Extracting collocations in context: the case of verb-noun constructions in Romanian

Amalia Todirascu, Christopher Gledhill, Dan Stefânescu

Extracting Collocations in Context: the case of Romanian VN constructions

Amalia Todirascu
LILPA / Université Marc Bloch Strasbourg
22, rue René Descartes, BP 80010
67084 Strasbourg cedex
todiras@umb.u-strasbg.fr

Christopher Gledhill
LILPA / Université Marc Bloch Strasbourg
22, rue René Descartes, BP 80010
67084 Strasbourg cedex
todiras@umb.u-strasbg.fr

Dan Stefanescu
Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie, 13,
Bucharest 050711, Romania
danstef@racai.ro

Abstract

We present here a linguistic analysis of verbo-nominal (VN) constructions in Romanian with a view to developing a system for the extraction of lexical collocations from large tagged and annotated corpora. We identify the salient morpho-syntactic properties not only of the collocation but also of the context surrounding the expression.

Keywords

VN constructions, collocation extraction.

1. Introduction

This paper presents an on-going project for the *Agence universitaire pour la Francophonie* (AUF), whose aim is to develop an extraction tool for a multilingual collocation dictionary (German, French, Romanian). We focus here on the specific properties of Romanian collocations and on the linguistic resources developed to extract them from texts. Collocations are sequences of frequently co-occurring words which have a specific syntactic behaviour and a specific sense. Their idiomatic use is difficult for non-native speakers, and especially for Natural Language Processing (NLP) systems. Few dictionaries, whether traditional or electronic, provide complete information about collocations. While most explain the sense of idiomatic expressions, they often do not give any information about the morpho-syntactic behaviour of the expression. However, several methods and tools for extracting collocations from text have been developed.

Several definitions have been proposed for 'collocation' and few definitions are appropriate for the purposes of NLP systems. Collocations have been seen as "frequent word co-occurrence" [5], "a conventional way of saying things" [17] or a "fixed phrase" [10] [11]. As proposed in [6], three interpretations of the notion of 'collocation' are: **cooccurrence**, a statistical view [25]; **construction** (or 'colligation'), in terms of lexico-syntactic relations [12], and **expression**, a semiotic unit from the point of view of pragmatics [18],[8]. We adopted the lexico-grammatical view of collocation, assuming that a collocation is made up of a base and a collocate, and whose syntactic relations can be described in terms of a generic pattern (such as V + N, N + ADJ, ADV + ADJ etc.), used to automatically extract collocations.

In this paper, we focus on verbo-nominal (VN) constructions such as *make a decision / a lua o decizie, to make an application / a pune în aplicare* etc. VN constructions are associated with a subset of morpho-syntactic properties, such as a preference for the definite article or zero-article, for singular or plural noun, for the presence of an indirect complement, etc. These subregularities are important for an automatic extraction tool, since by using contextual information of this type, an NLP system can filter out salient collocations from a larger set of candidates, identified by statistical measures.

There have been several approaches which only use statistical methods for collocation extraction ([19], [21]), while other approaches identify collocations by purely looking at syntactic relations [24] or using both syntactic and semantic properties [27] [4]. In this paper we adopt a hybrid approach to extract VN constructions, in that we use a statistical module to extract VN co-occurrences and then apply a set of language-specific filters. The linguistic filters we use here were defined as a result of comparative linguistic data, carried out on a parallel corpus.

2. Methodology

We have adopted here a method which has already been applied to extract collocations from German corpora [14], [20]. These studies assume that collocations have their own morpho-syntactic properties. Their methodology has been used to analyze a large corpus in which any relevant morpho-syntactic information (preference for DEF ART, specific PREPs, case in German) is taken into account from the surrounding context of the expression.

In our project, a similar analysis has been applied to Romanian and to French. First, we identify common morpho-syntactic properties in the three languages. This is necessary in order to develop parametrizable tools for the automatic identification of collocation candidates. The next step involves a statistical module to establish a complete list of candidates, from parallel, tagged corpora [28]. Next, non-salient candidates are filtered out, using morpho-syntactic information. We are currently adapting several tools which already exist for German [16], French, and Romanian [26]. However, this process is only semi-automatic, a final manual check of candidates is necessary.

3. The Corpus

In order to identify language specific filters, we require tagged and preferably syntactically annotated corpora. We have used a parallel corpus available in the languages of the EU was used: the *AcquisCommunautaire Corpus* (ACC) [22], containing all the main legal texts published by the EU member states since 1950. We selected a set of common documents from the ACC in French, German and Romanian (about 15 million words for each language). The style of the ACC is impersonal, and it contains many domain-specific terms and fixed expressions, typical of administrative texts. In order to compare and to select only relevant collocations, it is necessary to compare our specialized corpora with more general text archives.

