A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions

Amalia Todiraşcu[1], Dan Tufiş[2], Ulrich Heid[3], Christopher Gledhill[1], Dan Ştefanescu[2], Marion Weller[3], François Rousselot[4]

[1] LILPA, Université Marc Bloch, Strasbourg, France

[2]RACAI, Romanian Academy, Bucharest, Romania

[3]IMS Stuttgart, Universität Stuttgart, Germany

[4]INSA Strasbourg, France

E-mail: {todiras, gledhill}@umb.u-strasbg.fr, {tufis, danstef}@racai.ro, {uli, wellerm}@ims.uni-stuttgart.de, Francois.Rousselot@insa-strasbourg.fr

Abstract

We present the main findings and preliminary results of an ongoing project aimed at developing a system for collocation extraction based on contextual morpho-syntactic properties. We explored two hybrid extraction methods: the first method applies language-indepedent statistical techniques followed by a linguistic filtering, while the second approach, available only for German, is based on a set of lexico-syntactic patterns to extract collocation candidates. To define extraction and filtering patterns, we studied a specific collocation category, the Verb-Noun constructions, using a model inspired by the systemic functional grammar, proposing three level analysis: lexical, functional and semantic criteria. From tagged and lemmatized corpus, we identify some contextual morpho-syntactic properties helping to filter the output of the statistical methods and to extract some potential interesting VN constructions (complex predicates vs complex predicator). The extracted candidates are validated and classified manually.

1. Introduction

We present the main findings and preliminary results of a project aimed at developing a system for collocation extraction based on contextual morpho-syntactic properties. In this paper, we present two variants of a hybrid approach to the extraction of collocations from text corpora and to their classification. Collocations are lexical expressions composed of two or more items, each with their own at times unpredictable syntactic and semantic behavior. The term 'collocation' has been used in many different ways in linguistics (Bartsch 2004, Manning/Schütze 1999). Lexicographers (cf. Hausmann 2004) consider collocations as lexical expressions composed of at least two elements (a noun and a verb, in our case), which come in a specific grammatical relation. We adopt rather a contextualist approach (Williams, 2003), we consider that collocations are co-occurences (elements co-occuring frequently) as well as constructions (syntactic relations established between its elements), to be used in appropriate contexts.

Collocation is a crucial feature of idiomatic language use, and represents an important source of ambiguities and errors for NLP applications (such as parsing). They are a problem for lexical selection in NL generation (Wanner 1996), translation (Tufiş et al., 2006), and language learning. There has been much research work on collocation extraction over the last few years. Approaches to collocation extraction range from purely statistical ones, combined ones (statistics and linguistic pattern based extraction) to those relying on parsed corpora and detailed extraction patterns. Approaches combining linguistic and statistical knowledge can be distinguished according to the order of application of both types of knowledge: Smadja (1993) first identifies all statistically significant word pairs and then filters them in terms of part-of-speech combinations, e.g. to identify verb+noun collocations (e.g. *pay* + *attention*). Krenn (2000) and Heid (1998) invert the order and first extract, for example, all predicate + complement pairs, before ordering them by co-occurrence significance.

Our work follows these two lines of hybrid approaches, for three languages, namely two Romance (French, Romanian), and one Germanic (German). We extract not only data about lexical combinations, as it is done by, e.g. Smadja (1993), who delivers word pairs only, but we also provide data on the morpho-syntactic properties of each collocation, i.e. on its fixedness (see section 3.3 for details). Although our approach has some similarities with Fazly &Stevenson (2006), we do not only use the data about morpho-syntactic fixedness to classify noun+verb-combinations into compositional vs. non-compositional (i.e. idiomatic), but we also capture the morpho-syntactic specificities of each combination in detail, one goal being to build an electronic dictionary for NL analysis and generation. Moreover, we use the morphosyntactic features of the potentially interesting noun+verb constructions to differentiate, at least partially, between *complex predicators* (lexicalized SVCs, non-compositional, idiomatic ones) vs. *predicate + complement structures* (where the lexical combination is unpredictable, but the expression as such is not idiomatic). We extract lexical co-occurrence data and morpho-syntactic features of the extracted word groups in one single tool architecture.

As mentioned, we use two variants of the hybrid extraction method: the first approach is language independent and requires less linguistic knowledge. Both start from pos-tagged and lemmatized corpora.

Adjacency of a lexical items pair is not a requirement for selecting it as a potentially interesting noun+verb construction. For French and Romanian, we identify candidates by means of a statistical extractor which checks two criteria:

– the distance between the verbal and the nominal element in the sequential text: the less variation there is in the distance, the more likely the pair is a collocation;

– the strength of the association between both elements, as measured by the log likelihood ratio test (Dunning, 1993).

