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

Amalia Todirascu, Christopher Gledhill

Todiraşcu, Amalia & Christopher Gledhill. 2008. Extracting collocations in context: the case of verb-noun constructions in English and Romanian. *Recherches anglaises et Nord-américaines* 41. 107-122. **(PDF)**

Extracting Collocations in Context : The case of Verb-Noun Constructions in English and Romanian

Amalia Todirascu & Christopher Gledhill

Université Marc Bloch, Strasbourg.

Abstract

Verb-Noun (VN) constructions involve a 'generic' V plus a 'specific' N which is either (i) a structural qualifier of the Predicator as in *make* haste, or (ii) a syntactic Complement as in *to make* *a suggestion*. In each case the N expresses the semantic Range of the VN construction (Banks 2000, Gledhill 2007). However, despite much research on 'support verb' or 'light verb' constructions, there is no one morpho-syntactic feature which allows us to distinguish these constructions from any other sequence of V plus N, at least in English. From the point of view of corpus linguistics, this lack of formal marking makes it hard to extract them on a semi-automatic basis. As part of an on-going lexicographic project[1], we have examined various computational models in order to extract VN constructions from multilingual corpora. One of our main findings is that statistical methods alone are not sufficient; the collocations that are thrown up in the data extend from a few 'relevant' VN constructions, to rather too many 'semi-relevant' VN co-occurrences and 'irrelevant' noise.

Les constructions Verbo-Nominales (VN) sont composées d'un V générique et d'un N spécifique, lequel est soit (i) qualifieur structurel du Prédicateur comme faire faillite, *soit (ii) Complément syntaxique comme* faire des recommandations. *Dans chaque cas, le N exprime la Portée sémantique de la construction (Banks 2000, Gledhill 2007). Mais malgré quantité d'études sur les 'verbes supports' ou 'verbes légers', aucune propriété morpho-syntaxique ne permet de distinguer ces constructions des autres séquences V plus N, au moins en anglais. Du point de vue de l'analyste de corpus, ce manque de marque formelle rend difficile la conception d'un outil d'extraction automatique. Dans le cadre d'un projet lexicographique[1], nous avons étudié plusieurs modèles destinés à extraire des VN des corpus multilingues. Nos résultats préliminaires indiquent que des méthodes purement statistiques ne sont pas suffisantes : les données révèlent parfois de 'véritables' constructions VN, mais aussi pour la plupart des exemples 'non-pertinents' de cooccurrences VN, ou tout simplement du bruit.*

Introduction

This paper presents work in progress in the field of applied lexicography for the *Agence universitaire pour la Francophonie* (AUF). The aim of our project is to develop an extraction tool for a multilingual collocation dictionary (in German, French and Romanian). For the purposes of this paper, however, we focus on the properties of English and Romanian collocations and on the computational resources developed to extract them from electronic corpora. The particular phenomenon we are interested in is that of Verb-Noun (VN) constructions, a formal term which avoids the restrictions implied by the various competing

[1] Funded by the *Agence universitaire pour la Francophonie* (reported in Gledhill et al. 2007)

terms that have been used in the literature, such as ‘support verb’ (Gross 1989), ‘light verb’ (Pottelberge 2000), or ‘stretched verb’ (Allerton 2002).

Before embarking on our computational study, it is necessary to make clear two fundamental assumptions that have emerged during the course of this project:

(a) collocations are not items, collocation is a relation between items in the lexical system (Gledhill 2000).

(b) the only common denominator in VN constructions is that the N expresses the semantic Range[2] of the Predicate (Banks 2000).

