des informations syntaxiques et sémantiques pertinentes. Les collocations sont des expressions polylexicales, parfois discontinues, qui présentent un comportement syntaxique et sémantique propre (Gledhill, 2007) : *prendre en considération*, *argument de poids*, *peur bleue*, *battre un record*. Ces phénomènes sont difficiles à identifier par les systèmes de traitement automatique en raison de leurs propriétés syntaxiques et sémantiques propres. Si certaines de ces expressions sont figées (*nid d'anges*), d'autres sont plus variables (*battre rapidement le record*, *battre plusieurs records*). Leur traitement automatique nécessite des ressources spécialisées tels que les dictionnaires de collocations. Ainsi, des expressions idiomatiques sont présentes dans les exemples ou les définitions des dictionnaires électroniques, sans d'autres informations associées que la définition. D'autres ressources mettent l'accent sur les informations morphosyntaxiques ou syntaxiques de ces expressions. Chaque type

d'information nécessite des stratégies différentes d'accès et d'interrogation.

5 Dans ce contexte, nous nous intéressons à la modélisation des dictionnaires électroniques de collocations en format LMF. Nous proposons un modèle LMF adapté pour la représentation de dictionnaires multilingues de collocations, permettant de représenter leurs propriétés morphosyntaxiques, des exemples d'utilisation, des définitions associées à ces expressions. Nous avons transformé un dictionnaire multilingue de collocations, disponible en français, roumain et allemand (Todirascu *et al.*, 2008) en format LMF.

6 Nous allons présenter d'abord les dictionnaires informatisés et lexiques ainsi que leur choix de représentation des expressions polylexicales. Nous présentons la notion de collocations, leurs propriétés et les informations représentées dans les dictionnaires existants dédiés à la représentation de ces expressions. Nous présentons la structure de notre dictionnaire multilingue de collocations (Todirascu *et al.*, 2008), construit sur la base du matériel lexical identifié à partir de corpus monolingues et multilingues. Les normes TEI et LMF seront présentées dans la section suivante. Nous discuterons le modèle LMF proposé avec des exemples extraits de notre dictionnaire dans la dernière section de l'article.

1. Dictionnaires informatisés *versus* lexiques pour le TAL

7 Les dictionnaires informatisés, tels que le *TLFi*, le *Oxford Dictionary* ou le *Collins*, sont construits d'après le modèle des dictionnaires classiques. La microstructure et la macrostructure (Rey-Debove 1971, Wiegand 1988) reprennent cette organisation. En ce qui concerne la macrostructure (l'ensemble des lemmes ou la nomenclature), les entrées sont organisées par ordre alphabétique. En général, une entrée correspond à une unité lexicale, ce qui inclut les mots simples, les mots composés à l'aide d'un signe tel que le tiret (*porte-fenêtre*) ou l'apostrophe, ou les mots composés résultant d'une affixation (*décomposer*). Les expressions polylexicales, telles que les collocations et les expressions idiomatiques, sont parfois présentes à titre d'exemples

ou d'illustrations d'un sens particulier mais sans être utilisées comme mot vedette.

8 Les homonymes et des mots polysémiques demandent des stratégies différentes pour la représentation dans le dictionnaire. Ainsi, les homographes appartenant à des catégories lexicales différentes (*politique* — adjectif ou nom) sont représentés dans plusieurs entrées différentes. Pour les mots polysémiques, on fait souvent le choix de représenter tous les sens dans la même entrée.

9 La microstructure contient les informations attachées à chaque entrée lexicale. On y retrouve des informations concernant la catégorie lexicale, le genre (pour les noms), les définitions génériques et spécifiques à un ou plusieurs domaines, des citations illustrant le sens choisi et des expressions se formant à l'aide du mot vedette. Des informations concernant l'étymologie des mots et éventuellement des informations sur les procédés de formation de mots peuvent être présentes dans l'entrée, ainsi que des synonymes ou des antonymes. Parfois, ces informations sont complétées par des enregistrements sonores, par des informations concernant la fréquence d'usage dans des corpus de référence ou par des traductions proposées (*Collins* [3], *WordReference* [4]).

10 Par rapport au format papier, les dictionnaires électroniques sont dotés des fonctions avancées de recherche dans la macrostructure ou la microstructure du dictionnaire. Ainsi, il est possible d'extraire une partie de la nomenclature à l'aide de recherches par expressions régulières dans le mot vedette ou dans les diverses parties de la microstructure, dans les définitions, les exemples ou la partie étymologique. C'est le cas des dictionnaires tels que celui de l'Académie française [5], du *Trésor de la Langue Française informatisé* [6]. On peut mettre en valeur les liens de synonymie par des liens de type hypertexte ou encore des liens hyperonymiques/hyponymiques comme c'est le cas dans le *Digitales Wörterbuch der deutschen Sprache* (DWDS [7], allemand) ou le *Collins dictionary* (anglais).

11 Les lexiques utilisés dans le domaine du traitement automatique des langues se concentrent sur certaines informations présentes dans les dictionnaires encyclopédiques. Les entrées sont des collections de formes fléchies, contenant la partie de discours, le lemme associé, les informations morphosyntaxiques correspondantes (genre et nombre

pour les noms et les adjectifs, les modes et les temps pour les verbes) mais aussi la fréquence du mot dans des corpus de grande taille, tel que le *Glàff*[8] (Hathout *et al.*, 2014). Parfois, le lexique est une liste de lemmes. Pour chaque lemme, on représente la liste des formes fléchies correspondantes (*Morphalou*[9]) et les formes associées au lemme. En général, les collocations sont absentes de ces ressources.

12 Ces ressources limitent le nombre d'expressions polylexicales qui sont utilisées à titre d'exemple dans la plupart des ressources présentées et les recherches de ces expressions sont en général complexes. Les dictionnaires de collocations doivent palier ces problèmes d'accès à la ressource.

2. Dictionnaires de collocations

2.1. Collocations — problèmes définitoires

13 Les collocations ont fait l'objet de nombreuses études en TAL, en linguistique et en traduction. Du point de vue des linguistes, il s'agit des combinaisons de mots dont le sens n'est pas toujours compositionnel (Hausmann, 2004), « des associations de mots apparaissant souvent ensemble » (Firth, 1968 ; Sinclair, 1991) ou des expressions lexicalisées récurrentes qui sont reliées par des relations syntaxiques (Williams, 2003). Les collocations sont souvent considérées comme des cooccurrences privilégiées de deux mots (d'une base et d'un collocatif), reliées par des relations syntaxiques (Hausmann, 2004 ; Mel'čuk, 1992) ou une relation binaire entre deux éléments (Tutin, 2010 ; L'Homme, 2003). La base conserve son sens d'origine et le collocatif complète le sens de la base. Certaines collocations incluent comme base ou collocatif une autre collocation (Nerima *et al*, 2003). Sans prendre en compte en particulier une base, les collocations sont considérées comme unités polylexicales discontinues, ayant un comportement syntaxique spécifique et un sens souvent non-compositionnel (Gledhill, 2007 ; Odijk, 2013).