We set up various reference corpora containing similar genres (literature, newspapers, technical papers), to adjust the set of properties extracted from the ACC. We cleared these corpora of tables, pictures, irrelevant structural elements, and applied a uniform encoding to each. .For instance, the Romanian corpus about 10 million words: the RoCo corpus (newspapers); the NAACL corpus (newspapers, Romanian constitution and 2 novels); a philosophical treatise (Eliade); a medical corpus, the L4TE corpus (computer science). One problem was to select only texts with proper diacritics, because in Romanian the absence of diacritics might change the case or sense of the word, e.g. *fata* 'the girl' / *faţa* 'the face'.

In order to identify construction-specific morpho-syntactic properties, we use a tagged and syntactically annotated corpus. The French corpus has been tagged with a tagger trained on a corpus previously annotated using TreeTagger [23], while the Romanian corpus was tagged using the TTL platform [14].

Syntactic information is important to interpret the functional role played by a collocation or by various components co-occurring with the candidate. As the German corpus is annotated at chunk level, we annotated the French data at chunk level, using the Syntex parser [3]. and the Romanian data with the TTL platform.

4. V-N Collocations

As mentioned before, we hypothesize that VN constructions can be identified by finding collocations sharing several morpho-syntactic properties extracted from their immediate context. We are currently concentrating on Verb-Noun collocations, due to the productivity of this type of construction. For example, the light verbs [1] or support verbs that typically occur in VN constructions, such as *face / faire / make* or *lua / prendre / take*, have very different morpho-syntactic properties according to context, and a complete multilingual dictionary should explicitly represent this information. Generative grammarians [9] assume that these properties are determined by the specific type of 'predicate noun' alone, and they therefore minimize the role of the verb. Here we adopt a different perspective. As set out in [7], we propose that all VN constructions involve a 'generic' V which determines the argument structure of the predicate, and a 'specific' N which expresses the semantic process or 'range' ([2], [12]) of the predicate, as in *make a decision, take flight,* etc.

The most salient morpho-syntactic properties of VN constructions and the relation with the three levels of analysis can be seen in the following examples (from [7]):

V1. **Morphology.** Some VN constructions are related etymologically to a simple V (*to do work / to work, a se face noapte / a înnopta* 'to get dark'). But this equivalence is not always possible (*take a break / a face o pauză* is not the same as or is unrelated to *to break / *a pauza*)

V2 **Arguments.** Like simple Vs, VN constructions can take direct or indirect complements: *The candidate gave the electors a fright / Candidatul a băgat spaima în electorat, He put a brave face on the situation / A facut faţă situaţiei.*

V3 **Passive test.** Some VNs can have passive forms *(Pierre made a decision / Pierre ia o decizie vs. The decision was made by Pierre / O decizie a fost luată de Pierre*), but others do not : *to take flight / ?a flight is taken face obiectul, *obiectul a fost făcut* ' to be subject to …'. We have to mention that these examples are not translations of each other; they are intended to show the differences between Romanian and English.

V4. **Aspect.** Some VN constructions express perfective aspect [29]: *She laughed / She gave a laugh / She laughed for hours / ?She gave a laugh for hours*. In Romanian, this property is not available.

In addition, VN constructions also share some morpho-syntactic properties with Ns:

N1 **Determination**. The DET is often absent or fixed in many VN idioms (*take flight, a face obiectul* 'to be subject to'). When the N can be identified in referential contexts, the DET often becomes more variable (*to take an important decision, a luat o decizie importantă*).

N2 **Clefting**. The N in some VN constructions cannot be extracted (*He took flight / *It was the flight that he took El şi-a luat zborul / *Zborul pe care şi l-a luat*).

N3 **Expansion**. The N sometimes cannot be modified by relative clauses or other qualifiers (*He took the decision which was necessary / *He took the flight which was necessary*, *?El a luat decizia care se impunea, ?He took the flight which was necessary / *el şi-a luat zborul care se impunea*).

N4 **Conversion**. Some VN constructions cannot be nominalized (*The commission takes measures / Comisia a luat măsuri, The taking of measures by the commission / Luarea măsurilor de către comisie.*

So far we have evaluated these properties (V1-V4, N1-N4) in relation to French. In the following section we examine to what extent they apply to Romanian data, and we present some conclusions about the kinds of syntactic filters necessary to extract collocation candidates.

