Vers un dictionnaire de collocations multilingue

Amalia Todiraşcu*, Ulrich Heid**, Dan Ştefănescu***, Dan Tufiş***, Christopher Gledhill*, Marion Weller**, François Rousselot****

*LILPA, Université Marc Bloch Strasbourg, France,{todiras, gledhill}@umb.u-strasbg.fr
**IMS Stuttgart, Universität Stuttgart, Allemagne, {uli, wellermn}@ims.uni-stuttgart.de
***ICIA, Academia Română, Bucarest, Roumanie,{danstef, tufis}@racai.ro
****INSA Strasbourg, France, Francois.Rousselot@insa-strasbourg.fr

1. Introduction

Le projet « Collocations en contexte : extraction et analyse contrastive », financé par l'Agence universitaire de la francophonie (AUF), a comme objectif principal le développement d'un système d'extraction semi-automatique de collocations, qui exploite des corpus alignés. Un deuxième objectif est de constituer un dictionnaire multilingue de collocations, à partir des candidats collocationnels proposés par le système d'extraction. Dans cet article, nous présentons seulement les outils d'extraction et la structure du dictionnaire multilingue.

Les collocations constituent un problème central pour les traducteurs, ainsi que pour les systèmes d'aide à la traduction, tant sur le plan de leur contexte d'emploi, que sur celui du choix des composants. Ce qui nous concerne ici est le choix des composants dans des constructions verbonominales (VN) comme *prendre une décision.* Cette expression se traduit par une combinaison à verbe en roumain, *a lua o decizie*, mais en anglais la construction est réalisée par deux verbes *to make/to take a decision* « faire/prendre + décision », et en allemand par le verbe *treffen* « frapper, rencontrer » : *eine Entscheidung treffen.* De même, des verbes « équivalents » ont des distributions syntaxiques très différentes, par ex., *to make* est souvent employé dans des constructions causatives comme *to make good any damages* qui se traduit par *dédommager* en français, mais par *a compensa daunele* en roumain.

Pour ces raisons, les collocations ont fait l'objet de nombreuses études en lexicographie, en linguistique de corpus et en traitement automatique des langues (TAL). Ces domaines se concentrent sur la définition formelle des catégories de collocations, ainsi que sur l'identification de leurs propriétés syntaxiques, sémantiques et pragmatiques (voir Grossmann et Tutin : 2003 et L'Homme : 2003). Selon l'approche traditionnelle en lexicographie (voir Hausmann : 2004 et Manning et Schütze : 1991) la collocation est une relation binaire, asymétrique, entre deux lexèmes (une *base* et un *collocatif*) : le *collocatif* dépend de la *base* pour son interprétation (ainsi dans la construction *prendre une décision* l'acception spécifique du V *prendre* serait déterminé par le N *décision*). Cependant, dans ce projet, nous adoptons

plutôt l'approche « contextuelle » des linguistes britanniques (voir Halliday :1985, Gledhill : 2000 et Williams : 2003). Dans la perspective de Halliday (1985), la collocation est une corrélation sémantique que l'on peut observer dans un contexte donné entre un lexème (un *noyau*) et ses partenaires lexicaux (ses *collocateurs*) ou grammaticaux (ses *colligations*). Cette approche souligne surtout le fait que le phénomène collocationnel ne se réduit pas à une relation binaire entre deux éléments lexicaux, mais concerne également les schémas lexico-grammaticaux plus étendus dans lesquels ces lexèmes ont un rôle à jouer.

Vu les difficultés d'utilisation des collocations, beaucoup de projets de recherche se sont orientés vers la création de dictionnaires électroniques (voir Blumenthal : 2007 , Mel'cuk *et al.* : 1999) ou le développement d'outils d'extraction automatique des collocations. La plupart des dictionnaires électroniques proposent des listes de collocations « équivalentes », souvent dans la perspective de l'apprentissage des langues étrangères (p. ex., la Base lexicale du français (BLF)[1] et Verlinde *et al.* : 2003). Peu de ressources proposent une description complète des propriétés morpho-syntaxiques des collocations (voir Zingle et Brobeck-Zingle : 2003, qui se concentrent sur l'interaction entre collocations et sous-catégorisation syntaxique). Par contraste, d'autres dictionnaires proposent des données techniques sans grande utilité pour l'usager non averti. Nous pensons notamment au *Trésor de la langue française informatisé* (TLFI), qui associe, pour chaque entrée lexicale, les collocations dans lesquelles le mot intervient ou le DiCO (voir Polguère : 2006 et Mel'čuk *et al.* : 1994), qui propose une caractérisation des collocations à l'aide des « fonctions lexicales » qui font la particularité de cette dernière approche. Enfin, il est difficile d'adapter les dictionnaires existants pour un outil de TAL p. ex., les tables du Lexique-Grammaire développé au Laboratoire d'automatique documentaire et linguistique (LADL) (voir Gross : 1993) pour le français. Si nous disposons de dictionnaires de collocations pour le français, pour l'anglais et pour l'espagnol, ce n'est pas le cas pour le roumain, ni pour l'allemand.

Pour alimenter les dictionnaires électroniques, des outils d'extraction de collocations à partir des corpus ont été développés. Ces outils s'appuient, soit sur des méthodes statistiques (voir Evert : 2005), soit sur des techniques syntaxiques (voir Tutin : 2004 et Sereţan *et al.* : 2004). Si les méthodes statistiques identifient beaucoup de candidats non valides (la précision est alors faible) ; les méthodes syntaxiques nécessitent beaucoup de ressources linguistiques (la précision est meilleure, mais le rappel est faible et les ressources nécessaires sont très complexes). Pour trouver un bon compromis entre la précision et le rappel, des méthodes

[1] Katholieke Universiteit Leuven, - Institut interfacultaire des langues vivantes, base disponible en ligne http://ilt.kuleuven.be/blf/.

hybrides ont été développées (voir Smadja et McKeown : 1990). Dans la même catégorie d'approches, notre outil d'extraction applique d'abord des méthodes statistiques et utilise ensuite les informations morpho-syntaxiques sur le comportement linguistique des collocations pour extraire des candidats. Nous privilégions une approche faisant appel à peu de connaissances linguistiques qui exploite des corpus étiquetés du français et du roumain. Pour l'allemand, une approche plus "coûteuse" semble nécessaire. De plus, comme notre objectif est de construire un dictionnaire multilingue, nous explorons des corpus alignés pour sélectionner des candidats collocationnels.

Nous présentons la méthodologie adoptée dans le cadre de ce projet ainsi que les corpus étudiés pour définir le modèle linguistique que nous avons utilisé pour identifier des filtres linguistiques pour le repérage automatique des collocations. Les outils d'extraction monolingues et bilingues sont présentés en détail, ainsi qu'une série de données extraites à partir des corpus alignés et le modèle du dictionnaire créé.

2. La méthodologie

Pour atteindre l'objectif d'une extraction de collocations et de leurs contextes appropriés , nous avons adopté l'hypothèse de travail suivante : les collocations ont un comportement syntaxique propre (validée pour l'allemand par (Heid et Ritz : 2005) et (Ritz et Heid : 2006)). Selon cette approche, le contexte permet d'identifier les propriétés spécifiques pour chaque classe de collocations : le cas du complément direct ou indirect, le nombre, la détermination du *noyau* ou du *collocateur*[2] . Ces informations seront utilisées pour compléter le dictionnaire, en exploitant des corpus monolingues et multilingues. Par rapport à d'autres outils d'extraction automatique de collocations qui s'appliquent à des corpus monolingues, nous avons exploité également des corpus alignés pour identifier des candidats collocationnels communs.