14 Dans une perspective TAL, les collocations sont identifiées par leur contexte et par leurs propriétés morphosyntaxiques (Ritz et Heid, 2006) ou alors par des critères statistiques (Manning et Schütze,

1999) ou une combinaison de deux approches (Ramisch, 2012). Des liens sémantiques s'établissent entre le noyau et les collocatifs (Polguère, 2003), décrits à l'aide de fonctions lexicales. Ces définitions montrent la diversité des propriétés à représenter dans un dictionnaire ou un lexique.

2.2. Les propriétés de collocations

15 Les collocations ont un comportement lexical bien défini : le choix des verbes ou des noms n'est pas toujours libre, répondant aux critères sémantiques et pragmatiques spécifiques. Pour ces expressions, une traduction mot-à-mot est souvent incorrecte (en français, *poser une question* et non **demander une question*, mais *ask a question* est tout à fait acceptable en anglais). Souvent les collocations ont des propriétés morphosyntaxiques propres : certaines manifestent une forte préférence pour un nom avec déterminant zéro (*tenir compte*) ou défini (*faire l'objet*), alors que d'autres combinaisons sont plus variables et acceptent des modifieurs (*prendre des mesures drastiques*). En ce qui concerne le sens, il est plus ou moins compositionnel : les expressions idiomatiques ont un sens complètement différent de leurs éléments composants (*jeter l'éponge* = *abandonner*) ; mais pour certaines collocations, le sens reste encore déductible de ses éléments composants (*battre un record*, *prendre des mesures*).

16 De nombreuses études en linguistique (G. Gross 1996 ; Mel'čuk, 1984, 1988, 1992, 1999) identifient des propriétés syntaxiques et sémantiques pour les diverses catégories de collocations. M. Gross (1993) propose une constellation de propriétés lexico-syntaxiques et sémantiques, représentée dans les tables LADL[10] (Laporte, 2000) pour décrire l'environnement syntaxique de certaines locutions verbales et expressions idiomatiques en français. Gaston Gross (1994) classifie certaines locutions par rapport aux critères d'opacité et de compositionnalité et propose une description de leurs propriétés syntaxiques et sémantiques contextuelles.

17 Malgré le manque de définition consensuelle, on peut constater que les collocations sont caractérisées par :

- des cooccurrences fréquentes des mots qui composent la collocation. Il s'agit des mots qui manifestent une forte association lexicale (Hausmann, 2004 ;

Hoey, 2005) ;

- des relations lexico-syntaxiques qui s'établissent entre les composants de la collocation. Par exemple dans une collocation verbo-nominale, le nom est l'objet direct du verbe, ou dans une combinaison N de N (*argument de poids*), le deuxième nom est complément du nom de la base ;
- des combinaisons lexicalisées. Les collocations se combinent avec d'autres constituants syntaxiques, comme un mot simple (*fait l'objet* accepte un objet direct qui est introduit par la préposition *de*) ;
- un rôle pragmatique spécifique et un sens parfois opaque, non déductible à partir des sens des éléments qui la composent. En effet, l'usage impose l'appel à une collocation qui prend un sens différent de ses composants.

2.3. Quelles informations à représenter dans les dictionnaires de collocations ?

18 Les collocations sont peu représentées dans les dictionnaires informatisés ou dans les lexiques pour le TAL en raison de leur grande variabilité syntaxique. Les informations qui sont représentées sont dépendantes de l'objectif pour lequel le dictionnaire a été construit. Il est parfois difficile de représenter les propriétés mentionnées dans la section précédente dans un dictionnaire. De plus, il faut tenir compte de leur usage en contexte.

19 Pour un traducteur ou un apprenant d'une langue, il faut les lister dans un dictionnaire informatisé, avec leurs définitions et leurs propriétés. La *Base lexicale du français* (Verlinde *et al.*, 2003) propose une liste complète de collocations, avec le sens associé. Les informations présentes dans ce dictionnaire sont très riches : les contextes syntaxiques d'utilisation, les patrons de sous-catégorisation, le sens et la définition de la collocation. Ces informations sont complétées par des requêtes faites sur corpus : *JRC-Acquis* (Steinberger *et al.*, 2006), *Europarl* (Koehn, 2005), proposant des exemples extraits de corpus et des informations de fréquence.

20 Les informations représentées dans les dictionnaires de collocations reflètent le point de vue théorique adopté pour modéliser les collocations. Les dictionnaires qui sont inspirés par la théorie Sens-Texte (Mel'čuk, 1996) tels que *LAF* (Polguère, 2007), *DicoWeb* (Polguère, 2003) ou *DICE* (Alonso *et al.*, 2010) représentent les liens entre des

éléments à l'aide des fonctions lexicales, et la définition y est systématiquement proposée. En plus, les combinaisons syntaxiques disponibles sont représentées dans ces dictionnaires.

21 Pour les systèmes de traitement automatique, les dictionnaires doivent fournir des informations syntaxiques et sémantiques détaillées pour chaque collocation. On retrouve des informations dans certains lexiques développés pour le TAL. Dans les tables Lexique-Grammaire (M. Gross, 1994), on représente les contextes syntaxiques de chaque mot, avec un ensemble de contraintes qui s'appliquent sur les arguments (sujet, objet direct, objet indirect). Ces contraintes sont syntaxiques (les types de constituants acceptés), sémantiques (les sujets humains, non-humains) ou lexicales (Laporte, 2000). De même, le *Lefff*[11] (Sagot, 2010) présente des propriétés syntaxiques du collocatif. Le *Lexicoscope* (Kraif et Diwersy, 2012) propose d'extraire des contextes associés à une analyse syntaxique.

22 D'autres dictionnaires de collocations tel que le *DuELME*[12] (Odijk, 2013), disponible pour le néerlandais, regroupent les collocations par leur comportement syntaxique pour éviter les redondances. Ce dictionnaire représente les collocations par leurs patrons syntaxiques (qui s'appliquent à plusieurs collocations), les éléments qui composent la collocation, une glose et la traduction. La fréquence d'apparition des patrons et les propriétés morphologiques sont aussi représentées dans ce dictionnaire. D'autres dictionnaires de collocations sont développés pour un domaine spécialisé : le dictionnaire danois (Braatsch, Olsen, 2000), le dictionnaire franco-allemand de collocations nominales (Blumenthal, 2007), un dictionnaire extrait à l'aide d'un système d'extraction (Nerima et Wehrli, 2008). Ces ressources représentent l'ensemble des propriétés morphosyntaxiques associées aux collocations.