1. Preliminary remarks on the notion of ‘collocation’

Let us first examine assumption (a). Several definitions have been put forward for ‘collocation’ in the lexicographic and phraseological tradition, as well as by NLP researchers, grammarians and corpus linguists. Rather than select a single definition, Gledhill (2000) and Frath & Gledhill (2005) propose that collocation involves at least three different perspectives: (i) **cooccurrence**, a statistical view, which sees collocation as the recurrent appearance in a text of a *node* and its *collocates* (Firth 1957, Sinclair, 1996), (ii) **construction,** which sees collocation either as a lexical-grammatical colligation (Hunston and Francis, 2000), or as a relation between a *base* and its *collocative partners* (Haussmann 1989) and (iii) **expression**, a pragmatic view of collocation as the relation between a sign and its function (Moon 1998, Gledhill and Frath, 2007). It should be pointed out here that these different perspectives contrast with the usual way of presenting collocation in phraseological studies. Traditionally speaking, collocation is explained in terms of all three perspectives at once, in a continuum:

‘Free Combination’ ↔ ‘Bound Collocation’ ↔ ‘Frozen Idiom’

We would argue that this linear view promotes the idea that only some items are ‘collocations’, while others are uninteresting ‘combinations’, or unusual ‘idioms’. To understand this point, let us take a series of examples, as set out in the following table:

A	*I kicked the dog*	*I’m making a cake*	*What do I do now?*
B	*You kicked the habit*	*You made a suggestion*	*Did you do the washing-up?*

[2] Terms which are specifically part of the Systemic Functional model (Halliday 1985, Halliday & Matthiessen 2004) are presented here in capitals.

C	*He nearly kicked the bucket*	*She just made it!*	*That'll do me*

The traditional, phraseological, view would be to argue that examples (A) are 'free combinations', (B) 'interesting collocations' and (C) 'idioms'. However, we would argue the following:

- All examples (A, B and C) involve *co-occurrences* of a V plus N (V+N)
- All examples (A, B and C) involve *constructions* of the sequence V and N (V ^ N)
- All examples (A, B and C) are *expressions* (as signs, it is possible to 'promote' any of the examples in A and B to the status of 'idioms', just as it is possible to 'demote' the C examples the status of syntagmatic constructs).

So what is difference between examples (A), (B) and (C)? We would suggest the following:

- Examples (A) all involve more productive grammaticalised constructions
- Examples (B) all involve 'Process Range' constructions ; (this point is developed below).
- Examples (C) all involved less productive lexicalised constructions

In the lexicographic project that we are engaged in, we are particularly interested in examples such as (B) *make a suggestion* or *do the washing-up*, and (less crucially) examples such as *kick the habit.* Our problem therefore is how to identify the VN constructions which appear to lie at the centre of the lexical-grammatical scale, that is to say somewhere between examples (A) and (C).

2. Preliminary remarks on the notion of 'Verb-Noun construction'

We now turn to assumption (b). Constructions such as *make a suggestion* and *do the washing-up* have been analysed in various ways. Formalist grammarians (e.g. Giry-Schneider 1987, Grimshaw & Mester 1988, Kearns 1989, Di-Scullo & Rosen 1991, Gross 1998, Kim 1998, Allerton 2002.) assume that VN constructions are idiomatic in nature, in that their verbal nature is essentially determined by the N (a 'predicative noun'). Thus formal analysis assumes that a VN construction is the functional and structural equivalent of a simple V, indeed that the VN construction is derived from a deep structure Predicate, as reflected in the terminology (light verb, support verb, stretched verb …). In this paper, we adopt an alternative analysis from the point of view of systemic functional grammar (Halliday, 1985, Banks, 2000). This approach emphasises the textual role of VN constructions. For example, a communicative choice is often made between a congruent form (simple V) emphasizing a lexical product (*to suggest something*), or a metaphorical form (VN) emphasizing a lexical process (*to make a* *suggestion*). Gledhill (2007) has argued that VN constructions are best analysed from three different points of view: Syntactic Function (where the N constitutes an independent Complement as in *make a suggestion*), Lexical Structure (where the N is integrated into the VG as in *make haste*), and Semantic Role (also known as Transitivity). It is at the level of transitivity that VN constructions differ from all other V+N collocations. As Gledhill (2007) has pointed out, the N in all of these expressions, regardless of Function or Structure, the N expresses or qualifies the semantic Process in the Predicate. This role is termed Range (Halliday 1985, 149, Banks 2000), a term which is used both for 'cognate' Complements, in such examples as *sing a song, live a long life, tell tales,* as well as 'process' Complements such as *make a suggestion* and *do the washing-up*. The cognate examples in particular show that Range is not limited to generic or 'light' VN constructions. In the construction *kick the habit*, *kick* prototypically expresses a concrete Material Process, but in the context of *habit* the Predicate is metaphorically specified as a Behavioural Process.