[2] Nous adoptons les termes *noyau/collocateur* qui sont plus appropriés à une approche empirique qui extrait des candidats à partir d'une analyse statistique

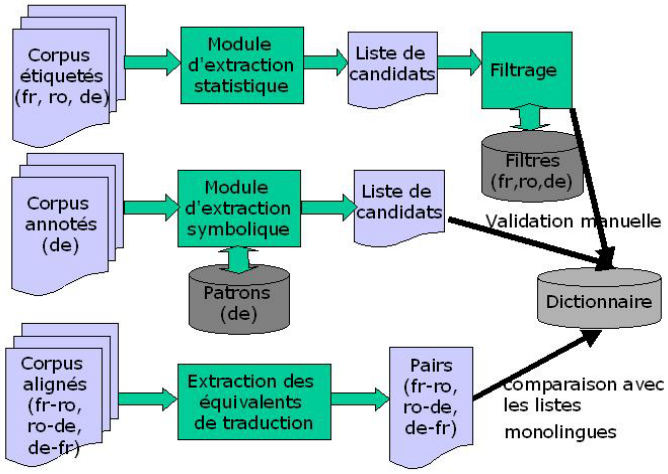

Fig.1. L'architecture du système

La méthodologie que nous avons adoptée ici comporte donc plusieurs étapes de travail, et combine des méthodes d'extraction monolingues et multilingues. D'abord, nous avons procédé à une analyse linguistique d'une classe particulière des collocations (les constructions verbo-nominales (VN)) pour le roumain, le français et l'allemand, à partir des corpus monolingues et multilingues. Les corpus multilingues ont été alignés au niveau de la phrase et des mots (il en sera question à la section 5). À partir des propriétés morpho-syntaxiques contextuelles identifiées dans les corpus, nous avons défini des filtres linguistiques pour le repérage automatique des classes de collocations, pour chacune des langues étudiées. Nous avons appliqué l'extraction statistique des candidats collocationnels, suivi d'un filtrage linguistique, pour chaque langue. Pour l'allemand, nous avons également évalué une méthode d'extraction symbolique, qui s'applique à des corpus annotés syntaxiquement. Nous avons alors établi une liste de candidats collocationnels (validés manuellement) pour chaque langue. Nous avons identifié des collocations communes aux langues étudiées, à partir des corpus alignés (roumain, français, allemand). L'anglais a été utilisé pour générer les alignements lexicaux roumain-français, français-allemand et roumain-allemand. Nous avons alors sélectionné les candidats communs, mais aussi les candidats trouvés dans une seule langue, pour constituer le dictionnaire multilingue de collocations.

2.1. Les corpus

Pour identifier les éléments du dictionnaire multilingue, nous avons exploité des corpus alignés. Nous utilisons un corpus multilingue, *AcquisCommunautaire-ACQ* (voir Steinberger *et al.* : 2006), disponible en 21 langues européennes, dont le français, le roumain, l'allemand et l'anglais. Il s'agit d'un corpus très spécialisé, portant sur la législation européenne publiée depuis 1950 et caractérisé par un style spécifique du langage juridico-

administratif. Nous avons sélectionné un ensemble commun de documents dans les trois langues, ainsi qu'en anglais (environ 16 millions de mots pour chaque langue).

Vu que le corpus multilingue est spécialisé, nous avons également utilisé des corpus monolingues, pour valider les résultats de l'analyse linguistique. Il s'agit de journaux en français (*Le Monde* 2004 et *Le Monde diplomatique* 1980-1998 : environ 44 millions de mots) et en allemand (*Stuttgarter Zeitung* et *Frankfurter Rundschau* 1992-1993 : environ 76 millions de mots). Pour le roumain, le corpus monolingue utilisé est composé de plusieurs textes journalistiques, de romans et de textes médicaux (10 millions de mots). Pour l'anglais, nous disposons du corpus British National Corpus Baby (BNC Baby) (environ 4 millions de mots) et du British National Corpus complet. Ces corpus ont été étiquetés et lemmatisés à l'aide du TreeTagger (voir Schmid : 1994) pour l'allemand, le français et l'anglais (*ACQ* et les corpus monolingues), et de l'outil Tokenzing Tagging Lemmatizing Free Running Text (TTL) pour le roumain (voir Ion : 2007). Vu que les outils d'étiquetage ont été entraînés sur des textes journalistiques, et que nous avons identifié beaucoup d'erreurs d'étiquetage et de lemmatisation, nous avons validé et corrigé manuellement les erreurs sur les corpus monolingues et multilingues.

Le corpus *ACQ* est également aligné au niveau de propositions et des mots (voir Tufiş *et al.* : 2005). Ce corpus est utilisé pour établir une liste de collocations communes aux trois langues, mais aussi pour identifier des collocations spécifiques d'une langue. Nous avons sélectionné un ensemble de propositions alignées qui respectent la condition suivante : chaque proposition a été traduite par une seule proposition dans la langue cible. Nous avons validé manuellement les alignements lexicaux et nous avons utilisé cette base de propositions alignées pour extraire la liste de collocations communes. Les outils d'extraction monolingues, ainsi que les outils d'alignement seront présentés dans les sections suivantes.

3. Les constructions verbo-nominales (VN)

Dans cette section, nous adoptons la définition des « constructions verbo-nominales » (VN) proposée par Gledhill (2007). Alors que les linguistes formalistes (voir Gross : 1993) ont reconnu plusieurs sous-catégories de « verbes supports », « prédicats légers », etc., Gledhill (2007) propose un seul critère pour établir une catégorie homogène de « constructions VN ». Il base son analyse sur le modèle systémique-fonctionnel de Halliday (1985), qui suppose que plusieurs niveaux de signification participent simultanément à la composition d'un message : i) fonction syntaxique (prédicateur, complément) ; ii) structure lexicale (groupe verbal, groupe nominal), et iii) transitivité sémantique (procès, participant).

C’est au niveau de la transitivité sémantique que nous pouvons distinguer les simples « co-occurrences VN », comme (1) *Pat a fait un gâteau*, des « constructions VN », comme (2) *Pat a fait une remarque.* La seule différence, alors, entre (1) et (2) est que, dans (2), le complément exprime la **portée**, une forme de métaphore grammaticale où le procès sémantique du prédicat entier est désigné ou délimité par un élément qui n’est pas le prédicateur (le V lexical). Le verbe *faire* porte dans (1) et (2) la morphologie associée à un prédicateur. De même, sur le plan sémantique, *faire* a toujours un rôle sémantique : dans (1) *faire* réfère à un procès « Matériel » qui dépend de la présence de deux participants appropriés (un sujet avec un rôle sémantique d’« Agent » et un complément avec un rôle d’« Affecté »). Par contre, dans (2) *faire* réfère à un procès « Mental », où le complément désigne la « Portée » de ce procès (de communication) alors que le sujet exprime le rôle d’« Expérienceur ». Il est aussi important de noter que la portée n’est pas toujours exprimée par un complément autonome comme dans (1) et (2), c’est-à-dire des **prédicats complexes**. Le terme s’applique également aux N qui sont intégrés dans un groupe verbal (GV), ce que Gledhill (2007) appelle un **prédicateur complexe**. Ainsi, dans (3) *Pat fait peur aux électeurs*, nous avons la même configuration de rôles sémantiques que dans (2), avec une seule modification : *Pat* est conçu comme un « Phénomène », et *aux électeurs* exprime l’ « Expérienceur ». À part cette différence structurelle, c’est surtout la notion de « portée » qui sert à regrouper des locutions comme *faire peur* et des constructions VN plus libres comme *faire une remarque.*

catégorie	**attributs pertinents**
V1	<possible par système morphologique>
V2	marqueurs de cas pour le roumain et l’allemand
V3	identifiable à l’aide de patrons morpho-syntaxiques
N1	détermination (patrons)
N3	modification du nom (adjectif, clause relative), nombre
N4	<possible par système morphologique>

Tableau 1. Les propriétés des constructions VN et les possibilités de les extraire automatiquement

Pour identifier les prédicats complexes et les prédicateurs complexes de manière semi-automatique, nous avons identifié un faisceau de propriétés morpho-syntaxiques, basé sur les propriétés V1-4 et N1-4 identifiées par Gledhill (2007). Ainsi, les propriétés des constructions