23 Nous identifions plusieurs catégories d'informations qui nous semblent nécessaires pour l'identification automatique des collocations ou pour consultation par un utilisateur humain :

1. les propriétés morphosyntaxiques de collocations (les types d'arguments) ;
2. les propriétés morphosyntaxiques des éléments qui composent la collocation (par exemple les modifieurs possibles, les préférences pour certains déterminants ou diathèse) ;
3. la définition et des exemples illustrant cette définition ;

4. les contextes d'utilisation et leur fréquence.

24 Les ressources lexicales présentées développent en général un seul aspect (syntaxique, sémantique). L'accès et l'extraction de ces informations est difficile en raison des formats spécifiques adoptés par chaque ressource. Une solution permettant d'améliorer la recherche et l'extraction des propriétés des collocations peut être représentée par les normes de représentation des dictionnaires, telles que la TEI et la LMF.

3. Normes pour la représentation des dictionnaires électroniques

25 Les standards et les normes de représentation facilitent le partage et la portabilité des ressources lexicales (Aristar-Dry *et al.*, 2012). Parmi les normes utilisées pour structurer ces ressources (Mangeot et Enguehard, 2013), nous présentons la norme TEI et la norme LMF.

3.1. La norme TEI

26 La norme TEI a été créée pour la représentation standardisée des ressources électroniques en format numérique. La TEI est construite sur le langage à balise XML[13] et est adoptée pour plusieurs ressources lexicales gérées par le projet ORTOLANG[14].

27 Pour représenter les données des dictionnaires (dans une balise <lexicon>), la TEI prévoit deux types d'entrées : <entry>, ayant une structure rigide, comprenant des informations orthographiques, lexicales, définitions et exemples, étymologie, prononciation, et <entryFree> qui permet de rédiger l'article du dictionnaire d'une manière très libre, mais en combinant les mêmes éléments dans le désordre. Un dictionnaire est une liste d'éléments <entry> ou <entryFree>.

```
<entry>
  <form><orth>axe</orth></form>
  <gramGrp><pos>subst.</pos><gen>masc.</gen></gramGrp>
  <sense n="1">
    <def> Ligne qui partage un objet, un corps en deux parties symétriques dans le sens de la plus grande dimension </def>
    <cit type="example"><quote>L'axe du corps humain.</quote>
```

```
</cit>
  </sense>
</entry>
```

28 Une entrée <entry> contient des informations concernant la forme <form> (destinée à représenter la forme et la prononciation), les informations morphosyntaxiques (<gramGrp> contenant la partie de discours <pos> et le genre <gen>). Le <sense> regroupe la définition <def> et des exemples <cit>. Cette représentation ne laisse pas la place à la description des variations syntaxiques éventuelles (pour les expressions polylexicales). Le format proposé est plus adapté pour la représentation des dictionnaires électroniques que pour la consultation manuelle, mais il est aussi possible de représenter des lexiques pour le traitement automatique des langues. Toutefois, la représentation sous format TEI peut varier (par exemple, en cas d'homographes, on représente plusieurs possibilités sous une seule <entry>) ce qui limite parfois la possibilité de partager et de réutiliser les données. Par exemple, il est possible de changer l'interprétation de la balise <cit> pour identifier un lien vers une autre ressource électronique (corpus ou dictionnaire) ou vers une traduction (dans ce cas, le dictionnaire est orienté et la direction de traduction est importante).

29 Plusieurs dictionnaires s'inspirent de la TEI pour structurer les données dans un dictionnaire. Pour *TLFPhraseo* (Jacquey *et al.*, 2018), les entrées sont des expressions idiomatiques et des collocations présentes dans le *TLFi*. Dans le *TLFi*, ces expressions sont accessibles via un mot vedette qui en fait partie : *débarrasser la table* et *débarrasser le plancher* sont disponibles dans l'entrée du verbe *débarrasser*. Dans *TLFPhraseo*, ces expressions deviennent des mots vedette. Pour une expression, plusieurs formes normalisées sont présentes (par exemple des mots composés écrits avec un tiret ou sans tiret), et on définit la propriété de contiguïté dans <gramGrp>.

30 Cette norme a l'avantage de respecter la structure et les informations que l'on trouve dans les dictionnaires en format papier, permettant des recherches similaires. Par contre, il n'est pas possible de représenter des variations syntaxiques ou morphologiques sous forme synthétique : il faut faire la liste exhaustive des variantes dans chaque entrée.

3.2. La norme *Lexical Markup Framework* (LMF)

31 *Lexical Markup Framework* (LMF), une norme ISO-24613:2008 (Francopoulo, 2013 ; Francopoulo *et al.*, 2006a ; Romary, 2002) propose un modèle générique pour la représentation de toutes catégories de ressources lexicales : des lexiques pour le TAL, des bases lexico-sémantiques, des dictionnaires en format électronique. LMF représente les informations classiques à représenter dans le dictionnaire (lemme, définition) mais aussi des informations multilingues, syntaxiques ou sémantiques. De plus, cette norme prévoit de décrire les variations morphologiques et le comportement morpho-syntaxique des expressions polylexicales, par plusieurs extensions (Francopoulo *et al*, 2006a).

32 Selon la norme LMF[15], une ressource lexicale contient un ou plusieurs lexiques monolingues <*Lexicon*>. Un lexique contient plusieurs entrées lexicales <*LexicalEntry*>. Une entrée contient la partie de discours du mot (<*partOfSpeech*>), le lemme (<*Lemma*>) et toutes les formes fléchies du mot (<*Word Form*>). Une entrée lexicale intègre un ou plusieurs sens (<*Sens*>). Un <*Sens*> est illustré par des exemples (<*Sense Example*>) et par une définition sémantique (<*Semantic Definition*>).

33 **Les informations syntaxiques** (<*SyntacticBehavior*>) sont associées à l'entrée lexicale <*LexicalEntry*>. Le comportement syntaxique de l'entrée est décrit par des patrons génériques <*Subcategorization Frame*>, composé par plusieurs arguments <*Syntactic Argument*> (on indique la fonction et le type de constituant).

34 **Les entrées multilingues** sont représentées dans le modèle LMF par l'intermédiaire d'un pivot <*SenseAxis*> (Francopoulo *et al.*, 2006b). Chaque partie monolingue du dictionnaire a la même macrostructure et microstructure. Les correspondances entre les langues sont indiquées par la balise *SenseAxis*. Il est possible de changer facilement de direction de traduction. Dans l'exemple présenté dans l'annexe A, on indique les correspondances entre deux termes spécialisés du domaine médical, en format XML (*gonadotrophine — gonadotropin*).