In summay, Range is the only factor which allows us to distinguish between VN 'co-occurrences' of the (A) type *make a cake, to do something,* and VN 'constructions', of the (B) type *make haste, make a suggestion*. Contrary to the claims of formal grammarians, there appear to be no particular morpho-syntactic properties which allow us to distinguish VN constructions and VN co-occurrences. This does not mean however these linguists have not tried to find one, as Pottelberge (2000) points out. And of course, it does not follow that because we cannot find specific features in English, they cannot be found in another language. However, as we see below, we believe that the situation also appears to apply to a Romance language such as Romanian.

In the rest of this section, we summarise some of the features that have been proposed as defining features of VN constructions (as set out in Gledhill 2007), and we examine to what extent they do or do not apply to both English and Romanian. In the list below, features V1-4 compare the properties of VN constructions with those of simple Vs:

V1 **Equivalence.** VN constructions consist of a 'generic' V and a 'specific' N. Some VNs are related etymologically to a simple V (*to do work / to work, to make a suggestion / to suggest,* `a se face noapte` *'to make night',* `a se înnopta` *'to become night'*). But this equivalence is not always possible (*take a break* =? *to break,* and the Romanian equivalent `a face o pauză` / `*a pauza`).

V2 **Argumentation.** Like simple Vs, VN constructions can take a variety of arguments, all realized as indirect Complements: *The candidate gave the electors a fright,* Candidatul a băgat spaima în electorat, or *He 'called / made a call' to his collegue,* A făcut apel la colegi.

V3 **Voice.** Some VNs allow for the passive *(Pat made a decision / The decision was made by Pat*), but others do not (*They took flight / ?Flight was taken by them.*). Where the N is an extension of the Predicator, as in the Romanian face obiectul *'to be subject to'*, N is not a Complement and the passive is blocked, hence *Obiectul a fost făcut.

V4 **Aspect.** As various linguists have suggested (Wierzbicka 1982, Vivès 1984, Cotte 1998), VN constructions in English express lexical aspect : *She laughed / She gave a laugh / She laughed for hours / ?She gave a laugh for hours* (Achievement). It is interesting to note that the presence of a Range element appears to have an effect on aspect (*Mary rolled for / ?in three seconds. (*Activity) / *Mary did a forward roll ?for / in three seconds*. (Accomplishment). This does not appear be a feature of Romanian.

A second set of criteria relate to the function of the N in the VN, or the role of the V when in nominal form:

N1 **Determination**. The determiner is often absent or fixed, especially when the N is an integrated qualifier of the Predicator (*take flight, make haste,* face obiectul / *face obiect / *'to be subject to'*). In discourse, however, the situation becomes variable (*He took an important decision / He took the decision which was necessary,* and in Romanian: ia o decizie *'take a decision',* ia decizia necesară, *'take the necessary decision'*).

N2 **Clefting**. The N in some VN constructions cannot be extracted in a cleft clause (*He took flight / *It was the flight that he took*), but it can in others (*This is the suggestion he made*). In Romanian (Profesorul a luat cuvîntul *'Professor-the has taken word-the' 'The professor made a speech',* but not *Cuvîntul pe care l-a luat profesorul *'The word that the professor has taken'.*)

N3 **Qualification**. The N in some VN constructions can be modified by relative clauses or other qualifiers (*She made sense, She made more sense than him, He took the decision which was necessary,* in Romanian el a luat decizia care se impunea), but there are exceptions (*They took flight / ?They took several flights / ?They took the flight which was necessary,* *el a luat cuvîntul care se impunea *'he has taken word which was imposed'.*)

N4 **Conversion**. The N in some VNs can be nominalised and made into a discourse referent (*The commission took measures / The taking of measures by the commission*). Once again there are exceptions, in both Romanian and English.