VN propres aux verbes sont : (V1) être équivalent à un seul verbe ; (V2) la valence (les compléments directs, indirects, etc.) ; (V3) le passif ; (V4) l'aspect. En ce qui concerne les propriétés caractéristiques du nom, nous étudions : (N1) le déterminant (l'absence ou la présence de l'article) ; (N2) le clivage ; (N3) la modification par une clause relative (qui n'est pas possible pour les prédicateurs complexes) ; (N4) la conversion vers un groupe nominal. Vu que les propriétés V4 et N2 sont trop spécifiques pour être utilisées fréquemment dans les corpus, nous avons utilisé seulement les propriétés présentées dans le tableau 1.

propriétés	**valeurs pour le français**	**valeurs pour le roumain**	**valeurs pour l'allemand**
Nombre	singulier, pluriel	singulier, pluriel	singulier, pluriel
Cas		nom., gén, dat., acc.	nom., gén, dat., acc.
détermination (articles)	aucun, défini, indéfini, possessif, démonstratif	aucun, défini, indéfini, possessif, démonstratif	aucun, défini, indéfini, possessif, démonstratif, quantitatif
modification du nom	adjectif, clause relative	adjectif, par génitif ou par groupe prépositionnel	adjectif, par génitif, par groupe prépositionnel
fusion entre préposition et article défini	-	-	oui, non
négation dans le déterminant quantitatif	-	-	oui, non

Tableau 2. Les valeurs des propriétés utilisées pour le repérage automatique et pour alimenter le dictionnaire

Aucune de ces propriétés n'est suffisante pour identifier d'une manière unique la classe des constructions, mais certaines sont identifiables dans le jeu d'étiquettes utilisé dans l'annotation morphosyntaxique ou en faisant appel à des outils de flexion. Parmi les propriétés utilisées, nous identifions les marqueurs de cas, le nombre, les prépositions, les déterminants (voir Todirascu *et al.* : 2008). Nous présentons ici les propriétés et leurs valeurs qui sont utilisables pour la définition de filtres linguistiques (voir tableau 1) et qui vont être stockées dans le dictionnaire (voir tableau 2).

Si la classe dont fait partie le candidat peut être validée seulement manuellement, les propriétés morpho-syntaxiques sont identifiées automatiquement dans un corpus étiqueté, ce qui produit de « bons candidats ». Les prédicateurs complexes sont caractérisés par une invariabilité morpho-syntaxique prononcée (absence du déterminant, pas de passivation possible, etc.), alors que les prédicats complexes sont plus flexibles : le complément est plus autonome sur un plan syntagmatique.

Ces propriétés ont été utilisées pour définir des filtres linguistiques utilisés par les outils d'extractions monolingues, dont les résultats sont présentés dans la section suivante.

4. Les outils d'extraction monolingues

Nous avons développé et évalué des outils d'extraction monolingue des collocations, considérant, comme nous avons indiqué plus haut, que les collocations VN sont à la fois des *cooccurrences* et des *constructions*. Nous avons ainsi développé un outil d'extraction hybride, pour les trois langues étudiées. Compte tenu des propriétés spécifiques de l'allemand, nous avons également développé une méthode d'extraction symbolique, à base des corpus annotés.

4.1. La méthode hybride

Nous avons développé un outil d'extraction qui applique une méthode statistique, indépendante des langues, suivie d'un filtrage linguistique, à base de patrons. Cet outil s'applique aux trois langues, même si l'approche symbolique s'avère mieux adaptée à l'allemand.

4.1.1. La méthode statistique

Cette méthode s'applique à des corpus étiquetés et lemmatisés et identifie des paires de mots qui ne sont pas toujours adjacents (voir Todirascu *et al.* : 2007), dont on peut évaluer les propriétés suivantes :

a) la distance entre les deux mots est relativement stable : nous utilisons la distance et la moyenne (voir Smadja et McKeown : 1990) ;

b) la cooccurrence des deux éléments est statistiquement significative : les deux mots apparaissent ensemble plus souvent que par hasard (mesuré par une mesure d'association comme le LogLikelihood Ratio Test (LL) (voir Dunning : 1993).

noyau	**collocateur**	**LL**	**art.**	**nombre**	**cas**	**classe**
aduce/ 'apporter'	atingere/'atteinte'	51567.34864	-	sing.	datif	A
înlocui/'remplacer'	text/'texte'	43992.3067	déf	sing., pl.	acc.	B
intra/'entrer'	vigoare/'vigueur'	42527.03736	-	sing.	acc. (în/'en')	A
avea/'avoir'	tratat/'traité'	32050.11219	déf	sing., pl.	acc.	non
face/'faire'	obiectul/'obiectul'	30729.47663	déf	sing.	datif	A
modifică/'modifier'	regulamentul/ 'règlement-le'	29141.39454	déf, -	sing., pl.	acc. (la/'à', din/'de')	B
lua/'prendre'	considerare/ 'considération'	27062.0349	-	sing.	nom.	A

Tableau 3. Les paires V-N roumaines les plus fréquentes, identifiées par un ensemble de propriétés morpho-syntaxiques plus ou moins invariables (nombre, déterminant, cas du complément indirect). A- prédicateur complexe ; B – prédicat + complément.

L'outil propose en sortie une liste de cooccurrences des V et des N, triée par le score LL, ainsi que les contextes de chaque paire. Employant cette méthode sur le corpus ACQ roumain et français, nous avons obtenu les collocateurs candidats présentés dans les tableaux 3 et 4. Nous trouvons ainsi un ensemble de collocations commun aux deux langues (*tenir compte/ține cont ; prendre des décisions/lua decizii ; faire l'objet/face obiectul*). Beaucoup de ces collocations, comme *a intra în vigoare/entrer en vigueur, adoptat la Bruxelles/fait à Bruxelles* correspondent à des expressions spécifiques du langage juridique du corpus ACQ.

noyau	**collocateur**	**LL**	**art.**	**nombre**	**cas**	**classe**
viser	article	80514.1578711753	déf.	sing.	acc.	-
modifier	lieu	65160.6523994904	-, indéf.	sing.	acc.	-
instituer	communauté	60852.4413396388	déf.	sing.	acc.	-
avoir	lieu	45860.0441745073	-	sing.	acc. (de)	A
remplacer	texte	45714.492150083	déf.	sing.	acc.	B
faire	Bruxelles	41099.1449169632	-	-	-	-
faire	objet	39398.1847211485	déf.	sing.	acc.(de)	A

Tableau 4. Les cooccurrences VN françaises, leurs propriétés contextuelles et les classes (A- prédicateur complexe ; B- prédicat complexe)

Nous pouvons constater la présence des propriétés morpho-syntaxiques mentionnées en section 3 : l'absence du déterminant (-), la préférence pour le nombre singulier et pour le cas accusatif, pour les prédicateurs complexes dans les deux langues (*cf.* tableau 4, qui présente les résultats de l'extraction pour le français).