35 **Les expressions polylexicales** sont représentées par des patrons permettant de combiner plusieurs mots dans une expression polylexicale et de représenter leur structure syntaxique. Les entrées du lexique peuvent être des expressions polylexicales, à l'aide de l'élément <*ListOfComponents*> qui regroupe plusieurs entrées lexicales <*LexicalEntry*> déjà présentes dans le dictionnaire.

36 Cette représentation modulaire permet une extraction facile des informations dans la ressource lexicale, d'une manière générique : on peut ainsi représenter les informations concernant les dictionnaires informatisés mais aussi des lexiques pour les applications de TAL. Il est possible de travailler avec les dictionnaires monolingues ou multilingues, et on peut facilement changer de direction de traduction. Les données ainsi représentées peuvent être réutilisées entre plusieurs applications. Si l'on a uniquement besoin d'informations morphologiques pour une application simple, on peut extraire ces informations sans utiliser la totalité des informations du dictionnaire. Ce modèle proposé par LMF est exhaustif, permettant la représentation d'une grande variété de ressources multilingues et monolingues.

37 Nombre de projets liés à la construction de ressources lexicales l'adoptent, qu'il s'agisse de lexiques en format électronique, de dictionnaires contenant des expressions multiples ou simples ou encore des dictionnaires multilingues : dictionnaires multilingues de noms propres (Bouchou et Maurel, 2008), ressources lexico-sémantiques (Eckle-Kohler *et al.*, 2012), dictionnaire de synonymes (Henrich et Hinrichs, 2010), dictionnaires collaboratifs, tel que W*iktionnaire* (Serraset, 2012), dictionnaires des langues peu outillées (Aristar-Dry *et al.*, 2012 ; Salmon-Alt *et al.*, 2005). Plusieurs projets de conversion des dictionnaires monolingues ont eu comme objectif la représentation en format LMF : les tables du Lexique-Grammaire (Laporte *et al.*, 2013) ou le dictionnaire *DuELME* pour le néerlandais (Odijk, 2013). Dans la même lignée de travaux, nous présentons un dictionnaire de collocations (Todirascu *et al.*, 2008) qui sera adapté à la norme LMF.

3.3. Un dictionnaire multilingue de collocations

38 Nous avons construit un dictionnaire (Todirascu *et al.*, 2008) qui contient des collocations verbo-nominales et leurs équivalents en

trois langues différentes (français, allemand, roumain), afin qu'il puisse être utilisé par un système de traitement automatique de langue. Les collocations sont des expressions polylexicales, parfois discontinues, ayant un comportement syntaxique et sémantique propre (Gledhill, 2007). Les collocations sont caractérisées par deux aspects : la fréquence d'apparition et les relations syntaxiques qui s'établissent entre les mots (prédicat-objet direct, etc.). Par ailleurs, les collocations manifestent une préférence marquée pour un nombre de propriétés contextuelles (dépendantes de langue) (Heid & Ritz, 2005) : le nom manifeste une préférence pour l'article défini ou apparaît sans déterminant, le complément direct du verbe est identifié par une préposition spécifique ou par une marque de cas (en allemand ou en roumain), le verbe apparaît souvent au passif, etc. Une analyse linguistique détaillée a permis d'identifier les propriétés morphosyntaxiques les plus pertinentes pour une extraction automatique, pour les trois langues étudiées (français, allemand, roumain) (Gledhill, 2007).

39 Deux classes de collocations verbo-nominales sont représentées dans le dictionnaire (Todirascu *et al.*, 2008), suivant la définition de Gledhill (2009) :

1. les prédicateurs complexes, qui ont des propriétés contextuelles figées : absence ou préférence pour le déterminant zéro, impossibilité de modifier le nom, impossibilité de mettre le verbe à la diathèse passive (*mettre en œuvre*, *tenir compte*, *faire l'objet*). De plus, le sens de ces expressions n'est pas compositionnel et expriment un procès relationnel (Halliday, 1985) ;
2. les prédicats complexes, qui acceptent un degré plus important de variabilité (le nom peut être modifié par des adjectifs ou par des relatives). Au niveau sémantique, le verbe et le nom expriment ensemble un procès mental (Halliday, 1985) (*prendre des mesures*, *arriver à un accord*).

3.3.1. La macrostructure du dictionnaire

40 Nous avons sélectionné un nombre d'environ 250 collocations verbo-nominales pour chaque langue. Chaque entrée contient des équivalents de traduction qui nous ont permis de mettre en évidence plusieurs cas représentés dans le dictionnaire :

- l'équivalent de traduction est une collocation de la même catégorie qu'en langue source (ayant un sens non-compositionnel) (da naştere/ donner lieu) ;
- plusieurs collocations "libres" traduites mettent en évidence la préférence pour un verbe ou nom particulier (prendre des décisions, a lua decizii, make decisions) ;
- plusieurs collocations une seule unité comme équivalent (a repara daunele 'réparer dommages-le' = dédommager).

41 Le dictionnaire contient la fréquence d'apparitions des collocations et de leurs propriétés dans plusieurs corpus multilingues disponibles dans les trois langues étudiées : un extrait du corpus parallèle *JRC-Acquis* (Steinberger *et al.*, 2006) et des corpus journalistiques et littéraires, de taille comparable (15 à 20 millions de mots/langue). Les corpus parallèles ont été alignés au niveau propositionnel et au niveau lexical (phrases et mots de la langue source et de la langue cible sont mis en correspondance, Todirascu *et al.*, 2008).

42 À partir des corpus alignés au niveau lexical, nous avons extrait les équivalents de traductions des candidats collocationnels fréquents trouvés dans la langue source, dans les deux sens. Par l'intersection des listes d'équivalents de traduction, nous avons pu établir une liste d'entrées trilingues dans le dictionnaire.

3.3.2. La microstructure du dictionnaire multilingue

43 Notre dictionnaire est composé d'entrées multilingues (Todirascu *et al.*, 2008). Chaque entrée contient des informations morphosyntaxiques et sémantiques pour chaque collocation, dans chaque langue étudiée. Une entrée regroupe des équivalents de traduction qui partagent le même sens (dans l'élément <te>).