It can be seen, especially in the second series of features (N1-4), that the relative 'fixedness' of VN constructions is largely dependent on whether the V or the N can be used as a discourse referent. As far as we are aware, this sort of textual variation is not something that computational linguists have examined in any detail. We believe therefore that no one morpho-syntactic feature stands out as unique in any of the expressions we have examined so far. However, we are at least in a position to say why we find constructions such as *make a suggestion* ‘relevant’ for our project, whereas co-occurrences such as *make a cake* are ‘not as relevant’, at least from a semantic point of view: VN constructions involve a Range element which fundamentally affects our interpretation of the Process.

We now turn to the question of computational analysis. In the following section, we set out the existing possibilities for extracting collocation candidates from a corpus of texts. Although we are not currently in a position to find VN constructions automatically, we can at least begin to form judgements about the quality of VN collocations that are ‘thrown up’ by the various tools that have been developed. The question we ask in the following sections is therefore: to what extent are these results ‘relevant’ VN constructions, ‘semi-relevant’ VN co-occurrences, or ‘irrelevant’ noise.

3. Methodology

In this section we discuss the tools that have been designed to extract collocations from texts and the electronic corpora we have analysed in our project. The tools we use to do this were initially developed by our project partners for German corpora (Heid and Ritz, 2005, Ritz and Heid, 2006). In order to extract collocation candidates from texts, a statistical module is used to establish a complete list of candidates, using parallel, tagged corpora (Tufis and al, 2005). This is a necessary step, because in order to obtain VN co-occurrences in the first place it is necessary to identify and mark up all the possible Ns and Vs in the corpus. We are currently adapting several tools existing for this process for German (Kermes, 2003), French and English (Rousselot and al, 2004), as well as Romanian (Todirascu et al, 2007, Stefanescu and al, 2006).

We have so far used several corpora for our main lexicographic project, although for the purposes of this paper we need only discuss one: the *Acquis Communautaire Corpus* (ACC), a very large parallel corpus of legal texts available in all the official EU languages (Steinberger and al, 2004). For each language, the corpus contains around 20 million tokens. The ACC contains all of the main legal texts published by the EU member states since 1950. The ACC is not a reference corpus: it is a highly specialised ‘LSP’ corpus, with a highly impersonal style and containing many domain-specific terms and fixed expressions which are typical of administrative texts. This point becomes particularly salient when we consider some of the specific co-occurrences and constructions that emerge from the data analysis.

For our analysis and tools, we require tagged corpora. Part-of-Speech (POS) taggers use morpho-syntactic information and are usually very robust. However, this process is still fraught with technical difficulties, not the least of which is the fact that each tool is designed for different languages, with different degrees of tagging success. The French, English and German corpora were tagged using TreeTagger (Schmid, 1994). While this tagger had previously been trained on newspaper texts, many lemmas or tags proposed for the ACC were wrong. We trained the tagger for the new domains, after correcting lemmas and tags. A manual validation was then done after automatic tagging. The Romanian corpus was tagged using TTL – a complex tool for pre-processing texts (Ion, 2006), and the tagged data were also validated manually. To give an idea of how problematic and how detailed this process can become, here are some examples of tags used to describe Romanian filters: *NxRY* – Noun (plural or singular), in direct case (Nominative or Accusative definite form); *NSOY* – Noun, singular, oblique case (Genitive or Dative case definite form); *V3* – Verb (3rd person), and so on.