Les contextes, étiquetés et lemmatisés, ordonnés par fréquence, servent à identifier les propriétés morpho-syntaxiques spécifiques et à appliquer les patrons de filtrage pour éliminer les candidats. Nous constatons souvent le N *vedere* 'vue' sans déterminant et au singulier (identifié par l'étiquette **nsrn**) (99,97 %), accompagné de la préposition *în* « en ». Il s'agit ici d'un candidat pour la classe prédicateur complexe.

```
Mot1=avea Mot2=vedere     dist=2                    LL=25533.14309
având/vg/avea în/s/în vedere/nsrn/vedere            17786
avut/vp/avea în/s/în vedere/nsrn/vedere             130
aibă/v3/avea în/s/în vedere/nsrn/vedere             128
avea/vn/avea în/s/în vedere/nsrn/vedere             51
au/va3p/avea în/s/în vedere/nsrn/vedere             41
au/v3/avea în/s/în vedere/nsrn/vedere               31
având/vg/avea in/nsn/in vedere/nsrn/vedere          11
avea/v3/avea în/s/în vedere/nsrn/vedere             6
aibă/v3/avea o/tsr/un vedere/nsrn/vedere            4
avea/vn/avea o/tsr/un vedere/nsrn/vedere            1
avem/v1/avea în/s/în vedere/nsrn/vedere             1
```

Par contre, *marque* est plus variable (déterminant défini 65,87 %, déterminant indéfini 34,13 %, pluriel 5,55 %, singulier 94,45 %). Notre système ne nous permet pas de décider si ce candidat est une construction ou une cooccurrence VN, mais nous pouvons garder les valeurs les plus fréquentes (pour le singulier).

```
Mot1= porter mot2=marque  dist=2                    LL=2764.12131446324
portant/porter/ver:ppre la/le/det:art marque/marque/nom    66
portent/porter/ver:pres la/le/det:art marque/marque/nom    60
porter/porter/ver:infi une/un/det:art marque/marque/nom    43
porter/porter/ver:infi la/le/det:art marque/marque/nom     42
portant/porter/ver:ppre une/un/det:art marque/marque/nom   16
porter/porter/ver:infi des/du/prp:det marques/marque/nom   14
portent/porter/ver:pres une/un/det:art marque/marque/nom   13
```

Ces contextes sont utiles pour établir une liste de valeurs possibles pour chaque propriété morpho-syntaxique et pour compléter le dictionnaire (*cf.* section 6).

4.1.2. Le filtrage

Les candidats identifiés par l'outil d'extraction statistique monolingue seront filtrés à l'aide des propriétés morpho-syntaxiques identifiées dans les contextes : la variabilité du déterminant, la préférence pour une classe de prépositions, etc., mais aussi à l'aide de règles heuristiques : la distance entre le verbe et le nom est supérieure à 5, les signes de ponctuation ne peuvent pas intervenir entre les deux mots, un nombre trop important de prépositions qui est employé entre le verbe et le nom, etc. Les filtres éliminent successivement des contextes qui ne peuvent pas être associés à des candidats pertinents. L'élimination de tous les contextes d'une paire implique l'élimination de ce candidat.

Suivant la fonction syntaxique jouée par le verbe et le nom, nous avons identifié plusieurs catégories de candidats qui ne peuvent pas être valides (voir Todirascu *et al.* : 2008) : prédicat + sujet, prédicat + modificateur, groupes nominaux, etc. Nous avons défini des filtres qui éliminent certaines classes, mais pour certaines fonctions il est nécessaire de faire appel à des corpus annotés syntaxiquement (p. ex., prédicat + modificateur). Pour définir les filtres, nous avons utilisé un langage de description propre à l'outil, qui utilise le jeu d'étiquettes MSD pour le roumain et le jeu d'étiquettes de TreeTagger pour le français :

Exemples de filtres pour éliminer des candidats qui sont :

a) des groupes nominaux en français : (*jour suivant*, *procédure prévue*, *Nations unies*), identifiables par `NOM VER:pper` où `NOM` – substantif ; `VER:pper` – verbe participe passé ;

b) des structures prédicat + complément indirect en roumain (*adresează statelor membre*) sont reconnus par le patron `V3 NxRO NxRO` où `V` – verbe 3e personne ; `NxRO` – substantif cas génitif.

4.2. Une approche symbolique pour l'extraction de candidats collocationnels allemands

4.2.1. Motivation pour une approche symbolique : l'ordre des mots en allemand

À la différence de l'anglais et du français, langues essentiellement configurationelles, l'allemand présente un ordre de constituants relativement libre. Néanmoins, environ 20 % seulement des syntagmes nominaux allemands ont une forme morpho-syntaxique dont on peut déduire le cas et le nombre de façon univoque (voir Evert : 2004).

Par conséquent, l'identification des valeurs de cas et de la valence des candidats collocationnels (*cf.* critère V2 du tableau 1) ne peut pas reposer uniquement sur les marques morphologiques des mots, ni sur leur ordre. La grammaire de l'allemand connaît trois modèles de placement du verbe :

- verbe en tête de phrase (questions et certaines conditionnelles) :
  Erfüllt die betreffende Person zu einem bestimmten Zeitpunkt nicht die Voraussetzungen [..., so...] (*ACQ*-DE) 'si la personne en question ne remplit pas, à un moment déterminé, les conditions [..., alors ...]' ;
- verbe en deuxième position, phrase déclarative standard :
  Die betreffende Person erfüllt [...] nicht die Voraussetzungen 'la personne en question ne remplit pas [...] les conditions' ;
- verbe à la fin de la phrase (phrase subordonnée) :
  wenn die betreffende Person [...] die Voraussetzungen [...] nicht erfüllt, so... 'si la personne en question [...] ne remplit pas les conditions [..., alors]'.

Pour l'extraction de candidats collocationnels du type VN, ces différents modèles d'ordre des constituants ont des conséquences majeures.

- Dans le modèle de la phrase complétive (verbe à la fin de la phrase), le groupe nominal et le complexe verbal (*die Voraussetzungen erfüllt*) se trouvent adjacents ; seuls des groupes nominaux au génitif, attachés au groupe nominal noyau de la collocation, ainsi que certains adverbes ou groupes prépositionnels peuvent intervenir entre les deux éléments collocationnels.

- Par contre, les modèles à verbe en tête de phrase et à verbe en seconde position impliquent qu'un ou plusieurs constituants peuvent se trouver entre le syntagme nominal et le verbe qui font partie de la collocation.

Il est donc bien moins facile d'appliquer des mesures de distance et/ou de simples filtres pour extraire les collocations. Au passif (construit avec un auxiliaire) ainsi que dans les constructions à auxiliaire modal, l'emplacement des deux éléments de la collocation est pourtant identique à celui de la complétive.

Dans nos outils, nous essayons de tenir compte de ces phénomènes : pour une extraction dont l'objectif est d'arriver à une précision élevée, nous ne considérons que les modèles syntaxiques où les deux éléments des collocations se trouvent dans des constituants adjacents.

4.2.2 Principes de l'extraction symbolique

Pour extraire des candidats collocationnels, nous faisons appel à une série de patrons d'extraction formulés sur la base des annotations présentes dans les corpus, c.-à-d. les étiquettes de catégorie[3], les lemmes, ainsi que les marques de début et fin de chunks (syntagmes sans recursion après leur tête) et l'annotation disjonctive des propriétés morphosyntaxiques des mots. Les patrons modélisent les parties pertinentes des phrases subordonnées et/ou passives.

L'application de chaque patron produit des candidats phrases, dont les lemmes du verbe et du nom, ainsi que les propriétés morpho-syntaxiques pertinentes sont identifiées et notées, dans une base de données (voir Heid et Weller : 2008 et Ritz et Heid : 2006).

Ceci nous permet, une fois que nous avons analysé une quantité suffisante de phrases, de comparer les propriétés des phrases extraites, selon les candidats collocationnels, c.-à-d. selon les couples de verbe et nom. Parmi les propriétés ainsi analysables se trouvent les critères suivants mentionnés plus haut, au tableau 1 : V3 (passivation possible ou non), identifié à travers les patrons spécifiques pour le passif ; N1 (déterminant), identifié dans le syntagme nominal ; N3 (modification), identifié dans le syntagme nominal (jusqu'à présent, nous n'identifions que les modifieurs adjectivaux) ; V2 (valence), identifiée, en partie, au moyen des marques de cas ou des prépositions. L'ordre des deux types de connaissances, linguistiques et statistiques est donc inversé par rapport aux outils décrits plus haut.

[3] Le jeu d'étiquettes STTS est utilisé par TreeTagger (voir Schmid : 1994).

4.2.3 Classification des candidats : prédicateurs complexes vs prédicats complexes

La distinction postulée plus haut entre prédicateurs complexes et prédicats complexes est évidemment aussi valable pour l'allemand. De plus, en allemand, les préférences morpho-syntaxiques semblent indicatives des prédicateurs complexes : ceux-ci n'acceptent guère de modification, préfèrent un seul type de déterminant (souvent l'article zéro), etc.