The candidates identified in this way are filtered by means of linguistic patterns, so as to remove unwanted candidates. For the task of classification, i.e. to decide to which type of collocation a given candidate pair belongs (see below for the classification used), we rely on the morpho-syntactic fixedness properties of the candidates.

For German, the same method is used, but we apply, in addition, a second hybrid approach, which inverts the order of linguistic and statistical knowledge, and which requires more linguistic knowledge to be encoded in extraction tools. The classification task relies on morpho-syntactic fixedness, the same way as for the Romance languages.

The precision evaluation for the three languages, was estimated by manual checking of the top-500 candidates ordered by log likelihood. As we use two different strategies, we can compare their output and assess the relationship between linguistic "investments" and the "dividends" earned (cf. Section 5). Given that the texts we work with are parallel (French, German, Romanian plus English as a hub), a partial cross-linguistic comparison is also possible.

The remainder of this paper is organized as follows: we first summarize the targeted linguistic classification of verb+noun collocations and the main morpho-syntactic properties we consider for the classification (section 2 and section 3). Section 4 is devoted to brief descriptions of both methods, and section 5 contains preliminary results and their partial evaluation. A full evaluation will be achieved by the end of the project and will be available at the conference presentation.

2. Verb+Noun Constructions

The work reported here is concerned with verb+noun (VN) constructions where the noun is involved in a complement or a prepositional complement of a predicator. As many VN constructions show morpho-syntactic idiosyncrasies, it is necessary to study the morpho-syntactic properties of each collocation and its constituents. This may be done once the lexical associations are identified (Tutin 2004), or together with the detection of collocation candidates (Ritz, Heid 2006). Among the verb+noun collocations we concentrate on the specific set of collocations known as **VN constructions**. We adopted Gledhill's (2007) functional criterion in order to arrive at a single category. His analysis is based on systemic functional grammar (Halliday 1985). This approach supposes that various lexico-grammatical systems contribute simultaneously to the construction of a message. Three systems are relevant to VN constructions, namely: i) Syntactic Function, ii) Lexical Structure and iii) Semantic Roles. These can be seen in example (1):

1)	*I*	*/ 'm*	***making***	*/ money*[1]
Function	S	F	P	C
Structure	Pro	Aux	V + ING	N
Role	AGT		MAT	MED

In this example, a prototypical transitive verb *make* expresses various levels of meaning, namely: i) assertion at the syntactic level (finite + predicator), ii) aspect at the lexical level (progressive), and iii) a dynamic, material process (MAT). The process expressed by the predicator (MAT) determines the roles played by the other elements in the clause, namely a 'process-external' participant or agent (AGT), and a 'process-internal' participant or medium (MED). We can compare this example with (2), whose analysis only differs at the level of semantic roles:

2)	*You*	*/* ***made***	*/ a*	***suggestion?***
Function	S	F-P	C	
Structure	Pro	V	Det	N
Role	MED	MEN	(PROC)	

In this case, the main difference is that we now have a communicative, or mental process (MEN), which is specified by the complement (PROC) with a process-internal participant (MED) now expressed by the subject. But the complements in (1) and (2) have the same syntactic status. Few probes can be used to distinguish them formally (they allow the passive, but they resist interrogation) The only difference between the two is semantic. In (1) the complement expresses a canonical modified object, whereas in (2) the complement expresses what Halliday calls '**process range**'. Process range (as defined by Gledhill 2007) is a form of grammatical metaphor in which a semantic process is designated or delimited by an element in the predicate which is not the predicator (the main lexical verb).

Our final point regarding terminology concerns a sub-category of VN construction which is particularly prevalent in our corpus, especially in French, Romanian and German, but perhaps less so in English. We make a distinction between **complex predicates**, that is to say constructions in which the process range is expressed by a complement, as in *make a suggestion, do the washing up,* and **complex predicators**, that is to say constructions in which the process range is not expressed by a complement, but by an element such as an extension of the predicator or an adjunct, as in *make fun of someone,* or *take oneself seriously.* The fact that these two terms are close in form is deliberate: we argue that they are simply two sub-types of the same family of VN construction.

From a lexical point of vue, VN constructions act as lexical items, being caracterized by a set of morpho-syntactic properties (Gledhill 2007):

- noun properties: determiners (presence or absence of determiner), qualifiers (nouns could be modified), conversion
- verb properties: arguments, verb equivalents or passivation;

[1] in SFL '/' identifies a group structure (nominal, verbal)