It is only after these initial 'pre-processing' stages, that we are then able to identify lists of VN co-occurrences or, in NLP terms, 'collocational candidates'. From a statistical perspective, all VN co-occurrences, that is to say all collocations of a V and N *in any order* and *separated by one or several words* are potential collocation candidates. In order to calculate the most salient VN pairs, we applied a statistical module (Stefanescu et al, 2006, Todirascu et al, 2007) for extracting VN co-occurrences from corpora, based on a solution proposed by Smadja & McKeown, (1990). This programme looks for pairs of words for which the standard deviations of distances are small. The next stage involves filtering out some of the pairs using Log-Likelihood (LL) score and then computing the LL score for all the pairs obtained using Smadja's method. Using LL filtering, we finally obtained a candidate list of VN co-occurrences, ordered by LL score and the distance between the base and the collocate, for English and Romanian. These results are set out in the following section (3.1).

Although our ultimate objective is to design a semi-automatic method for identifying VN constructions, at present the final stages of our analysis must be carried out manually, that is to say by a linguist. In this case we have sorted VN co-occurrences into two broad categories, 'relevant' constructions and 'non-relevant' co-occurrences. These are set out and discussed in section (3.2) and (3.3).

3.1 VN co-occurrences in the corpus

As mentioned, our approach to the identification of VN collocations is hybrid, combining statistical techniques and pattern-based matching in order to filter candidates. In the following tables, we present the first 20 VN co-occurrences that emerge from first the English ACC corpus and then the Romanian ACC corpus (it can be seen that some items have been lemmatized, and thus some examples involve Ns which have been mistagged as Vs, or vice-versa):

TABLE 1. VN co-occurrences in the ACC Corpus (English).

	W1	**W2**	**DIST.**	**LL**
1	have	regard	1	139337.613681525
2	do	brussels	2	58421.7707215154
3	treaty	establish	1	55994.7655599668
4	regard	establish	4	48903.9598951192
5	have	treaty	4	40571.0339233016
6	bind	entirety	3	39298.8622283224
7	have	european	7	32703.6372333056
8	regulation	bind	3	30228.6672675113
9	replace	following	3	28022.7117842671
10	day	follow	1	27337.5653023155
11	bind	states	10	27317.2237989513
12	take	account	1	26833.7653197018
13	bind	member	9	25419.1121599778
14	address	states	4	21930.0382277717
15	provide	opinion	10	21920.2165640043
16	publish	journal	4	21307.1284199788
17	have	opinion	4	21093.5576942016
18	enter	day	6	20579.6098506301
19	publish	official	3	20549.9777229096
20	address	member	3	19934.5794559521

TABLE 2. VN co-occurrences in the ACC Corpus (Romanian).

	W1	**W2**	**DIST.**	**LL**	**(Literal Equivalent)**	
1.	aduce	atingere	1	51567.34864	'bring	prejudice'
2.	înlocui	text	3	43992.3067	'replace	'text'
3.	intra	vigoare	2	42527.03736	'enter'	'vigour'
4.	avea	tratat	3	32050.11219	'have'	'treaty'
5.	face	obiect	1	30729.47663	'make, do'	'object'
6.	modifica	regulament	4	29141.39454	'modify'	'rule'
7.	modifica	dată	2	27658.4116	'modify'	'date'
8.	lua	considerare	2	27062.0349	'take'	'consideration'
9.	ține	cont	1	26635.12649	'take'	'account'
10.	adresa	membră	2	25844.0428	'adress'	'member'
11.	articol	intra	4	24921.96291	'article'	'enter'
12.	adresa	stat	1	23343.86292	'address'	'state'
13.	ține	seamă	1	22825.70709	'take'	'account'
14.	element	aplica	4	21924.02349	'element'	'apply'
15.	adopta	bruxelles	2	21792.22915	'adopt'	'brussels'
16.	adopta	regulament	2	20847.73793	'adopt'	'rule'
17.	articol	adresa	5	19716.52613	'article'	'adress'
18.	lua	măsură	1	19207.12849	'take'	'measure'

19.	regulament	intra	1	18726.54795	‘rule’	‘enter’
20.	iunie	privi	2	14661.84913	‘June’	‘related’

[W1 W2 = co-occurrence of V+N or N+V, the order being determined by the most frequent item. DIST = average distance between each co-occurrence. LL = log-likelihood score.]