F	**n_lemme**	**v_lemme**	**type-dét.**	**nombre**	**actif/passif**	**type-mod.**
5	Rechnung	ausstellen	déf.	sing.	passif	vfinal
4	Rechnung	ausstellen	indéf.	sing.	actif	vfinal
4	Rechnung	ausstellen	déf.	sing.	actif	vfinal
1	Rechnung	ausstellen	déf.	pl.	actif	vfinal
1387	Rechnung	tragen	nul	sing.	actif	vfinal
262	Rechnung	tragen	nul	sing.	passif	v-1
136	Rechnung	tragen	nul	sing.	passif	vfinal

Tableau 5. Exemples de candidats extraits du corpus *ACQ* allemand, avec leurs préférences morpho-syntaxiques observées

Dans le tableau 5, nous proposons deux candidats extraits du corpus *ACQ*, tous les deux à noyau *Rechnung* (compte). Le tableau contient une ligne par type de préférence morpho-syntaxique, et les colonnes contiennent les données observées, à savoir (de gauche à droite) : la fréquence absolue du modèle préférentiel observé (f), le lemme du nom et du verbe, le type de déterminant observé (type-dét.), le nombre du syntagme nominal, la diathèse (actif/passif), ainsi que le modèle d'ordre des constituants (v-1, v-2, vfinal).

Or, l'analyse du tableau 5 montre certaines différences entre les deux entrées *Rechnung ausstellen* (*établir un compte, présenter une facture*) et *Rechnung tragen* (*tenir compte*) : *Rechnung ausstellen* a été rencontré aussi bien à l'actif qu'au passif, au singulier et au pluriel, avec un article défini aussi bien qu'avec un article indéfini. Cette variabilité coïncide avec une classification manuelle de *Rechnung ausstellen* en tant que construction à prédicat plus complément. D'autre part, les données obtenues pour *Rechnung tragen* présentent une préférence marquée pour l'absence d'article et pour le singulier, même si le candidat apparaît aussi bien à l'actif qu'au passif. Nous avons tendance à classifier *Rechnung tragen* parmi les prédicateurs complexes, tout comme *tenir compte*/*a ţine cont*.

candidat	**A : Vfinal**	**P : V-1**	**P : Vfinal**	**P : V-2**

Auffassung vertreten (être d'avis)	1321	53	97	48
Bezug nehmen (faire référence)	783	439	492	0
Rechnung tragen (tenir compte)	2287	481	492	0
Gebrauch machen (faire usage)	2095	216	430	0
Sorge tragen (assurer)	241	31	43	0

Tableau 6. Quelques candidats collocationnels extraits de textes allemands avec leur distribution au passif à travers les trois modèles d'ordre des constituants (V-1, V-2, Vfinal)

Même si la présence de préférences suggère une première classification, nous sommes loin d'une classification automatique, parce que, d'une part, nous ne disposons pas d'une quantité suffisante de données pour tous les candidats collocationnels, et, d'autre part, parce que certains groupes composés d'un prédicat et de son complément ont également des préférences morpho-syntaxiques. Pourtant, un moyen intéressant d'identifier les prédicateurs complexes réside dans la prise en compte de leur comportement à la forme passive, par rapport aux trois modèles d'ordre des mots mentionnés plus haut : dans le tableau 6, nous présentons les fréquences absolues observées à l'actif (A:) et au passif (P:), pour certains candidats à haute fréquence ; pour le passif, nous distinguons les trois modèles d'ordre des mots (V-1, V-2, Vfinal).

Le tableau 6 montre clairement que certains candidats, bien que très fréquents au passif, ne se retrouvent jamais sous la forme V-2 du passif. Ceci semble s'expliquer par le fait que les éléments nominaux des prédicateurs complexes ne peuvent pas se trouver dans la position à gauche du verbe (ou auxiliaire) flechi. Pour ces exemples, cette distribution semble coïncider très clairement avec la classification en tant que prédicateur complexe.

5. L'outil d'extraction bilingue

Nous avons utilisé un corpus aligné, à la fois pour réaliser une analyse contrastive des propriétés des constructions VN, et pour choisir des candidats collocationnels à inclure dans le dictionnaire.

5.1 Les méthodes d'alignement

Pour ce projet, nous avons aligné les corpus ACQ au niveau des phrases et au niveau lexical. Les relations qui sont établies entre diverses unités textuelles alignées (des paragraphes, des phrases ou des mots) nous aident à identifier les candidats collocationnels intéressants à inclure dans le dictionnaire. Dans notre cas, l'alignement lexical (au niveau des mots et des unités lexicales) est utilisé pour trouver des constructions VN équivalentes dans

les trois langues du projet (français, allemand, roumain). Puisque nous ne disposons pas de ressources suffisantes pour l'alignement (lexiques, corpus annotés), pour toutes les paires de langues étudiées dans le projet, nous avons utilisé l'anglais comme langue pivot pour générer les autres alignements. Ainsi, les alignements sont initialement réalisés à partir des paires anglais-français, anglais-allemand, anglais-roumain. À partir de ces alignements, nous générons les paires français-allemand, français-roumain et allemand-roumain (voir Tufiş et Koeva : 2007).

Avant l'alignement lexical, il est nécessaire de passer par une étape d'alignement propositionnel (qui n'est pas dépendante des langues source et cible). Nous avons adapté une méthode d'alignement propositionnel qui est très efficace (voir Moore : 2002), pour prendre en compte également les cas d'alignement où une phrase est traduite par plusieurs phrases dans la langue cible (voir Ceauşu *et al.* : 2006). Cette méthode hybride fonctionne en trois étapes : d'abord, l'application utilise des méthodes qui s'appuient sur la longueur de la proposition et l'alignement géométrique. La deuxième étape filtre les candidats pour retrouver les alignements qui sont sûrs et, sur cette base, l'algorithme construit une liste d'équivalents de traduction. La troisième étape utilise les listes d'équivalents de traduction pour corriger les autres alignements. Nous utilisons un classifieur de type vecteur support, indépendant de la langue. Son entraînement doit se faire à l'aide d'une partie restreinte du corpus (200 propositions, alignées manuellement). L'application utilise la distribution LIBSVM (voir Fan *et al.* : 2005), avec les valeurs implicites pour les paramètres d'entraînement.

L'alignement propositionnel seul n'est pas suffisant pour identifier les candidats collocationnels intéressants. Raison pour laquelle nous avons réalisé l'alignement lexical pour plusieurs paires de langues. Nous avons utilisé des corpus segmentés, étiquetés et lemmatisés. L'alignement lexical est indépendant du corpus et des unités lexicales qui le composent. Pour l'alignement lexical du corpus anglais-roumain, nous avons utilisé un outil développé par RACAI (voir Tufiş *et al.* : 2005) qui utilise un algorithme itératif. Pour chaque itération, l'alignement lexical va aligner diverses catégories de mots comme les entités nommées, les nombres, les dates, les mots pleins, les mots grammaticaux et les signes de ponctuation.

L'alignement identifie également des syntagmes non recursifs (groupes nominaux, prépositionnels) qui sont mis en correspondance. Un alignement entre deux unités lexicales (mots ou expressions) est caractérisé par une série de paramètres comme une liste de formes-« cognats », des expressions équivalentes, des catégories lexicales similaires, la localité (la distance entre les mots, à l'intérieur d'un « chunk »). Cette dernière hypothèse a été appliquée après une étude de corpus identifiant les situations où la traduction par une autre catégorie

lexicale est possible (un V traduit par un N ou par un Adjectif, un N traduit par un V, etc.). Pour les autres paires de langues, nous avons utilisé des corpus alignés manuellement pour entraîner Uplug (voir Tiedemann : 2003).