Even if none of these properties is sufficient to identify VN constructions, we might use them for an automatic extraction, followed by a manual validation (using semantic criteria). Then, **complex predicators**, i.e. combination of V+N which as a whole act as a predicator, such as FR *faire l'objet de* or RO *a face obiectul* ("[to] concern"), *a ține cont* ("take into account"), or DE *Gebrauch machen* ("make use (of)") are typically characterized by a high degree of morpho-syntactic fixedness (e.g. the noun may occur only in singular/plural, be only in definite/indefinite form and never takes a modifier; the verb may be used only in specific tense, mood or diathesis). **Complex predicates** (i.e.) predicate+complement structures, are lexical collocations which show more morpho-syntactic variability and are therefore more compositional; examples include FR *prendre des mesures*, RO *a lua măsuri*, DE *Maßnahmen ergreifen* (all: "take measures"). In these examples, the complement may occur with different grammatical features and may take a modifier.
Morpho-syntactic fixedness concerns several parameters, applicable to the collocation as a whole, and partly to its components:

- collocation property: type of construction: *V+NP, V+PP*
- noun properties:
  - determination: *definite, indefinite, possessive, null,* etc.
  - number: *singular, plural*
  - modifiability: by *adjectives, prepositional phrases, relative clauses,* etc.
- verb properties:
  - preference for certain forms: *tense, mood*
  - preference with respect to voice: *active, passive*

To define linguistic patterns used by the extraction systems, we study various monolingual and multilingual, aligned corpora to detect relevant morpho-syntactic properties, across languages.

3. The Data

3.1. Corpus Description

For our experiments we used both multilingual parallel and monolingual texts.
The *Acquis Communautaire* corpus (Steinberger et al., 2006) contains parallel documents in 22 languages about laws, conventions, treaties etc. adopted by EU member states since 1950. It is one of the largest multilingual parallel corpora freely available. From the *Acquis Communautaire (JRC-Acquis)* multilingual parallel corpus we extracted a sub-corpus containing only the English sentences which are 1-1 aligned with corresponding sentences in all our target languages: FR, RO and GE. Based on the alignment transitivity we generate the language pairs we were interested in (FR-GE, RO-FR, RO-GE) for our investigation.

The sub-corpus, extracted as mentioned, contains 60389 sentences and around 1,4 million words per language (see table 1 below).

We used the parallel corpus to cross-lingually check our hypotheses, but because the language in the Acquis corpus is rather formulaic, we also made separate monolingual analysis, using monolingual newspaper corpora containing extracts from:

- *Le Monde*, *Le Monde Diplomatique* (FR);
- *Agenda, Evenimentul Zilei* (RO);
- *Frankfurter Rundschau* and *Stuttgarter Zeitung* (GE).

Language	number of tokens (60389 sentences)	average number of tokens per sentence
English	1466912	24.29
French	1527241	25.29
German	1314441	21.76
Romanian	1422995	23.56

Table 1. Compositions of the extracted sub- corpus.

However in this paper we refer only to the analysis of the data extracted from the *JRC-Acquis* sub-corpus.

The Romanian data was tagged and lemmatized by TTL (Ion 2007), while the French and German texts were tagged using TreeTagger (Schmid, 1994). Due to the fact that TreeTagger is trained on newspapers, French and German (Acquis sub-corpora) contain many tagging and lemmatization errors. While TreeTagger provides a reduced set of French tags, we applied Flemm (Namer, 2000) to obtain correct lemma and to complete the tags with morpho-syntactic information (number, gender, case etc.), required to define filtering patterns. A post-tagging manual validation eliminated most of the tagging/lemmatization errors.

The German part of the parallel data, besides being tagged and lemmatized (STTS tagset (Schiller et al, 1995)), is additionally syntactically annotated. These annotations are used, based on manually defined patterns, by the symbolic extraction method for German.

While the combination of morpho-syntactic attributes characterising the constituents of a VN construction could be specific to each language, we found that lexical idiosyncrasies, although not identical, could be pinpointed also cross-lingually. That is to say that if one finds a VN construction in one language, its translation equivalent in the other language has all the chances to be a VN construction as well. In our experiments we monolingually computed the lists of potential candidates for VN constructions and than by using bilingual word alignment we checked whether, for instance, a VN construction candidate found in FR could be also found, via word alignment equivalence, in RO. Since morpho-lexical criteria for checking a VN construction were language specific and developed independently, whenever we were able to establish a cross-lingual lexical translation equivalence, this has been interpreted as an evidence of considering it as a valid construction. We present here the word alignement methodology applied to *JRC-Acquis* corpus.