The next stage in our analysis involves manually analysing the contexts of these VN pairs and identifying several various categories of data. As can be seen in tables 1 and 2, very few of the examples thrown up by the corpus analysis actually correspond to ‘valid’ or ‘relevant’ VN constructions. Generally speaking, it would appear that ‘valid’ VN constructions represent only a minority of possible VN co-occurrences; there is however a tendency for VNs to occur more frequently in Romanian.

3.2 ‘Relevant’ VN constructions

We present here three main types of ‘valid’ or ‘relevant’ VN construction (A-C) which emerge from our data analysis, as well as their properties in both English and Romanian.

(A) **Predicator + N (Range)**. This category involves VN constructions in which the N expressing the semantic Range of the clause is structurally speaking a qualifier within the Verb Group, and not a syntactic Complement. In many examples, an indirect Complement is present. This element is usually marked in English by prepositions such as *to*, *of*, *over* and has the Semantic Role of Goal or Object:

(1) *to **take account** of the provisions of this Regulation*
(2) *this Article shall **give rise** to an obligation to export to the destination indicated*
(3) *Member States shall ensure that insurance claims **take precedence** over other claims on the insurance undertaking according to one or both of the following methods…*

In Romanian, the indirect Complement is signalled by the genitive or a preposition:

(4) Dispoziţiile prezentului regulament nu **aduc atingere** îndeplinirii oricăror obligaţii
‘The articles of this rule do not **affect** [‘bring predjudice’] the accomplishment of any duty’
(5) …întrucât se cunoaşte că anumite probleme care **fac** în prezent **obiectul** măsurilor tranzitorii nu vor putea fi soluţionate…
‘because it is well known that some problems which **are subject** to transitory measures could not be solved…’
(6) perioada normală de timp necesară pentru deplasare **ţinând cont** de mijloacele de transport şi distanţele implicate
‘the regular time interval required to travel, **taking** into **account** the travel means and the distances’

(B) **Predicator + Complement / Adjunct (Range)**. In these constructions, Range (the semantic Process of the Predicate) is expressed by a Complement. Few examples emerge in the data analysis, although there are many to be found lower down the frequency table, with a very productive cluster of examples around the generic verbs *make* and *take*:

(7) *…the methods of sampling and analysis used for this purpose can **have** direct **repercussions** on the establishment and functioning of the common market…*
(8) *it is appropriate to **make** this **distinction** specifically in the case of the creation of joint ventures*
(9) *Article 5 The Commission may **make suggestions** to the Member States as to the coordination of their control activities in accordance with Community regulations*
(10) *the competent authority shall **take** appropriate **action**.*
(11) *Member States should **take** due **consideration** of the requirements of Articles 12 and 13 of this Directive …*
(12) *The Council shall **take** a **position** on these Commission proposals by 30 June 1989.*
(13) `partidul său` **`va lua măsuri`** `împotriva senatorului Pruteanu`
'his party will **take measures** against Senate member Pruteanu'

An important variation on this theme in English (reported in Gledhill 2005) involves an Adjunct which expresses a (Mental) Process Range , followed by a direct Complement which is the Goal or Phenomenon.

(14) *Member States do not exhaust the fishing possibilities provided for in the Protocol, the Commission may **take into consideration** licence applications from any other Member State.*
(15) *under the legislations of two or more Member States, the benefits shall be granted to him, **taking** the aggravation **into account**, in accordance with the provisions of Article 40 (1).*

(C) **Predicator + N + Projected Clause**. This category, which includes complex Predicators as well as complex Predicates, involves legitimate VN constructions which introduce a further clause complex, in particular the very productive sequence *have + effect of* + V-ING:

(16) *These differences **have** the **effect** in particular of increasing the risk of misuse and for that reason, the Schengen States are introducing a document incorporating features aimed at preventing counterfeiting and falsification.*
(17) *a number of the sugar-beet or cane producers directly affected by one of the operations referred to in paragraph 1 expressly **show their willingness** to supply their beet or cane to a sugar-producing undertaking which is not party to those operations….*

It would appear that no candidates of this type were identified in Romanian.