Ainsi, l'alignement lexical peut servir à extraire les équivalents de traduction pour les collocations de type VN. Les mots alignés sont utiles pour établir les cas où les candidats sont traduits par les mêmes mots ou pour valider les diverses classes de constructions dans les trois langues. Nous comparons les cas où nous avons des expressions qui sont alignées avec un seul mot et les cas où les candidats collocationnels étaient extraits par les outils monolingues.

5.2. Un exemple d'extraction

Les candidats confirmés par l'outil d'extraction bilingue, sont, pour la plupart, des collocations pertinentes, où chaque élément de la collocation est traduit de la même façon qu'en dehors de cette combinaison. Les candidats collocationnels non confirmés sont également intéressants pour les traducteurs : la majorité de ces cas sont non compositionnels ; c'est-à-dire que leur traduction implique souvent des traductions non standard (mots uniques ou expressions) (voir Villada Moiron et Tiedeman : 2006).

Nous présentons quelques données extraites à partir des corpus *ACQ* roumain, français, allemand, mais aussi anglais. Les équivalents de traduction communs peuvent être intégrés dans le dictionnaire multilingue de collocations, puisque nous retrouvons des spécificités dans chaque langue : le verbe utilisé qui n'est pas toujours la simple traduction de l'autre verbe, certaines prépositions *în 'en', en, in, into, on* sont intégrées au groupe verbal (*cf.* tableau 7).

Dans les corpus alignés, nous avons identifié quelques candidats qui, étant traduits par un V ou un N simple, n'ont pas d'équivalent VN. L'alignement lexical permet alors d'identifier les traductions correctes, y compris pour les expressions idiomatiques :

Turned a blind eye/lit. 'tourner un oeil aveugle'= a ignorat/ignorer

Meets the eye/lit. 'rencontrer l'oeil'= a se vedea / voir

Give a hard time to/lit. 'donner un temps difficile'=a priponi / attacher

ou pour les prédicateurs complexes :

- *a pune în aplicare/mettre en application/Anwendung vorzusehen = to implement*
- *a aduce atingere/porter atteinte = berühren/to affect*

roumain	**français**	**Allemand**	**équivalent anglais**	**classe**
a ţine cont	tenir compte	Rechnung tragen	to take into account	A

a intra în vigoare	entrer en vigueur	in Kraft treten	to enter into force	A
a lua decizii	prendre des décisions	Entscheidungen treffen	to make decisions	B
a compensa pagubele	réparer les dommages	Schaden ersetzen	to make good (any) damage	B
a exercita activităţile	exercer des activités	Tätigkeit ausüben	to carry on (the) activities	B

Tableau 7. Une liste de candidats communs aux 3 langues, de type prédicateur complexe (A) et prédicat complexe (B)

Si un candidat est proposé par l'outil d'extraction monolingue pour la langue source et qu'il a un équivalent collocationnel dans une des langues cibles, et inversement, alors il peut être inclus dans le dictionnaire.

6. Le dictionnaire

Le dictionnaire en construction sur la base des résultats de l'extraction monolingue et bilingue est conçu comme un dictionnaire multilingue des collocations. Bien qu'il ne contienne pour l'instant que des collocations, il pourra être relié à un dictionnaire de lexèmes simples.

6.1. Un dictionnaire multilingue

Nous partons de l'hypothèse que beaucoup de collocations se traduisent par des collocations ; par conséquent, l'entrée lexicale du dictionnaire est conçue comme une entrée multilingue dès le départ : elle consiste en plusieurs sous-entrées (<te>) dont chacune permet la description d'une collocation, avec un attribut qui en indique la langue :

```
<entry id= "1">
    <te  lang= "fr">...</te>
    <te  lang= "ro">...</te>
    <te  lang= "de">...</te>
</entry>
```

Cette disposition est compatible avec le fait que certaines collocations se traduisent par des mots simples (*turn a blind eye to/ignorer*), parce que les sous-entrées peuvent contenir aussi bien des collocations que des mots simples. Nous partons de l'hypothèse que les collocations et/ou les mots simples réunis dans une entrée expriment le même contenu. Cette équivalence est implicite dans la structure du dictionnaire, mais elle peut être rendue explicite, si besoin est, par un reformatage.

6.2 Description des collocations et de leurs éléments

Dans notre approche, les collocations verbonominales (comme toutes les collocations) sont constituées de deux éléments : le noyau et le collocateur. La description lexicale dans le dictionnaire (<complexitem>) prévoit alors quatre zones dans l'entrée : deux zones décrivent les propriétés de la collocation tout entière (<c_spec>) et les propriétés imposées par la présence d'une préposition (<prep>) ; les deux autres zones sont consacrées à la description de chacun des deux éléments qui composent la construction (dans le cas des constructions VN, ce seront <v_spec> et <n_spec>). En outre, l'entrée contient un élément '<construction>' qui accueille le lemme de la collocation.

```
<entry id= "1">
    <te  lang= "fr">
              <complexitem>
              <construction> ...  </construction>
                    <v_spec> ...  </v_spec>
                    <prep>...</prep>
                    <n_spec> ...  </n_spec>
                    <c_spec> ...  </c_spec>
              </complexitem>
       </te>
</entry>
```

Pour les cas où nous représentons des mots simples dans le dictionnaire, nous remplaçons le <complexitem> par <simpleitem>, caractérisé par ses propriétés spécifiques, selon sa catégorie lexicale (<spec>).

6.3 Propriétés et données lexicographiques des collocations

Parmi les données lexicographiques utilisées pour décrire la collocation dans son entièreté, on peut en distinguer deux: d'une part, les données lexicographiques à proprement parler (surtout les phrases exemples données, ainsi que le type de collocation) ; d'autre part, les propriétés lexicales et syntaxiques de la collocation. Les données du premier type sont essentiellement de nature documentaire et interprétative (éléments <colloc-type> et <examples>), tandis que les dernières sont plus strictement linguistiques et concernent l'usage de la collocation en tant qu'objet linguistique à part entière.

Les propriétés linguistiques dont le dictionnaire tient compte, et qui se trouvent à l'intérieur d'un élément <colloc_spec> sont de trois types :

- la valence de la collocation (élément <required_args>), exprimée en termes du cas et/ou de la préposition requise par la collocation : pour *tenir compte*, on mettra donc :

<required_args prep = "de"> p-object </required_args>

et pour *Rat+erteilen* (‘donner+conseil’), on aura :

<required_args case = "dat"> indirect_object </required_args>

pour exprimer que l’expression prend un complément au datif ;

- la tête lexicale d'un usage restreint de la collocation (élément <lexical_restriction>) : *auf Halbmast setzen* « mettre en berne » n’accepte guère comme complément d’autres lexèmes que *Flagge*, *Fahne* « drapeau ». Dans ces cas, on ajoutera l’élément <lexical_restriction> pour indiquer la restriction ; sa valeur peut être une énumération de lemmes. Pour la documentation lexicographique des collocations, un conteneur <colloc_documentation> permet de grouper les données en question ; nous prévoyons la possibilité de donner plusieurs phrases exemples et de renseigner le LL calculé sur un corpus donné. La description de la collocation entière est donc structurée comme suit :

```
<entry id= "1">
      <te  lang= "de">
      <complexitem>
            [...]
            <c_spec>
            <colloc_spec>
                  <required_args case = "acc"> object </required_args>
                  <lexical_restriction     compl="object">     Flagge,     Fahne
                  </lexical_restriction>
                  <colloc_type> compl_pred </colloc_type>
            </colloc_spec>
            <colloc_documentation>
                  <colloc_LL value="2999.854" corpus="ACQ"/>
                  <examples>
                        <example> ... </example>
                  </examples>
            </colloc_documentation>
            </c_spec>
            </complexitem>
</te>
</entry>
```

6.4. Propriétés et données lexicographiques du nom

Outre l’indication du lemme nominal <n_lemma>, la description des propriétés du nom explicite les critères N1 et N3 mentionnés plus haut, c.-à-d. des préférences en matière de

détermination <det>, nombre <number> et de modification <modifier>. En outre, on mentionne la fonction grammaticale que prend l'élément nominal à l'intérieur de la collocation (<case>).