3.2. Aligned Corpus

The word alignment system (Tufiş et al., 2006) uses a statistical alignment model and a statistical translation dictionary. For the statistical translation dictionary we use GIZA++ (http://www.fjoch.com/GIZA++.html) and

lemmatized parallel corpora (for languages with productive inflectional morphology, in order to increase statistical confidence, the translation equivalence probabilities are computed for lemmas not for wordforms). The alignment model consists of various weights and thresholds for different features and they are supposed to work for most Indo-European languages (cognates, translation equivalence entropy, POS-affinities, locality etc.). Based on our previous translation Ro-En model and the Ro-En translation dictionary extracted from the JRC-Acquis, we aligned several Ro-En parallel documents. From the Fr-En sub-corpus we extracted a Fr-En translation dictionary but since we do not have yet a Fr-En word-alignment model we used the model built for English-Romanian alignment and given the languages closeness the accuracy of the resulted alignment of Fr-En texts was acceptable (but obviously lower than in the case of Ro-En). For the evaluation purposes there were selected 1000 sentence-pairs (Ro-En, Fr-En) and their word alignment was manually validated and corrected. The Ge-En 1000 sentences were aligned (Fraser and Marcu, 2007) and were manually validated, using an alignment editor, very similar to the one used for Ro-En and Fr-En (see Fig. 2).

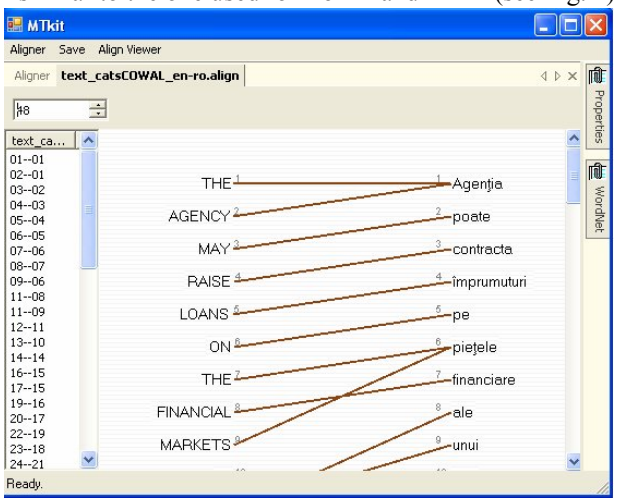


Figure 2. The lexical alignment editor

The word alignment links are representations of translational equivalence between the respective tokens and we rely on the heuristics called TH, according to which if M words in language L1 are aligned to N words in the hub language L2, and these N words are aligned to Q words in language L3, then it is highly probable that the N words in language L1 are aligned to the Q words in language L3. We decided to take a hub approach with English as the language to which all the lexical alignments were done for multiple reasons: it will be simpler to extend our approach to several other languages represented in the Acquis corpus; for evaluations and corrections is easier to find experts understanding English and the other language; linguistic resources and the processing tools available for English, as well as the ever improving alignment technologies allow for cross-lingual annotation transfer and thus rapid prototyping of linguistic knowledge for the target language, etc. Once the VN construction candidates have been monolingually identified in one language, the TREQ-AL system, largely described in (Tufiş, 2004), uses the lexical alignments and the TH heuristics to find their translations in the other two languages.

In order to extract collocation candidates for each language, we identified several relevant language-dependent morpho-syntactic properties from the tagged, lemmatized monolingual corpora. Aligned corpora were used to cross-lingually compare the morpho-syntactic preferences.

3.3. Interpreting Data

As mentioned in the introduction, the output of the statistical extraction method should be filtered in order to eliminate invalid candidates. To define these filtres, we studied the contextual morpho-syntactic properties for each language. We select a set of most frequent VN constructions for their language, involving very frequent verbs (*faire/a face/machen* 'to make', *mettre/a pune*/'to put' etc.). We manually identified their contextual morpho-syntactic properties, and their VN construction category. The contextual relevant morpho-syntactic properties were looked upon the verb+noun comstruction, as well as the indirect complement or circumstantial complement. Most of these properties expressed in terms of atribute-values existing in the tagsets used by each language. We then identified preferences for some properties and values, and we use them to define selection patterns to extract relevant candidates from the output of the statistical module.

Verb	Noun	Art	Case	Nr	Type
faire 'make'	objet 'subject'	definite	acc(de)	sg	A
tenir 'take'	compte 'account'	null	acc(de)	sg	A
remplacer 'replace'	texte 'text'	def, indef	-	sg, pl	B
prendre 'take'	considération 'account'	null	-	sg	A
prendre 'make'	décision 'decision'	def, indef	-	sg, pl	B

Table 3. The most frequent FR VN constructions extracted from the *JRC-Acquis: A – complex predicator; B – complex predicate*

Even if none of these properties itself is sufficient to decide that the VN cooccurrence is a VN construction, they are useful to automatically select appropriate candidates, using few linguistic ressources (the tagged and lemmatized corpus). Some of these properties are common to all languages (determiner, number, mood or tense). While the absence of the determiner or the preference for the definite article represent strong criteria to identify complex predicator for all the studied languages, preference for possessive article are specific to German data. Cases of the direct or indirect complement are useful properties for Romanian and German, while gender is only relevant for German VN constructions.