3.3 'Non-relevant' VN co-occurrences

The following categories of VN co-occurrence (D-J) are 'non-relevant' in the sense that the N in the VN pair does not contribute to the expression of semantic Range. Such negative data are important to our study, in the sense that these occurrences often belong to constructions of a different type and our study must eventually rule them out in some way.

(D) **Preposition Group + Completive**. This category involves examples which resemble legitimate Range constructions of the (A) type (Predicator + N), but which are in fact non-Finite verbs forming complex Prepositional Groups often with the syntactic function of Adjunct:

(18) ***having regard*** *<u>to the Treaty</u> establishing the European Economic Community…*
(19) *this sum has to be determined* ***having regard*** *to prices recorded on the markets during a reference period…*

This usage is frequent in the ACC; the expressions involved being so formulaic that a number of long-range collocations emerge as secondary data, as in the following:

(20) ***Having*** *regard to the* ***Treaty***
(21) ***Having*** *regard to the* ***opinion*** *of the Committee of Regions*

(E) **Predicator + Complement**. This category involves cases of VN co-occurrence which resemble type (B) constructions, but in which the Complement has a semantic role other than Range (this category includes many examples of Complement ^ Predicator order, associated with various types of relative structure):

(22) *No animal admitted to the semen collection centre may* ***show any clinical sign of disease*** *on the day of admission.*
(23) *in the light of* ***experience gained*** *<u>by the Commission</u>, at the time…*
(24) **Regulamentul adoptat** de cãtre Comisie privind stocarea datelor..
'The **rule adopted** by the Commission concerning data storage…'

(F) **Predicate + Complement (Intensive) / Predicate + Adjunct**. This category of VN co-occurrence involves Complements of a Relational Process or Adjuncts following a passive construction. This accounts for the majority of collocations emerging in the English results, including examples such as:

(25) *This Directive is* ***addressed*** *to the Member* ***States***
(26) *The date of entry into force of the Agreement will be* ***published*** *in the Official* ***Journal*** *of the European Communities by the General Secretariat of the Council*
(27) *This Decision is* ***addressed*** *to the* ***Member*** *States*
(28) Fiecare stat membru **adresează** un raport celorlalte state **membre** şi Comisiei
'Each member state adresses a report to the other members and to the Commission…'

(G) **Subject + Predicator**. In these cases, the N in the VN co-occurrence has the function of Subject, and is therefore strictly speaking a NV construction:

(29) Acest **Regulament** va **intra** în vigoare la trei zile după publicarea lui în Jurnalul Oficial
'this Rule will come into force three days after its publication in the Official Journal'

(H) **Predicate + Adjunct.** In this category, the N is an element in an Circumstantial Adjunct, in many cases indicating Location (24, 26), Temporal Extent (25, 28) or Manner (27):

(30) ***done*** *at* ***Brussels*** *on 14 June 1983*
(31) *This Directive shall* ***enter*** *into force on the 20th* ***day*** *following that of its publication in the Official Journal of the European Communities*
(32) *In the text* ***published*** *in the* ***Official*** *Journal of the European Communities a material error occurred in the date for the bringing into force of the laws*
(33) Articolul 4 a fost **modificat** la **data** de 20 mai 1994
'the article 4 has been modified on the 20th May 1994'

(I) **Modifier (V) + Head (N)**. In this category, the V is a non-Finite Epithet or Classifier of the following N.