44 Pour chaque collocation verbo-nominale, trois types d'informations sont présentes, représentés en langage XML propre :

- les informations concernant le verbe et ses propriétés <v_spec> (la préférence pour la diathèse passive ou impossibilité d'appliquer le passif) ;
- les informations concernant le nom et ses propriétés <n_spec> (préférence pour un déterminant particulier ou pour l'absence du déterminant, pour le singulier ou le pluriel). Pour chaque propriété, sa fréquence est calculée et représentée sous forme de pourcentage (l'attribut *freq*) ;

- une section pour représenter les informations des propriétés morphosyntaxiques propres aux collocations (<c_spec>):

```
<entry id= "1">
   <te lang= "fr">
      <complexitem>
         <construction>mettre+en+berne</construction>
         <v_spec><lemma>mettre</lemma></v_spec>
         <prep>en</prep>
         <n_spec>
            <lemma>berne</lemma>
            <det freq="100">null</det>
            <nb freq="100">sg</nb>
         </n_spec>
         <c_spec>
            <colloc_spec>
               <required_args case="acc"> object </required_args>
               <lexical_restriction compl="object">berne</lexical_res-
triction>
               <colloc_type> complex_predicate</colloc_type>
            </colloc_spec>
            <colloc_documentation>
               <colloc_LL value="2999.854" corpus="ACQ"/>
               <examples><example> ... </example></examples>
            </colloc_documentation>
         </c_spec>
      </complexitem>
   </te>
</entry>
```

45 Les propriétés linguistiques spécifiques à la collocation (<colloc_spec>) sont :

- les arguments de la collocation (élément <required_args>), qui ont une préférence exprimée pour le cas et/ou de la préposition requise :
- *tenir compte* demande un objet direct (introduit par la préposition *de*) <required_args prep = "de"> p-object </required_args> ;
- *pune+în+evidenţă* (‘mettre+en+evidence’) entraîne un objet direct à l’accusatif : <required_args case = "acc"> direct_object </required_args> ;
- la tête lexicale d’un usage restreint de la collocation (élément <lexical_restriction>) : « mettre en berne » n’accepte guère comme compléments d’autres

lexèmes que « drapeau » ou « pavillon ».

- les exemples. Un conteneur <colloc_documentation> permet de donner plusieurs exemples et de renseigner le Log-Likelihood (LL)[16] calculé sur un corpus donné.

46 Cette représentation est adaptée à la classe de collocations que nous avons étudiée, les collocations verbo-nominales, mais elle reste peu réutilisable par d'autres applications. Pour ajouter d'autres catégories de collocations (nominales, adjectivales), et pour compléter la description des collocations, nous avons procédé à une transformation de cette structure de dictionnaire dans le format standardisé LMF.

4. Normalisation du dictionnaire

47 Afin de rendre la ressource développée compatible avec d'autres ressources et outils, nous avons choisi de la transformer selon la norme LMF (ISO ISO-24613:2008). Mis à part les informations de base, nous avons représenté plusieurs catégories d'informations spécifiques :

- nous utilisons <*Lexical Entry*> et <*List Of Components*> pour représenter à la fois des unités lexicales simples et les collocations.
- Les éléments <*Syntactic Behavior*>, <*Subcategorisation Frame*> et <*Syntactic Argument*> représentent le comportement syntaxique des collocations. Les collocations ont des comportements syntaxiques spécifiques, des préférences pour certaines catégories d'arguments.
- Les éléments <*Sense*>, <*Sense Example*> *et* <*Semantic Definition*>. L'information représentée dans <Sense> permet simplement de relier les entrées multilingues par <*Sense Axis*>.
- L'extension pour représenter les expressions polylexicales (*MWEPattern*). Cette extension permet de représenter la structure interne de la collocation et les relations qui s'établissent entre les divers éléments (<*MWELex*>, <*MWENode*>).
- L'extension pour représenter les patrons morphologiques (<*Morphological-Pattern*>, <*TransformSet Process*>, <*GrammaticalFeatures*>, <*SynPattern*>) permet de présenter les spécificités de chaque mot intégrant une collocation. De plus, on regroupe les mots par patrons morphologiques.

48 Nous avons ajouté un élément supplémentaire <*Frequence*> qui représente la fréquence d'apparition d'une configuration syntaxique ou d'une propriété particulière.

49 Notre dictionnaire de collocations contient une liste d'entrées multilingues, les expressions lexicales sont utilisées comme mot vedette dans ce dictionnaire.

4.1. Un exemple de représentation

50 Nous présentons un exemple de représentation pour la collocation *tenir compte*. L'entrée lexicale contient <ListOfComponents> qui font référence aux entrées lexicales du verbe *tenir* et du nom *compte*. Nous avons proposé la propriété « collocation » pour représenter cette classe d'expressions. <Sense> contient un identifiant unique, utilisé dans une balise <SenseAxis> qui définit les correspondances avec les autres langues.

```
<LexicalEntry mwePattern="complex">
   <feat att="collocation" val="prédicateur complexe"/>
   <Lemma><feat att="writtenForm" val="tenir compte"/></Lemma>
   <ListOfComponents>
      <Component entry="E1"/>
      <Component entry="E2"/>
      <Component entry="E3"/>
   </ListOfComponents>
   <Sense id="fra:sens1">
</LexicalEntry>
<LexicalEntry id="E1" morphologicalPatterns="verb1">
   <feat att="partOfSpeech" val="V"/>
   <Lemma><feat att="writtenForm" val="tenir"/></Lemma>
</LexicalEntry>
<LexicalEntry id="E2" morphologicalPatterns="det0">
   <feat att="partOfSpeech" val="D"/>
   <Lemma>
      <feat att="writtenForm" val="— "/>
   </Lemma>
</LexicalEntry>
<LexicalEntry id="E3" morphologicalPatterns="nom1">
   <feat att="partOfSpeech" val="N"/>
   <Lemma><feat att="writtenForm" val="compte"/></Lemma>
   <Frequency><feat att="frequency" val="99"/></Frequency>
</LexicalEntry>
```

51 Chaque élément *Component* fait référence à une entrée lexicale simple. Pour générer les formes qui correspondent aux contraintes imposées par la combinaison de mots, on fait appel à plusieurs patrons morphologiques et syntaxiques, communs à plusieurs entrées. Dans cette représentation, nous avons deux attributs associés aux entrées multiples, « mwepattern » et « morphologicalPatterns » pour les composants. On peut alors décrire une classe de comportements syntaxiques et morphologiques pour les noms (dans l'exemple, pour les noms au singulier sans déterminant) et pour les verbes séparément. Ces comportements peuvent être partagés par plusieurs mots.

```
<MorphologicalPattern id="nom1">
  <feat att="partOfSpeech" val="N"/>
  <TransformSet>
    <Process>
      <feat att="operator" val="addLemma"/>
    </Process>
    <GrammaticalFeatures>
      <feat att="grammaticalNumber" val="s"/>
    </GrammaticalFeatures>
  </TransformSet>
</MorphologicalPattern>
```

52 La balise <MorphologicalPattern> propose des règles de transformations communes pour retrouver toutes les formes possibles du mot (toutes les formes du verbe ou du nom, dans notre cas) : <Process> indique l'action à faire sur le lemme (ajout d'un -s pour le pluriel), <GrammaticalFeatures> indique le cas qui s'applique pour cette action. De plus, dans <GrammaticalFeatures>, on garde l'information de la fréquence de la propriété, fréquence trouvée dans le corpus. Pour d'autres exemples, voir l'annexe B.