5. The Romanian Data

Romanian grammar is very close to Latin. Ns are characterized by the following properties: number, gender, and 5 cases. Case is marked by a specific ending (if the N is determined by an enclitic definite ART) or indefinite ART (*unei / unui / unor / unora* / = of some) or PREP (*pe*- literally 'on', for the accusative). The DEF ART is added as an ending for definite nouns (om***ului***, case***i***, oameni***lor***, case***lor***). Verbal morphology is characterized by mode (indicative, subjunctive etc.), tense (present, past, future…), number and person. The subject is not mandatory as in other Romance languages, and the perfect is usually formed with the auxiliary '*a avea* / to have'.. The passive is always made up of the auxiliary *a fi* / 'to be' followed by the past participle form of 'be', and by the past participle of the verb. The order of syntactic components is free.

5.1 The Case of *a face* (*to do* or *make*)

In order to identify the specific properties of VN constructions in Romanian, we studied the specific contextual properties presented in section 4. We looked in particular at morphology (V1, N1), the syntactic functions of the V and of the N, as well as their semantic roles. We searched for relevant information in the Romanian ACC corpus and in the general Romanian corpus.

VN constructions have several V-specific properties in Romanian. While V1-V3 are still valid tests for VNs, V4 (aspect) could not be used. For example, V1 applies to Romanian (the predicate can be replaced by a simple V), as in *a se face noapte* > *a înnopta* ('night falls', literally 'it makes dark'), *a face dovada / > a dovedi / to prove*). Several idiomatic expressions cannot be replaced by a simple verb (*a face faţă / *to make face > a faţa / to face*, *a face obiectul / to be subject to* but this is not the same meaning as *?a obiecta / to object*. The passive test (V3) is used to show that many of these expressions are idiomatic.

If the properties V1-3 apply to Romanian, although in different ways, as we have seen, the nominal properties N1-4 present some specific features. Extraction is not possible in Romanian. Expansion of the complement (N3) is however possible by modifying nouns with relative clauses: *al cărui obiect îl face* ('whose object is …'), *a cărei dovadă este...* ('whose proof is…') . The determiner (N1) is fixed in several idiomatic expressions: *a face obiectul* – *'*be subject to', *a face dovada* – *'*to make proof of'* (definite article), *a face faţă* – *'*to face' (no definite article),

5.2 Semantic Properties

In systemic functional grammar [13], the semantic role played by many nouns in VN constructions is known as 'process range'. The process range expresses the semantic process of the predicate, and is often integrated into the verb group [7] (as in *a face obiectul* *'*to be subject to…'). Any indirect complement which follows this element then becomes the semantic object (or 'goal'). In French and English, this indirect complement is usually introduced by a PREP, but in Romanian this role is filled by the genitive case. In (1), the complement expresses a simple relational process. However, in (2) we have more complex situation (subject reading):

(1)*...să facă obiectul unei proceduri administrative...*
'is the subject of an administrative procedure'

(2*) ...la instituţiile financiare, care fac parte din categoria….*
'in financial institutions which are part of this category'

The most frequent collocations of *face* in the Romanian Acquis Communautaire are VN constructions where the N has been integrated into the verb group (VG). In French, it is possible to establish a relation between specific types of ART (definite, indefinite and zero) and a specific process type (e.g. material processes tend to be definite) [7]. But again, this is not possible for Romanian; VN constructions with a definite suffix (*face obiectul, face dovada, face legătura*) are mostly relational process, and the process range is expressed by the indirect complement:

(3)…*Trece peste graniţa dintre statele membre şi care face legătura între sistemele de transport…*
'…crosses the border between member states and which joins the transportation system…'

In VN constructions where Ns have indefinite ART (*fac+un / o / unele / nişte* + N) several semantic processes can be identified: mental (verbal communication, as (4) or material as in (5):

(4) *se face un proces verbal al fiecărei şedinţe a …*
'Minutes shall be taken of all meetings'

(5) *Comisia poate să facă orice modificări la prezentul Regulament care …*
'The commission should make some changes in the present rules…'

Among VN constructions without articles, we found several relational process: (*a face faţă / to face, a face parte / be part of, a face obiectul / is subject to)* :

(6) *Pentru a putea face faţă unor situaţii de urgenţă*
'in order to deal with emergency situations'…

Other VN constructions where the DET is absent are mostly material intransitive processes: *face vizite/ to pay visits, face comerţ/*to make trade.*

We conclude that in Romanian, as with English and French, there is a certain tendency for groups of words to lexicalize with a corresponding rigidity of morpho-syntactic features (preference for indefinite ART, systematic use of some specific classes of PREP etc.). These features are relevant to a module for filtering such expressions.