In table 3 and 4, we present some of the most frequent VN co-occurrences for French and Romanian and the most salient properties (determiner, number, and the case of the indirect complement). For French, the cases are identified only by the preference for some specific

prepositions. Complex predicates are then identified by more variable preferences.

Verb	Noun	Art	Case	Nr	Type
aduce 'affect'	atingere	null	genitive	sg	A
înlocui replace'	textul 'text'	null, definite	acc	sg,pl	B
face 'make'	obiectul 'subject'	definite	genitive	sg	A
lua 'take'	măsuri 'measures'	null,definite, indefinite	acc	sg	B
ține 'take'	cont 'account'	null	acc(de)	sg	A

Table 4. The most frequent RO VN co-occurences extracted from the *JRC-Acquis.*

In table 5, we present German data, characterized by specific properties as voice (active or passive), or sentence type (v-1, v-2 or Vfinal). These properties should be identified on syntactically annotated corpus and they are used by the symbolic extraction system.

Noun	Verb	Art	Nr	Voice	Type	Class
Rechnung 'account'	ausstellen 'establish'	def, indef	sg	passive, actif	vfinal	B
Bezug 'reference'	nehmen 'make'	null	sg	active, passive	vfinal, v-1	A
Rechnung 'account'	tragen 'take'	null	sg	active, passive	vfinal, v-1	A
Gebrauch 'use'	machen 'make'	null	sg	passive	v-1	A

Table 5. Some of the most frequent VN co-occurences extracted from the German *JRC-Acquis*

Although we identified relations between morphosyntactic fixedness and the process type (Todirascu et al, 2007), it is not possible to have an automatic extraction of VN classes and thus a manual validation is then necessary.
As we already discussed, we used these properties to define linguistic filters to select candidates extracted by statistical methods. In addition, we use syntactic information for the symbolic extraction method. We then present the extraction methods evaluated in our project, and we focus on linguistic filtering.

4. Collocation Extraction Methods

4.1. The statistical method

For all three languages, we use a statistical collocation extractor (Ştefănescu et al, 2006) which is not bound to word adjacency, being able to detect noun+verb coocurrences which are not contiguous. The criteria for considering a noun+verb as a possible interesting construction are:

- the stability of the distance between noun and verb within texts (judged by a low standard deviation of these distances): this parameter is particularly useful for complex predicates in configurational languages with a relatively fixed constituent order, and for German verb final sentences.
- the co-occurrence significance of noun and verb (in terms of loglikelihood - LL).

This module proposes a list of the most frequent VN cooccurrences (order by LL), their contexts and their frequency (fig. 6):

V=avea N=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/v3/avea în/s/în vedere/nsrn/vedere	31

Fig 6. Various contexts extracted for *a avea în vedere* ('having regard to'): **vg** – gerund verb; **s** – preposition; **nsrn** – noun, singular, accusative, no article; **vp** – past participle; **v3** – verb 3rd person; **vn**- infinitive

The output of the statistical method should be filtered to eliminate irrelevant candidates, but as well to select valid candidates. As we presented in section 3.3, we use morpho-syntactic preferences to define linguistic filters. We applied two categories of patterns on the extracted contexts:

(1) patterns used to identify invalid candidates. We apply some heuristic rules: a longer distance (more than 5 words), the occurrence of a sentence boundary or of several prepositions bewteen the verb and the noun are signs of invalid candidates;
(2) patterns used to select potential relevant candidates. Morpho-syntactic fixedness is a relevant criteria to select complex predicators. For the complex predicates, characterized by variable morpho-syntactic properties, extraction patterns select as well irrelevant candidates (for example, the cases where the noun is the circumstantial complement and the noun is the indirect complement of the predicate could not be distinguished automaticaly).

The filtering module uses the contexts extracted by the statistical module. First, we match eliminatory patterns to the contexts of each candidate and we delete the matched contexts. If all the contexts were deleted, then the candidate should not be selected. Secondly, we apply selection patterns to get potential VN constructions.
We defined a simple language to describe the patterns. A pattern is composed of tags or lemmas, of operators, inspired by regular expressions syntax:

<tag>|<lemma> (<tag>|<lemma>)$^{<op>}$
where <op> could be:

- {n,m} – means minimum n and maximum m tags or lemmas;
- + means at least 1 tag or lemma;
- * means 0 or several tags or lemmas

Examples of patterns eliminating candidates:

a) VER PRP <tag>$^{+}$ PRP NOM

where VER is the verb; PRP is a preposition; at least 1 tag occurs between the prepositions; NOM is the noun;
This pattern eliminates any candidate where the verb and the noun is separated by at least 2 prepositions as : *le texte modifié$_V$ en dernier lieu par la Comission$_N$*'the text changed at the last moment by the Commission'.