53 Pour représenter les propriétés syntaxiques des collocations, nous avons utilisé la balise <MWEPattern> constituée d'un <MWENode> qui indique des propriétés spécifiques à chaque élément de la collocation, alors que <MWELex> indique l'ordre et les séparateurs (espace, tiret etc.) apparaissant entre les éléments. Dans cet exemple, on indique que le nom NP est l'objet direct du verbe VP (la relation est donné par <MWEEdge>) et il est utilisé au singulier. Ainsi, plusieurs

collocations peuvent partager la même configuration interne, représentée une seule fois dans le dictionnaire.

```
<MWEPattern id="complex">
  <MWENode>
    <feat att="syntacticConstituent" val="VP"/>
    <MWELex>
      <feat att="rank" val="1"/>
      <feat att="structureHead" val="yes"/>
    </MWELex>
    <MWEEdge>
      <feat att=”function” val=”directObject”/>
      <MWENode>
        <feat att="syntacticConstituent" val="NP"/>
        <feat att="grammaticalNumber" val="singular"/>
        <MWELex>
          <feat att="rank" val="2"/>
          <feat att="graphicalSeparator" val="space"/>
        </MWELex>
      </MWENode>
    </MWEEdge>
  </MWENode>
</MWEPattern>
```

54 Enfin, pour représenter les informations syntaxiques contextuelles associées avec la collocation, nous utilisons l'élément <*Subcategorisation Frame*>. L'objet direct doit être un groupe nominal et il doit être à l'accusatif, contraintes exprimées par <SyntacticArgument> :

```
<SubcategorizationFrame id="directobject">
  <SyntacticArgument>
    <feat att="id" val="0"/>
    <feat att="syntacticFunction" val="directobject"/>
    <feat att="syntacticConstituent" val="NP"/>
    <feat att="restriction" val="accusative"/>
  </SyntacticArgument>
</SubcategorisationFrame>
```

55 Le regroupement des comportements syntaxiques et morphologiques permet d'éviter les redondances. De plus, les patrons morphologiques permettent la création des formes spécifiques du nom ou du verbe

dynamiquement au moment où on interroge le dictionnaire. Cela représente l'avantage d'avoir une représentation synthétique du dictionnaire.

4.2. Comparaison avec d'autres ressources lexicales en format LMF

56 Parmi les ressources qui sont disponibles en format LMF et qui contiennent des collocations, nous mentionnons LG-LMF (Laporte *et al.*, 2013) et *DuELME* (Odjik, 2013). Pour le premier lexique, il s'agit d'une modélisation des tables à l'aide de <*SyntacticBehavior*> et *de* <*SubcategorisationFrame*>. Pour les expressions figées, LG-LMF propose quatre patrons *MWE Pattern* pour décrire la structure interne des 96 expressions figées présentes dans cette ressource. Dans notre approche, nous avons gardé une représentation par *MWE Pattern* pour décrire la structure interne de la collocation et <*SubcategorisationFrame*> pour décrire le comportement syntaxique en contexte. Pour le deuxième lexique, notre dictionnaire s'approche plus de la structure de *DuELME*, puisque les patrons sont groupés par leur fréquence et on représente aussi les propriétés morphosyntaxiques.

Conclusion et perspectives

57 Dans cet article, nous avons présenté un modèle LMF pour représenter les dictionnaires multilingues de collocations. À partir des informations morphosyntaxiques et sémantiques représentées dans les dictionnaires qui contiennent des collocations et dans notre dictionnaire, nous avons proposé une représentation du dictionnaire en modèle LMF utilisant les modules syntaxique, sémantique et les extensions permettant la description des comportements morphosyntaxiques des collocations. Ce modèle permettra dans le futur l'extension du dictionnaire vers d'autres classes de collocations et une recherche ciblée vers les informations syntaxiques et sémantiques.

BIBLIOGRAPHY

Alonso Margarita, Nishikawa Alfonso, Vincze Orsolya, 2010, « DiCE in the

web: An online Spanish collocation dictionary », dans Granger, Sylvaine, Paquot Magali (dir.), *eLexicograpy in the 21st century : New Challenges, New Applications. Proceedings of eLex 2009*, *Cahiers du Cental* 7, Louvain-la-Neuve, Presses universitaires de Louvain, p. 367-368.

Aristar-Dry Helen, Drude Sebastian, Windhouwer Menzo, Gippert Jost, Nevskaya Irina, 2012, « Rendering Endangered Lexicons Interoperable through Standards Harmonization: The RELISH Project », dans Calzolari Nicoletta *et al.* (dir.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, May 23-25, p. 766-770.

Blumenthal Peter, 2007, « A Usage-based French Dictionary of Collocations », dans Kawaguchi Yuji, Takagaki Toshihiro, Tomimori Nobuo, Tsuruga Yoichiro (dir.), *Corpus-Based Perspectives in Linguistics*, Amsterdam u.a., Benjamins, p. 67-83.

Bouchou Béatrice, Maurel Denis, 2008, « Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres », dans *TAL (Traitement Automatique des Langues)*, *Traitement automatique des langues*, 49(1), p. 61-88.

Braasch Anna, Olsen Sussi, 2000, « Formalised Representation of Collocations in a Danish Computational Lexicon », dans Heid Ulrich et al. (dir.) *The Ninth EURALEX Congress*, *Proceedings*, vol. II, Stuttgart, p. 475-488.

Eckle-Kohler Judith, GUREVYCH Irina, HARTMANN Silvana, MATUSCHEK Michael, MEYER Christian, 2012, « UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF », dans Calzolari Nicoletta *et al.*, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA). Istanbul, Turkey, May 23-25, p. 275-282.

Firth John Rupert, 1956, « Descriptive Linguistics and the Study of English », dans Palmer, F. R. (dir.), *Selected Papers of J.R. Firth*, Indiana University Press, p. 96-113.

Francopoulo Gil, Bel Nuria, George Monte, Calzolari Nicoletta, MonachinI Monica, PET Mandy, Soria Claudia, 2006, « Lexical markup framework (LMF) for NLP multilingual resources », dans *International Committee on Computational Linguistic and the Association for Computational Linguistics – COLING / ACL 2006*, Sydney/Australia.