6. Automatic Extraction

As presented in section 2, our extraction approach combines statistical techniques and pattern-based matching in order to filter candidates.

6.1 The Statistical Module

Verb Noun pairs co-occurring together frequently (separated by one or several words) are potential collocation candidates. We have applied a statistical module for extracting V-N pairs from the corpora, based on [21], using mean and variance. The mean is the average of the distances between the words forming the pair, while the variance measures the deviations of the distances with respect to the mean already computed. Collocations are pairs of words for which the standard deviations of distances are small. We computed the standard deviation for all V-N pairs (from the ACC corpus) within a window of 11 content words length for all the three languages involved in the project and we considered as good, all the pairs for which standard deviation was smaller than 2 [21] .

We want to further filter out some of the pairs so that we keep only those composed by words which appear together more often than expected by chance, using Log-Likelihood (LL). The idea behind the LL score is finding the hypothesis which describes better the data:

$H_0 : P(w_2|w_1) = p = P(w_2|\neg w_1)$
(null hypothesis - independence)
$H_1 : P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1)$
(non-independence hypothesis)

The LL score formula is:

$$LL = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} * \log \frac{n_{ij} * n_{**}}{n_{i*} * n_{*j}}$$

where nij represents the number of occurrences when the words wi and wj appear together, ni* is the number of occurrences for wi together with any wj, etc.

We computed the LL score for all the pairs obtained by the first method.. We kept in a final list the pairs for which the LL score was higher than 9 (see the table for Chi-square distribution with one degree of freedom). Using LL filtering, we obtained a list of candidates for Romanian collocations (table.1) Among the top pairs extracted, we identify some valid candidates, expressing processes (*face obiectul, aduce atingere, intra în vigoare, face modificări*), but the other candidates are not collocations (morpho-syntactic properties are variable). The *face+noun* constructions identified among the first 20 candidates are collocations and have specific morpho-syntactic properties (no article or definite, preference for singular). For all these pairs, we apply linguistic filters to select valid candidates.

Fig.1 First LL score

w_1 w_2	dist.	LL score	Process
Aduce atingere 'to affect/to prejudge'	1	51567.34864	Relation process
înlocui text 'replace text'	3	43992.3067	-
intra vigoare 'applied' (or literally 'placed in vigour')	2	42527.03736	Relational process
Face apel la 'call for' (or literally 'to make a call'))	3	32050.11219	Relational process
face obiect 'be subject (to)'	1	30729.47663	Relational process
Face modificări 'make changes'	4	29141.39454	Material process

6.2 The Filtering Module

As we saw in section 5, some Romanian collocations have specific morpho-syntactic and semantic properties. We use these properties to extract relevant candidates from the statistical module output. We mainly use a set of patterns, manually defined, based on linguistic analysis.

One example of an extraction pattern identifies the sequence P (predicate) + C (complement) (direct) + C (indirect), or in tagged code « *a face* NxRY *{1,5} NxOY»,: NxRY means Noun (plural or singular), in direct case (Nominative or Accusative definite form); NSOY means Noun, singular, oblique case (Genitive or Dative case definite form); {1,5} means 1 up to 5 words. This sequence alone can identify four valid VN constructions among the candidates proposed by the statistical module: *face obiectul, face dovada, face subiectul, face transferul*. Another pattern for *face* constructions combined with the prepositon *cu* (with) (face NxRY *{1,5} cu) identifies some interesting candidates: *a face legătura cu (makes a link with), a face declaraţia cu privire la (make a declaration in relation to…)*. These candidates involve various relational processes: *a face legătura cu* ('relate'), *a face transferul* ('transfer'), but also some communicative processes as well *a face declaraţia cu privire la (*'to declare'). In addition, V+*în / in* selects candidates as *înlocui în text* ('to place in text'), *intra în vigoare* ('to apply / to enter into force').

7. Conclusion

The paper has presented some features of VN constructions in Romanian. Generally speaking, Romanian shares most of the properties of VN constructions that have been identified for Western European languages. The difference is that the specific configuration for each VN construction is different. The verb *a face* (equivalent to French *faire*) operates syntactically in the same way as *faire*, but does not cover the same semantic ground. It is also clear from this study that the relevant context for all of these expressions extends way beyond the basic V plus N collocation: in almost every case, the expression involves a specific morpho-syntactic configuration and has a phraseology and context of use which is highly consistent. Our conclusion must therefore

be that the contextual features of VN constructions are crucial to the semi-automatic extraction of collocations.