Si ce dernier élément est unique, les trois premiers ne le sont pas, même si très souvent, il y a de fortes préférences pour une forme morpho-syntaxique spécifique d'une collocation (par exemple, *tenir compte* toujours sans article, au singulier), il peut y avoir des alternatives. L'expression *in + Dienst + stehen* « travailler pour (qqn) » peut en effet se trouver aussi bien au singulier (*er steht im Dienst der Armee*), avec un article défini, qu'au pluriel, avec ou sans article (*er steht in (den) Diensten der Armee)*. Pour exprimer l'interrelation entre les deux préférences, il convient d'utiliser l'élément <alt> qui permet de répéter l'élément <n_spec> tout entier. Là où la variation ne concerne qu'une seule préférence, on peut aussi répéter seulement l'élément correspondant. Pour exprimer les préférences, nous donnons une indication quantitative partout où c'est possible (attribut 'freq'). La partie correspondante de l'entrée pour *in + Dienst + stehen* est reproduite ci-dessous :

```
<entry id= "1">
      <te  lang= "de">
            <complexitem>
                  <construction> in+Dienst+stehen </construction>
                  <v_spec> ...  </v_spec>
                  <alt>
                  <n_spec>
                        <n_lemma> Dienst </n_lemma>
                        <det freq = "64%"> nul </det>
                        <number freq = "64%"> pl </number>
                        <modifier> </modifier>
                        <case> p-obj (in) </case>
                  </n_spec>
                  <n_spec>
                        <n_lemma> Dienst </n_lemma>
                        <det freq = "36%"> def </det>
                        <number freq = "36%"> sg </number>
                        <modifier> </modifier>
                        <case> p-obj (in) </case>
                  </n_spec>
                  </alt>
                  <c_spec>
                        <colloc_spec>
                        <required_args case = "gen"> gen-attrib </required-args>
                        <lexical_restriction compl = "object"> …
                        </lexical_restriction>
                        <colloc_type>  p+c_constr </colloc_type>
                        </colloc_spec>
```

```
                        <colloc_docu>[...]
                        </colloc_docu>
                    </c_spec>
                </complexitem>
        </te>
</entry>
```

6.5. Propriétés et données lexicographiques des verbes

La logique de la description du V est parallèle à celle de la description du N : on indique le lemme du V `<v_lemma>`, ainsi que ses préférences de forme `<v_form>` et de diathèse `<act_pass>`. La préférence formelle concerne essentiellement des combinaisons fortement figées, comme l'expression *compte tenu (de)*. La partie correspondante de l'entrée pour la collocation *faire l'objet (de)* se présenterait alors comme suit :

```
<entry id= "1">
        <te  lang="fr">
        <complexitem>
        <construction> faire+objet </construction>
        <v_spec>
                <v_form> </v_form>
                <act_pass freq = "100%"> active </act_pass>
        </v_spec>
        <n_spec>
        <n_lemma> objet </n_lemma>
                <det freq = "98%"> def </det>
                <number freq = "98%"> sg </number>
                [...]
        </n_spec>
        <c_spec>
        <colloc_spec>
        <required_args case = "de"> p-obj </required-args>
        <lexical_restriction compl = "object"> </lexical_restriction>
        </colloc_spec>
        <colloc_docu> [...]</colloc_docu>
        </c_spec>
        </complexitem>
        </te>
</entry>
```

6.6. Comparaison avec d'autres modèles dictionnairiques

Dans ce qui suit, nous indiquons la structure complète de l'entrée lexicale du dictionnaire[4], (les mentions PCDATA sont remplacées par trois points de suspension (...)):

```
<entry id= "#">
       <te  lang= "fr">...</te><te lang= "ro">...</te>
       <te  lang= "de">
       <complexitem>
              <construction> ... </construction>
              <v_spec>
                     <v_form> ...</v_form>
                     <act_pass freq = #> active|passive </act_pass>
              </v_spec>
              <n_spec>
                     <n_lemma> ... </n_lemma>
                     <det freq = #> def | indef | nul |  poss | dem | quant </det>
                     <number freq = #> sg | pl </number>
                     <modifier> ... </modifier>
                     <case> dat | acc | pobj(...) </case>
              </n_spec>
              <c_spec>
                     <colloc_spec>
                     <required_args case = "gen | dat | acc | pobj"> </required_args>
                     <lexical_restriction compl = "subj | obj | ind_obj"> ...
                     </lexical_restriction>
                     <colloc_type> compl_pred | p+c_constr </colloc_type>
                     </colloc_spec>
                     <colloc_docu>
                     <examples>
                            <example> ... </example>
                     </examples>
                     </colloc_docu>
              </c_spec>
              </complexitem>
       </te>
</entry>
```

Comme nous l'avons indiqué plus haut, les dictionnaires électroniques qui tiennent compte des collocations ne sont pas très nombreux : le modèle DiCo/LAF (voir Polguère : 2000) (implémenté en format électronique http://olst.ling.umontreal.ca/dicouebe/) est certainement la réalisation la plus avancée dans le domaine. Dans DiCo/LAF, on trouve les mêmes types d'information que dans le dictionnaire présenté ici, pour le français. Pourtant, une dimension multilingue est absente du DiCo/LAF. De même en est-il pour le DiCE (*cf.* http://www.dicesp.com), qui propose une analyse sémantique selon les fonctions lexicales de Mel'čuk (1999).

En TAL, seul le modèle suggéré par Braasch et Olsen (2000), pour la base de données STO, contient un certain nombre de propriétés collocationnelles ; les propositions faites dans le cadre de l'initiative de standardisation LMF, *Lexical Markup Framework*, sont plus générales et moins détaillées que les nôtres : le module pour les entrées polylexicales du LMF (voir Francopoulou *et al.* : 2006) ne prévoit pas la description préférentielle des collocations proposée ici ; d'autre part, LMF étant extensible, il ne semble pas impossible de définir une

[4] La DTD est disponible à l'adresse : http://todirasamalia.neufblog.com/dico.dtd

extension particulière de LMF pour les collocations qui permette de transformer notre dictionnaire en une forme compatible avec le LMF. Du point de vue du formalisme, notre dictionnaire se rapproche considérablement du LMF, ou même de *Lexical Systems* (voir Polguère : 2006) : comme ceux-ci, nous nous servons d'une structure de réseau en XML, avec un vocabulaire contrôlé. Ceci nous donne beaucoup de liberté dans la description lexicale, mais l'approche n'est d'autre part pas très contrainte : à la différence de la maquette dictionnairique discutée dans Spohr et Heid (2006), nous renonçons aux possibilités de contrôle de cohérence qu'offrirait un formalisme plus élaboré. L'avantage de la flexibilité se paie donc en support automatique pour le contrôle de consistance.

7. Conclusion

Nous avons présenté ici un outil d'extraction de collocations, ayant comme objectif l'identification de leurs propriétés morpho-syntaxiques pertinentes. Nous avons développé des outils d'extraction monolingues : un système hybride, qui applique d'abord les techniques statistiques, suivies d'une étape de filtrage linguistique, et une approche symbolique, adaptée à l'allemand.

Nous avons également présenté un outil d'extraction de collocations multilingue, qui s'applique à des corpus alignés. La méthode d'alignement (propositionnel et lexical) est présentée en détail. Les candidats extraits pour chaque langue par l'outil d'extraction de collocations monolingue sont confrontés à une liste d'équivalents de traduction établie à partir de corpus alignés et sont utilisés pour alimenter un dictionnaire multilingue.

Remerciements

Ce projet est financé par le réseau « Lexicologie, Terminologie, Traduction » de l'Agence universitaire de la francophonie.

Bibliographie

BLUMENTHAL (Peter) : 2007, «A Usage-based French Dictionary of Collocations», in Kawaguchi (Y.), Takagaki (T.), Tomimori (N.), Tsuruga, (Y.), eds., *Corpus-Based Perspectives in Linguistics*, (Amsterdam u.a.: Benjamins), pp. 67-83.