Francopoulo Gil, George Monte, Calzolari Nicoletta, MonachinI Monica, Bel Nuria, Pet Mandy, Soria Claudia, 2006, « LMF for multilingual, specialized lexicons », dans Zweigenbaum Pierre, Schulz Stefan, Ruch Patrick (dir), *LREC 2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine*. Genova, Italy, ELDA, p. 223-236.

Francopoulo Gil, 2013, *LMF: Lexical Markup Framework* , ISTE / WILEY.

Gledhill Christopher, 2007, « La portée : seul dénominateur commun dans les constructions verbo-nominales », dans *Actes du 1er colloque Res per nomen*, Reims.

Gledhill Christopher, 2009, « Vers une analyse systémique des locutions ver-

bales, constructions verbo-nominales et autres prédicats complexes », dans Banks David (dir.) L*a Linguistique systémique fonctionnelle et la langue française*, Paris, L'Harmattan.

Gross Gaston, 1996, *Les expressions figées en français. Noms composés et autres locutions*, Ophrys, Paris.

Gross Maurice, 1989, « La construction des dictionnaires électroniques » dans *Annales des télécommunications*, t. 44, n° 1-2.

Gross Maurice, 1993, « Les expressions figées en français », dans *L'Information grammaticale*, vol. 59, n° 1, p. 36-41.

Gross Maurice, 1994, « Constructing Lexicon-grammars », dans ATKINS Beryl Sue, ZAMPOLLI Antonio (dir.) *Computational Approaches to the Lexicon*, Oxford University Press, Oxford, p. 213-263.

Hathout Nabil, Sajous Franck, Calderone Basilio, 2014, « GLÀFF, a Large Versatile French Lexicon », dans *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 1007-1012.

Hausmann Franz Josef, 2004, « Was sind eigentlich Kollokationen ? », dans SteyeR, K. (dir.) *Wortverbindungen - mehr oder weniger fest*, Institut für Deutsche Sprache, Jahrbuch 2003/2004, p. 309-334.

Halliday Michael, 1985, *An Introduction to Functional Grammar*, London, Arnold.

Heid Ulrich, Ritz Julia, 2005, « Extracting collocations and their contexts from corpora », dans *Proceedings of Conference on Computational Lexicography and Text Research*, Budapest.

Henrich Verena, HINRICHS Erhard, 2010, « Standardizing Wordnets in the ISO Standard LMF: Wordnet-LMF for GermaNet », dans *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, 2010, p. 456-464.

Hoey Michael, 2005, *Lexical Priming: A New Theory Of Words And Language*, Routledge.

ISO/TC 37/SC 4, « Language resource management - Lexical markup framework (LMF) », http://lirics.loria.fr/documents.html, 2007.

Koehn Philipp, 2005, « Europarl: A Parallel Corpus for Statistical Machine Translation », dans *MT Summit 2005*.

Laporte Eric, Tolone Elsa, Constant Mathieu, 2013, « Conversion of Lexicon-Grammar Tables to LMF: Application to French », dans Francopoulo, Gil (dir.) *LMF: Lexical Markup Framework*, ISTE / WILEY.

Laporte Eric, 2000, « Mots et niveau lexical », dans PierreL, Jean-Marie (dir.), *Ingénierie des langues. Série Informatique et systèmes d'information*, Paris, Hermès, p. 25-49.

L'Homme Marie-Claude, 2003, « Les combinaisons lexicales spécialisées (CLS). Description lexicographique et intégration aux banques de terminologie », dans Grossmann Francis, Tutin Agnès (dir.), *Les collocations : analyse et traitement*, *Travaux et Recherches en Linguistique Appliquée*, p. 89-105.

Manning Christopher D., Schütze Hinrich, 1999, *Foundations of statistical natural language processing*, MIT Press.

Mel'čuk Igor *et al.*, 1984-I, 1988-II, 1992-III, 1999-IV. *Dictionnaire explicatif et combinatoire du français contemporain*, Presses de l'Université de Montréal.

Mangeot Mathieu, Enguehard Chantal, 2013, « Des dictionnaires éditoriaux aux représentations XML standardisées » dans Gala Nuria et Zock Michael (dir.) *Ressources lexicales: contenu, construction, utilisation, évaluation*, John Benjamins p. 24.

Nerima Luka, Seretan Violeta, Wehrli Eric, 2003, « Creating a multilingual collocation dictionary from large text corpora », dans *Proceedings of the Research Notes Session of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest, Hungary*, p. 131-134.

Nerima Luka, Wehrli Eric, 2008, « Generating Bilingual Dictionaries by Transitivity », dans *Proceedings of LREC'2008 Conference*.

Odijk Jan, 2013, « DUELME: Dutch Electronic Lexicon of Multiword Expressions », dans Francopoulo Gil (dir.) *LMF: Lexical Markup Framework*, ISTE / WILEY.

Polguère Alain, 2003, *Lexicologie et sémantique lexicale. Notions fondamentales*, Presses de l'Université de Montréal.

Polguère Alain, 2007, « Lessons from the Lexique actif du français », dans GERDES Kim, REUTHER Tim, WANNER Léo (dir.), *Meaning-Text Theory 2007. Proceedings of the Third International Conference on the Meaning Text Theory*, Klagenfurt, May 20–24, Wiener Slawistischer Almanach, Sonderband 69, München–Wien, p. 397-405.

Rey-Debove Josette, 1971, *Étude linguistique et sémiotique des dictionnaires français contemporains*. The Hague.

Ritz Julia, Heid Ulrich, 2006, « Extraction tools for collocations and their morpho-syntactic specificities », dans *Proceedings of LREC'2006*, Genova, Italia.

Romary Laurent, 2002, *The ISO 16642 document (draft), Version ISO/TC 37/SC 3*, http://www.loria.fr/projets/TMF/tmf.html.

Sagot Benoît, 2010, « The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French », dans *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*, Istanbul, Turkey.

Salmon-Alt Susanne, Akrout Amine, Romary Laurent, 2005, « Proposals for a normalized representation of Standard Arabic full form lexica », *Second International Conference on Machine Intelligence*, Tozeur, Tunisia.

Sinclair John, 1991, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Serasset Gilles, 2012, « Dbnary: Wiktionary as a LMF based Multilingual RDF network », dans Calzolari Nicoletta *et al.*, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, May 23-25.

Silberztein Max, 1993, *Dictionnaires électroniques et analyse automatique de textes : le système* INTEX, Masson, Paris.

Steinberger Ralf, Pouliquen Bruno, Widiger Anna, Ignat Camelia, Erjavec

Tomaz, Tufiş Dan, Varga Daniel, 2006, « The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages ». Proceedings of the LREC, Genova, Italy, 24-26 May.