After deleting contexts matching eliminatory patterns, the selection patterns should match the remaining contexts associated to the candidates:

b) VER NOM:Ns de|à

where: VER is the verb; NOM:Ns is a common noun, singular; followed by a lemma (one of the prepositions *de, à*)

This pattern is used to select French complex predicators as *tenir compte 'take account', faire usage 'make use', faire face 'to face'*

c) V în NxN

where V is the verb; followed by a lemma (the preposition în/'in') NxN – noun, without determiner

This pattern selects Romanian complex predicator candidates: *intra în vigoare* 'enter into force', *pune în aplicare* 'bring into force', *lua în considerare* 'take account'.

Some of the invalid candidates could not be identified using only lemma and tag information. Syntactic annotation is then useful to improve extraction.

4.2. Symbolic extraction

For German we use, in addition to the statistical procedures described above, a symbolic, pattern-based approach for further filtering the noun+verb combinations:

- Candidates are first extracted by a set of relatively fine-grained symbolic extraction patterns, which are aimed at indentifying predicate + complement, verb + indirect complement and verb + prepositional complement pairs from those contexts where German word order allows to decide with good precision that the two elements may collocationally belong together. This step relies on recursive chunking (Kermes, 2003). The same patterns also capture morpho-syntactic details of the candidates and store them alongside the coocurrence data.
- In a second step, the lemma pairs are ordered according to log likelihood.

Regular expression queries informed by the peculiarities of German word order and verb placement rules account for those syntactic contexts (e.g. sub-clauses, passives), from which predicate+complement constructions can be extracted with good precision. The queries also extract the abovementioned features from the partially parsed text. The features of each sentence extracted are identified and stored in a database; preferences are then computed by summing up over the feature frequency of all available sentences for a given VN construction and comparing the values (Ritz, 2006), in order to arrive at preferences in terms of percentages (Evert 2005).

For the classification of the extracted candidates into **complex predicators** (**type A**) vs. **complex predicates (type B)** we rely on morpho-syntactic fixedness: the more restricted the candidate with respect to the morpho-syntactic features listed in section 2 (i.e. the less variation we recognize), the more likely the candidate is a complex predicator.

5. Results – Evaluation – Interpretation

5.1 Monolingual extraction

The interesting VN coocurrences, as found by the statistical extraction method, are ranked according to the loglikelihood score and the inflectional variations are grouped together as shown below. We cut off the selected candidates if the LL is less than 9. Note the variation with respect to number and definiteness in Table 7, giving support to consider the collocation as a predicate+complement (type B) construction. The Romanian tagset used here is the CTAG tagset of the tiered tagging methodology (Tufiş, 1999). This tagset is automatically expandable to the MSD tagset fully compliant with the MULTEXT-EAST specifications (http://nl.ijs.si/ME/): **vn**-verbe infinitive; **vp**-verbe past participle; **npn** – noun plural, no article; **npry** – noun plural, accusative, definite article.

Lemmas combination: lua+măsură	**LL: 19209,013**	**Av. Distance=1**
Occ/tag (lemma1)	**Occ/tag (lemma2)**	**Frequency**
lua/vn	Măsuri/npn	244
lua/vn	Măsurile/npry	148
luat/vp	Măsura/nsry	56

Table 7: interesting VN constructions in Romanian.

In Table 8, the invariability of the noun gives high confidence in considering the collocation as a complex predicator (type A). The French tagset is used by TreeTagger (Stein, Schmid, 1995) : ver:infi – verb infinitve; ver:pper – verb past participle; ver:pres – verb present tense; nom – noun; det:art – definite article.

Lemma combination: faire+le+objet		**LL: 46334.620**	**Av. distance=2**
Occ/tag (lemma1)	**Occ/tag (lemma2)**	**Occ/tag (lemma3)**	**Frequency**
faire/ver:infi	l'/det:art	objet/nom	1216
fait/ver:pper	l'/det:art	objet/nom	960
font/ver:pres	l'/det:art	objet/nom	932

Table 8: interesting VN construction in French

In Table 9 is shown a German collocation (*make use*) with statistical data implying that this is a complex predicator:

Lemma combination: Gebrauch + machen	**LL:5897,334**	**Av. distance=1**
Occ/tag (lemma1)	**Occ/tag (lemma2)**	**Frequency**
Gebrauch	machen (active)	278
Gebrauch	Gemacht (passive)	64

Table 9: interesting VN constructions in German

To evaluate the precision of the statistical (unfiltered) approach for all three languages, we manually validated the 500 top ranked word pairs suggested by the tool, applying the semantic criteria of Process Range (section 2). We did the same validation for the rule-based filtered German candidates. The validation considered only the correction of the collocation, without further distinction between predicate+complement and complex predicator constructions (to be done in the next evaluation step, using the semantic criteria presented in section 2). Table 10 summarizes the results.