BRAASCH, (Anna), OLSEN (Sussi) : 2000, "Formalised Representation of Collocations in a Danish Computational Lexicon", in Heid (U.) *et al.*, eds. *The Ninth EURALEX Congress, Proceedings, Vol. II*, (Stuttgart), pp. 475-488.

CEAUSU (Alin), ŞTEFANESCU (Dan) et TUFIS (Dan) : 2006, « Acquis Communautaire Sentence Alignment using Support Vector Machines», in *Proceedings of LREC 2006*, (Genoa).

DUNNING (Ted) : 1993, « Accurate Methods for the Statistics of Surprise and Coincidence», *Computational Linguistics*, 19,(1).

EVERT (Stefan) : 2005, *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, Ph.D. thesis, (Institut für maschinelle Sprachverarbeitung, University of Stuttgart).

FAN (Rong-En), CHEN (Pai-Hsune) et LIN (Chih-Jen) : 2005, « Working set selection using the second order information for training SVM». Technical report, (Department of Computer Science, National Taiwan University).

FRANCOPOULO (Gil), Bel (Nuria), *et al.* : 2006, « Lexical markup framework (LMF) for NLP multilingual resources», in *International Committee on Computational Linguistics and the Association for Computational Linguistics - COLING / ACL 2006*

GLEDHILL (Christopher) : 2007, « La portée : seul dénominateur commun dans les constructions verbo-nominales», in FRATH (P.), PAUCHARD (J.), GLEDHILL (C.), dir., *Actes du 1er colloque Res per nomen*, (Université de Reims), pp. 113-124.

GLEDHILL (Christopher) : 2000, *Collocations in Science Writing*, (Tübingen: G. Narr).

GROSS (Maurice) : 1993, « Les phrases figées en français», in *L'information grammaticale*, 59, (Paris), pp. 36-41.

GROSSMANN (Francis), TUTIN (Agnès), dir. : 2003, « Les collocations: analyse et traitement», Numéro special : « *Travaux et Recherches en Linguistique Appliquée* ».

HAUSMANN (Franz Josef) : 2004, « Was sind eigentlich Kollokationen? », in STEYER (K), eds., *Wortverbindungen – mehr oder weniger fest*, pp. 309-334

HALLIDAY (Michael) : 1985, *An Introduction to Functional Grammar*, (London, Arnold).

HEID (Ulrich) et RITZ (Julia) : 2005, « Extracting collocations and their contexts from corpora», in *Actes de COMPLEX, Conference on Computational Lexicography and Text Research*, (Budapest).

HEID (Ulrich) et WELLER (Marion) : 2008, "Extraction des collocations à partir des corpus : le cas de l'allemand", in Actes de JASR'07, (Tunis).

ION (Radu) : 2007, "TTL: A portable framework for tokenization, tagging and lemmatization of large corpora", (Bucharest : Research Institute for Artificial Intelligence, Romanian Academy).

L'HOMME (Marie-Claude) : 2003, «Les combinaisons lexicales spécialisées (CLS). Description lexicographique et intégration aux banques de terminologie», in GROSSMANN (F.) et TUTIN (A.), dir. *Les collocations : analyse et traitement, Numéro special: « Travaux et Recherches en Linguistique Appliquée »*, pp.89-105

MANNING (Christopher. D.) et SCHÜTZE (Hinrich) : 1999, *Foundations of statistical natural language processing*, (MIT Press).

MEL'CUK (Igor) et al. : 1999, « Dictionnaire explicatif et combinatoire du français contemporain», in *Recherches Lexico-Sémantiques*, (Presses Universitaires de Montréal).

MOORE (Robert C.) : 2002, « Fast and Accurate Sentence Alignment of Bilingual Corpora», in *Machine Translation: From Research to Real Users (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California)*, (Heidelberg : Springer-Verlag), pp. 135-244

POLGUERE (Alain) : 2000, "Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French", in *Proceedings of EURALEX'2000*, (Stuttgart), pp. 517-527.

POLGUERE (Alain) : 2006, « Structural properties of Lexical Systems: Monolingual and Multilingual Perspectives», in *Proceedings of the Workshop on Multilingual Language Resources and Interoperability* (COLING/ACL 2006, Sydney), pp. 50-59.

RITZ (Julia) et HEID (Ulrich) : 2006, « Extraction tools for collocations and their morphosyntactic specificities», in *Proceedings of the Linguistic Resources and Evaluation Conference*, (Genova).

SERETAN (Violeta), NERIMA (Luka) et WEHRLI (Eric) : 2004, « A tool for multi-word collocation extraction and visualization in multilingual corpora», in *Proceedings of EURALEX'2004*, (Lorient, France), vol. 2, pp.755-766

SMADJA (Frank A.) et McKeown (Katheleen. R.) : 1990, « Automatically extracting and representing collocations for language generation», in *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, (Pittsburgh, Pennsylvania), pp. 252-259.

SCHMID (Helmut) : 1994, « Probabilistic Part-of-Speech Tagging Using Decision Trees», in *Proceedings of International Conference on New Methods in Language Processing*.

STEINBERGER (Ralf), *et al.* : 2006, « The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages», *Proceedings of the 5th LREC Conference*, pp.2142-2147.

SPOHR (Dennis) et HEID (Ulrich) : 2006, « Modeling monolingual and bilingual collocation dictionaries in Description Logics», in *Proceedings of the EACL Workshop on Multiwords and Multilinguality*, (Trento, Italia), pp. 65-72.

TIEDEMANN (Jörg) : 2003, « Combining clues for word alignment», in *Proceedings of the 10th EACL*, (Budapest, Hungary), pp. 339-346.

TODIRASCU (Amalia), GLEDHILL (Christopher) et STEFANESCU (Dan) : 2007, « Extracting Collocations in Context: the case of Romanian VN constructions», in *Proceedings of RANLP'2007*, (Bulgaria).

TODIRASCU (Amalia), GLEDHILL (Christopher) et STEFANESCU (Dan) : 2008, « Collocations en Contexte: extraction et analyse contrastive», in *Actes de Journées d'Animation Scientifique Regionale*, (Tunis).

TUTIN (Agnès) : 2004, « Pour une modélisation dynamique des collocations dans les textes», in *Actes du congrès EURALEX'2004*, (Lorient, France), vol. 1, pp. 207-221.

TUFIS (Dan), ION (Radu), CEAUSU (Alin), et STEFANESCU (Dan) : 2005, "Combined Aligners", in *Proceedings of the ACL'2005 Workshop on Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond*, pp. 107-110.

TUFIS (Dan) et KOEVA (Svetlana) : 2007, « Ontology-Supported Text Classification Based on Cross-Lingual Word Sense Disambiguation», in MASULLI (F.), MITRA (S.), PASI (S.), eds. *Proceedings of WILF'2007*, (Berlin Heidelberg : Springer-Verlag) pp. 447-455.

VERLINDE (Serge), SELVA (Thierry) et BINON (Jean) : 2003, « Les collocations dans les dictionnaires d'apprentissage: repérage, présentation et accès», in GROSSMANN (F.), TUTIN, (A.), dir. *Les collocations: analyse et traitement*, (Amsterdam : De Werelt), pp. 105-115.

VILLADA MOIRON (Begona) et TIEDEMANN (Joerg) : 2006, « Identifying idiomatic expressions using automatic word-alignment», in *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*. April, 3rd, (Trento, Italy).

WILLIAMS (Geoffrey) : 2003, « Les collocations et l'école contextualiste britannique», in GROSSMANN (F.) et TUTIN (A.), dir., *Les collocations : analyse et traitement : Travaux et Recherches en Linguistique Appliquée* . (Amsterdam : DeWerelt).

ZINGLE (Henri) et BROBECK-ZINGLE (Marie-Louise) : 2003, *Dictionnaire Combinatoire du Français, Expressions, locutions et constructions*, (France : La Maison du Dictionnaire).