8. Acknowledgements

This work has been funded by Agence Universitaire pour la Francophonie (AUF). We thank Rada Mihalcea (University of Texas, United States) for the NAACL corpus, Dan Tufis (Romanian Academy) for the tagging tools, as well as Dan Cristea (University of Iasi, Romania) for the L4TE corpus.

9. References

[1] Allerton, D., 2002. Stretched Verb Constructions in English, London, Routledge.

[2] Banks, D., (2000). The Range of Range: A transitivity problem for systemic linguistics, *Anglophonia,* 8, 195-206.

[3] Bourigault D. & Fabre C.(2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaires,* n° 25, pp. 131-151.

[4] Bolshakov, I.A., Gelbukh. A. (2001). A Very Large Database of Collocations and Semantic Links. NLDB'2000. Lecture Notes in Computer Science N 1959, Springer-Verlag, pp. 103–114

[5] Cowie, A. P. (1981) The treatment of collocations and idioms in learner's dictionaries, in *Applied Linguistics*, 2(3), 223-235.

[6] Gledhill, C., (2000). Collocations in Science Writing*,* Gunter Narr Verlag, Tübuingen

[7] Gledhill (2007) La Portée : seul dénominateur commun dans les constructions verbo-nominales. In Frath, P. Pauchard, J. & Gledhill, C. (Eds.), 2007, *Actes du 1er Colloque, Res Per Nomen*, Université De Reims-Champagne-Ardenne, 24-26 Mai 2007 : 113-125.

[8] Gledhill, C, Frath, P., (2007) Collocation, phrasème, dénomination: vers une théorie de la créativité phraséologique, in *La Linguistique*. Vol 1/1.

[9] Grimshaw, J. & Mester, A., (1988). Light Verbs and θ-Marking, *Linguistic Inquiry,* 19, 205-232.

[10] Gross, M. (1993) Les phrases figées en français. *L'information grammaticale* 59, Paris, 36-41.

[11] Grossmann, F., Tutin, A.(eds.) (2003). Les collocations: analyse et traitement, Numéro special: *Travaux et Recherches en Linguistique Appliquée*.

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

[13] Halliday, M., (1985). *An Introduction to Functional Grammar*. London, Arnold.

[14] Heid, U., Ritz, J. (2005) Extracting collocations and their contexts from corpora, *Actes de COMPLEX-2005*, Budapest.

[15] Ion, R. (2007). TTL: A portable framework for to-kenization, tagging and lemmatization of large corpora. Research Institute for Artificial Intelligence, Romanian Academy, Bucharest (in Romanian), 22p.

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

[17] Manning, C., Schütze, H. (1999), Foundations of Statistical Natural Language Processing, MIT Press, Cambridge.

[18] Moon, R., (1998). *Fixed Expressions and Text.* Oxford, Oxford University Press.

[19] Quasthoff, U. (1998). Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values. In *Proceedings of LREC'02*, ELRA, S. 853-856.

[20] Ritz, J., Heid, U. (2006) Extraction tools for collocations and their morphosyntactic specificities, in: *Proceedings of LREC-2006*, Genova, Italia, 2006.

[21] Smadja, F. A., McKeown, K. R. (1990), Automatically extracting and representing collocations for language generation, *Proceedings of ACL*, 252-259, Pittsburgh..

[22] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C. Erjavec, T., Tufiş, D., Varga, D. (2006), The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, *Proceedings of LREC'06*, pp.2142-2147.

[23] Schmid, D. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.

[24] Seretan, V., Nerima, L., Wehrli, E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora, *Proceedings of EURALEX'2004*, Vol2, pp.755-766.

[25] Sinclair, J., (1991). *Corpus, Concordance, Collocation,* Oxford, Oxford University Press.

[26] Stefanescu D, Tufis, D, Irimia E. (2006) Extragerea colocatiilor dintr-un text, Atelierul « Resurse lingvistice si instrumente pentru prelucrarea limbii române », Iasi.

[27] Tutin, A (2004). Pour une modélisation dynamique des collocations dans les textes*, Actes du congrès EURALEX'2004*, Lorient, France, 2004, Vol. 1, 207-221.

[28] Tufiş, D., Ion, R., Ceauşu, A., Stefănescu D. (2005). Combined Aligners. In *Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*, Michigan, pp. 107-110.

[29] Wierzbicka, A., (1982). 'Why can you Have a Drink when you can't Have an Eat?', *Language,* 58.