Results n = 500	Statistical extractor			Rule-based extractor
	RO	FR	DE	Frequency
True positives	211	171	157	223
False positives	289	329	343	277
Precision	42,2%	34,2%	31,34%	44,6%

Table 10: multilingual evaluation of the noun+verb constructions

5.2. Error Types

While the statistical module extracts many invalid candidates, we examined invalid candidates and we classify them, in order to propose eliminating patterns. The vast majority of false positives fall into the following types:

- complements of the multiword expressions wrongly identified as MWE parts: *să informeze Comisia cu privire*$_{N}$ *la* ... (“to inform the Comission concerning…”) correct: *informeze+Comisia*), *le plan d'urgence interne prévu*$_{Vpart}$ *à l'article*$_{N}$ ("the emergency plan provided at article…") correct: *plan*$_{N}$ *prévu*$_{V:pper}$;
- subject+predicate combinations: *acest regulament*$_{N}$ *va intra*$_{V}$ *în vigoare* (“this rule will enter into force”, correct: *intra*$_{V}$*+vigoare*$_{N}$), *La Comission*$_{N}$ *propose*$_{V}$ *la modification des dispositions legales... ("*the Comission proposes the following change of the law provisions..."*) Diese Bestimmung*$_{N}$ *gilt*$_{V}$ *nicht* ("these provisions does not apply");
- predicate+adjunct combinations: *articolul a fost modificat*$_{V}$ *ultima dată*$_{N}$ ("The article has been modified last time...", correct *modifica+articol*) *les dates visées*$_{VPart}$ *au présent article*$_{N}$ (“the dates concerned by the present article”, correct: *viser+date*);
- mistagging adjectives as verb participles: *le jour*$_{N}$ *suivant*$_{V}$ (“the following day”) *dispozițiile*$_{N}$ *modificate*$_{V}$ ("the modified provisions"); correct: no extraction;
- GE separable verb prefixes not recognized *(“...teilen*$_{V}$ *der Kommission den Wortlaut*$_{N}$ *(mit)*: produced : “*Wortlaut*$_{N}$ *teilen*$_{V}$”; correct version : “*Wortlaut*$_{N}$ *mitteilen*$_{V}$”).

To eliminate some of these invalid candidates, we defined patterns matching each error class. The aim of using these patterns is to improve the results of the extraction tool:

a) NOM VER:pper

where NOM is a noun; VER: pper is past participle. This patterns eliminates the French noun groups composed of a noun and a verb past participle.

b) V3 <tag>{1,2} NxOy

where V3 is verb 3rd person; <tag>{1,2} means at most 2 tags might occur between the verb and the noun; NxOy is a genitive noun.

These patterns eliminate the Romanian candidates when the noun is the indirect complement, marked by the genitive case.

The comparison of the two methods suggests that the knowledge-poor statistical collocation extraction devices can effectively be used for languages for which no detailed grammatical knowledge is available, or for efficient probing into data, without expending effort in designing regular queries. However, syntactic information is required to identify invalid candidates (subject+predicate combinations). Nevertheless, it is possible to identify morpho-syntactic properties of the collocations along with the candidates themselves.

5.3. Multilingual extraction

As mentioned in section 3.2, we used sentence and lexical aligned corpus to find lexical translation equivalents for each of the pairs FR-RO; RO-GE; GE-FR. These results are compared to candidates extracted from the monolingual corpus. Table 11 presents some candidates having a collocation equivalent in all the languages, and the class is similar:

Romanian	French	German	English	Class
a ține cont	tenir compte	Rechnung tragen	to take account	A
A intra în vigoare	entrer en vigueur	treten in Kraft	enter into force	A
a lua decizii	prendre des décisions	Entscheidungen treffen	to make decisions	B
a da naştere	donner lieu	Anlaß zu geben	to give rise	A
a face referire	faire référence	Bezug nehmen	to refer to	B

Table 11. Common candidates and their classes

In order to validate the methodology presented in section 4, we manually analyzed the collocation equivalents extracted from a set of 1000 word-aligned sentences. For each language pair, we studied the lexical translation equivalents. This experiment shows as well that VN constructions are not always equivalent across languages. For example, *a compensa daunele* 'to compensate the dammages' is translated in French as single verbe *dedommager*, but in the French corpus we find *reparer les dommages*, translated in English as *to make good dammages* or in German *Ersetzen Schaden 'replace the dammages'*. The class of the VN construction is not similar across languages: *emmetre un avis* (FR) *or emite un aviz* (RO) are complex predicates, but in German the equivalent *Stellung nehmen* is complex predicator.

We evaluate manually the VN candidates extracted by the two systems from the word aligned set of sentences for the three languages. We applied then the semantic criteria to distinguish between complex predicators, complex predicates and simple V+complement constructions. We evaluate precision using the number of complex predicate and predicators.