Todirascu Amalia, Heid Ulrich, Stefanescu Dan, Tufis Dan, Gledhill Christopher, Weller Marion, Rousselot François, 2008, « Vers un dictionnaire de collocations multilingue », dans *Cahiers de Linguistique*, Université de Louvain.

Tutin Agnès, 2010, « Les collocations dans les dictionnaires monolingues spécialisés de collocations », *2e Congrès Mondial de Linguistique Française (CMLF-2010)*.

Verlinde Serge, Selva Thierry, Binon Jean, 2003, « Les collocations dans les dictionnaires d'apprentissage : repérage, présentation et accès », in Grossmann Francis, Tutin Agnès (dir.) *Les collocations : analyse et traitement*, Amsterdam, De Werelt, p. 105-115.

Wiegand Herbert Ernest, 1988, « Wörterbuchartikel als Text », dans *Das Wörterbuch: Artikel und Verweisstrukturen. Jahrbuch 1987 des Instituts für deutsche Sprache*. Hrsg. von Gisela Harras. Düsseldorf (Sprache der Gegenwart LXXIV), p. 90-120.

Williams Geoffrey, 2003, « From meaning to words and back: Corpus linguistics and specialised lexicography », dans *ASp*, vol. 39-40, p. 91-106

Ressources

Analyse et traitement informatique de la langue française — UMR 7118 (ATILF) (2016). *Morphalou* [Lexique]. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr, https://hdl.handle.net/11403/morphalou/v3.1.

APPENDIX

Discussions

Annexe A

Un exemple de représentation multilingue qui utilise *SenseAxis* pour relier des dictionnaires disponibles en plusieurs langues.

```
<Lexicon>
  <LexiconInformation>
    <feat att="language" val="fra"/>
  </LexiconInformation>
```

```
  <LexicalEntry>
    <feat att="partOfSpeech" val="noun"/>
    <Lemma>
      <feat att="writtenForm" val="gonadotrophine"/>
    </Lemma>
    <Sense id="fra#gonadotrophine">
      <SemanticDefinition>
        <feat att="text" val="Lycoprotéine d'un poids moléculaire d'environ
43 000 daltons produite par le syncytiotrophoblaste"/>
      </SemanticDefinition>
    </Sense>
  </LexicalEntry>
  <SenseAxis id="A1" senses="fra#gonadotrophine eng#gonadotropin">
  </SenseAxis>
  <LexiconInformation>
    <feat att="language" val="eng"/>
  </LexiconInformation>
  <LexicalEntry>
    <feat att="partOfSpeech" val="noun"/>
    <Lemma>
      <feat att="writtenForm" val="gonadotrophin"/>
    </Lemma>
    <Sense id="eng#gonadotropin">
      <feat att="domain" val="medicine"/>
      <SemanticDefinition>
        <feat att=”text” val=”that acts on the gonads to promote their
growth and function"/>
      </SemanticDefinition>
    </Sense>
  </LexicalEntry>
</Lexicon>
```

Annexe B

Exemple d'un comportement représenté dans la base

La balise <MorphologicalPattern> suivante décrit le comportement d'un déterminant zéro qui apparaît avec la forme singulier ou pluriel du mot. Cette information est utile pour décrire le comportement de plusieurs collocations incluant les noms *œuvre*, *face*, *compte*, *appel*.

```
<MorphologicalPattern id= "det0">
  <feat att="partOfSpeech" val="D"/>
  <Lemma><feat att="writtenForm" val="-"/></Lemma>
  <TransformSet>
    <Process><feat att="rule" val="//"/></Process>
    <GrammaticalFeatures>
      <feat att="grammaticalNumber" val="— "/>
    </GrammaticalFeatures>
  </TransformSet>
</MorphologicalPattern>
```

NOTES

1 http://www.tei-c.org/

2 http://www.lexicalmarkupframework.org/

3 https://www.collinsdictionary.com/

4 http://www.wordreference.com

5 https://academie.atilf.fr/9/

6 http://atilf.atilf.fr/

7 https://www.dwds.de/

8 Acronyme de Gros Lexique À tout Faire du Français.

9 https://repository.ortolang.fr/api/content/morphalou/3/LISEZ-MOI.html

10 Acronyme de « Laboratoire d'Automatique Documentaire et Linguistique ».

11 Acronyme de « Lexique des formes fléchies du français ».

12 Acronyme de « Dutch Electronic Lexicon of Multiword Expressions ».

13 Le langage à balises XML permet de structurer l'information et de représenter des informations sémantiques. Ainsi, un fragment de texte est annoté avec des balises (indiquées par les caractères <> et un nom). On attribue ainsi une interprétation au fragment marqué par <balise>texte</balise>. Ainsi <personne age= « 20»>Jean Dupont</personne> a été marqué comme étant un nom de personne. « age » est un attribut, une propriété avec une valeur donnée (20) associée à la personne.

14 https://www.ortolang.fr/

15 Par un souci d'uniformité, nous avons choisi une représentation en format XML du modèle LMF.

16 Le Log-likelihood évalue la probabilité que les deux mots (base et collocatif) sont utilisés ensemble plus souvent que le hasard.

ABSTRACTS

Français
L'article présente plusieurs normes utilisées pour la représentation des données dans les dictionnaires électroniques et lexiques destinés aux outils de Traitement automatique des Langues (TAL). Les normes présentées préconisent la représentation des informations linguistiques dans des dictionnaires électroniques selon le modèle TEI (Text Encoding Initiative) et le modèle LMF (Lexical Markup Framework). Nous nous intéressons en particulier aux dictionnaires de collocations à l'adaptation du modèle LMF pour la représentation de ce type de données.

English
We present the issues related to linguistic informations to be represented in the lexical resources. We present several standards for lexical ressources representation, such as TEI (Text Encoding Initiative) and LMF (Lexical Markup Framework). We present a multilingual collocation dictionary and the model we propose to make it compatible with the LMF standard.

INDEX

Mots-clés
collocation, dictionnaire, Lexical Markup Framework, propriété morpho-syntaxique, TEI

Keywords
Lexical Markup Framework, lexical ressource, multilingual collocation dictionary, morpho-syntactic propertiy, TEI

AUTHOR

Amalia Todirascu
Amalia Todirascu est professeur des universités en linguistique française et outillée à l'université de Strasbourg depuis 2016. Elle travaille dans le domaine du

traitement automatique de langues et de la linguistique de corpus, sur le développement d'outils de simplification automatique de textes, de détection automatique de la coréférence, et de création de ressources électroniques (corpus, lexiques).
IDREF : https://www.idref.fr/130431796
ORCID : http://orcid.org/0000-0002-3092-3549
HAL : https://cv.archives-ouvertes.fr/amalia-todirascu