Class	RO	FR	GE
complex predicator	7	8	5
complex predicate	49	53	47
V+Complement	57	49	52
Subject+Predicate	7	5	6
Other classes	48	53	45
Total	168	163	137
Precision	33,33%	37,2%	38%

Most of the candidates have a collocation equivalentin the other languages. The error rate is between 25 to 33% (subject+predicate and the other classes).

6. Conclusion

While our statistical approach relies exclusively on pos-tagging and lemmatization, our symbolic approach uses regular expression based patterns to extract collocation candidates from chunked material. The statistical approach is oriented towards recall, the symbolic one towards precision. For a lexicographic application, both provide raw material for manual inspection.

Future work includes implementation of French and Romanian symbolic patterns, and extension of the filtering patterns to improve classification as complex predicators or complex predicates. We are currently using the tools to create data for a trilingual dictionary of VN collocations; we are using word alignment on the *Acquis Communautaire* corpus to extract equivalence pair candidates for the three languages and compare them with the monolingually extracted data.

7. Acknowledgements

This work has been funded by Agence Universitaire pour la Francophonie (AUF).

8. References

BARTSCH, S. (2004). *Structural and Functional Properties of Collocations in English*. Tübingen: Narr.

EVERT, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, IMS, Universität Stuttgart

FAZLY, A, STEVENSON, S. (2006). Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations. Proceedings of EACL-2006, 337-344. Trento, Italy.

FRASER, A., MARCU, D. (2007) Measuring Word Alignment Quality for Statistical Translation, Computational Linguistics, 33 (3): 293-303.

ION, R. (2007). Metode de dezambiguizare semantică automată. Aplicații pentru limbile engleză şi română. Ph.D.Thesis, Romanian Academy.

GLEDHILL, C. (2007). La portée : seul dénominateur commun dans les constructions verbo-nominales, in Actes du 1er colloque Res per nomen, Université de Reims, 113-124.

HALLIDAY, M.A.K. (1985). An Introduction to Functional Grammar. London, Arnold.

HAUSMANN, F.J. (2004). Was sind eigentlich Kollokationnen?, en K.Steyer (eds.) Wortverbindungen – mehr oder weniger fest, 309-334

HEID U. (1998). Towards a corpus-based dictionary of German noun-verb collocations, in Proceedings of the EURALEX'1998. Liège, 301-312.

KERMES, H. (2003). *Off-line (and On-line) Text Analysis for Computational Lexicography*, Ph.D. thesis IMS, University of Stuttgart, AIMS, vol. 9, n. 3.

KRENN, B.(2000). The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations. PhD thesis, Universität des Saarlandes.

MANNING C. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press.

NAMER, F. (2000) Flemm: Un analyseur Flexionnel de Français à base de règles. Jacquemin, C. (éds) - *Traitement automatique des Langues pour la recherche d'information*. Paris: Hermes, pp.523-47.

RITZ, J., HEID, U. (2006). Extraction tools for collocations and their morpho-syntactic specificities, In: *Proceedings of LREC'2006*, Genova, Italia.

SMADJA, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143-177

SCHILLER, A., et al (1995) Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS, Technical report, Universität Stuttgart.

SCHMID D. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*

STEIN, A, SCHMID, H. (1995). Etiquetage morphologique de textes français avec un arbre de décisions. *Traitement automatique des langues*, Vol. 36, n. 1-2: Traitements probabilistes et corpus, 23-35.

STEINBERGER R. et al. (2006) : The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in Proceedings of LREC'2006, 2142-2147.

STEFANESCU D, TUFIS, D, IRIMIA E. (2006). Extragerea colocatiilor dintr-un text, in Resurse lingvistice si instrumente pentru prelucrarea limbii române, Universitatea Al.I.Cuza Iasi, 89-95.

TODIRASCU A. GLEDHILL C. STEFĂNESCU D. (2007). Extracting Collocations in Context: the case of Romanian VN constructions, in Proceedings of RANLP 2007, Sofia.

TUTIN, A (2004). Pour une modélisation dynamique des collocations dans les textes, EURALEX'2004, Lorient, France, 207-221.

TUFIŞ, D (1999). Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, pp. 28-33.

TUFIŞ, D (2004). Term Translations in Parallel Corpora: Discovery and Consistency Check. In Proceedings of the 4th LREC Conference, Lisabona, pp. 1981-1984.

TUFIŞ, D., et al. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In Proceedings of the EACL2006, Trento, Italy, 153-160

WANNER, L (1996). Lexical functions in lexicography and natural language processing, John Benjamins, Amsterdam/Philadelphia

WILLIAMS, G. 2003. Les collocations et l'école contextualiste britannique. In Grossmann, F. & Tutin, A. (éds). *Les collocations : analyse et traitement : Travaux et Recherches en Linguistique Appliquée* . Amsterdam : DeWerelt.