(34) *the* ***supporting documents*** *are in order…*
(35) *If the* ***measures envisaged*** *are not in accordance with the opinion of the committee the exercise of implementing powers conferred on the Commission*

(J) **Mistaken Identify.** This final category, which is still unfortunately very large, includes instances of collocation which do not even involve VN co-occurrence. Most of these are due to tagging errors (example 36), long-range collocations (as mentioned above, and in 37, 38), incorrect segmentation (missing punctuation – example 39), or conjunctions between N and V (example 40 etc.).

(36) *This Regulation shall be* ***binding*** *in its* ***entirety*** *and directly applicable in….*
(37) ***Having*** *regard to the opinions of the* ***European*** *Parliament*
(38) *Whereas the measures* ***provided*** *for in this Decision are in accordance with the* ***opinion*** *of the Standing Committee on Zootechnics*
(39) **Articolul** 2 Acest Regulament va **intra** în vigoare la trei zile după publicarea
'Article 2 This Decision will enter into force three days after its publishing'
(40) Prezentul regulament este obligatoriu în toate **elementele** sale şi **se aplică** direct în toate statele membre
'all the articles of this rule are mandatory and it is applied by all Member States'

Many of these invalid examples can be excluded by the simple expedient of ruling out 'long-range' VN co-occurrences, or by more rigorous application of orthographic boundaries. In general, the ‘noise’ which occurs especially in our latter categories shows that considerable ‘post-processing’ is still required in any semi-automatic approach to the identification of ‘valid’ VN constructions.

Conclusion

In this paper we have argued that the only common denominator in Verb-Noun constructions is that the Noun has the semantic role of ‘Process Range’, thus contributing fundamentally to the expression of Process in the Predicate. We have also argued that it is necessary to distinguish between cohesive collocations (= constructions), of the type set out in 3.2, such as *give rise to*, and statistically salient collocations (= co-occurrences), of the type set out in 3.3 such as *supporting documents*. The latter sometimes constitute valid constructions, but not of the type which we were hoping to emerge from the data analysis.

Our conclusion must be that VN constructions cannot be fished out of a corpus by simply looking for statistically significant co-occurrences of V plus N. The data we have previously examined for French (Gledhill 2007) suggest that constructions in which the N is a qualifier of the Predicator (*faire faillite* ‘to go bankrupt’, *faire l’objet de,* ‘to be subject to’, etc.) percolate to the top of the statistical list just as they appear to do here for Romanian. However, it turns out that in English these constructions are less statistically salient. One explanation for this may be that English just has less of these constructions. However, it may also be that our table of English VN co-occurrences (section 3.1) contains a large number of highly rigid, fixed expressions. Linguistically speaking, some of these correspond to valid constructions, but the majority appear to be ‘noise’. The lesson from this must be that relevant collocations do not always correspond to highly fixed sequences.

However there is a more positive outcome from our data analysis. Looking at the data, it has become clear to us that that the relevant context for any relevant VN construction extends beyond the basic collocation of V plus N. Our preliminary conclusion must therefore be that the contextual features of VN constructions that are to be found beyond the VN pairs themselves are crucial to the semi-automatic extraction of collocations.

Acknowledgements

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

References

ALLERTON D. (2002): *Stretched Verb Constructions in English,* London, Routledge.

BANKS D. (2000): "The Range of Range: A transitivity problem for systemic linguistics", in *Anglophonia*, 8, 195-206.

COTTE P. (1998): "*Have* n'est pas un verbe d'action : l'hypothèse de la réélaboration", in Rousseau, A. (ed): *La Transitivité,* Lille, Presses Universitaires du Septentrion, 415-439.

DI-SCULLO A-M & ROSEN S.T. (1991): "Constructions à prédicats légers et quasi-légers", in *Revue québécoise de linguistique*, 20:1,13-37.

FIRTH J.R. (1957): *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.

FRATH P. & GLEDHILL C. (2005): "Free-Range Clusters or Frozen Chunks? Reference as a Defining Criterion for Linguistic Units," in *Recherches anglaises et Nord-américaines,* vol. 38 :25-43.

GIRY-SCHNEIDER J. (1987): *Les prédicats nominaux en français*. *Les phrases simples à verbe support.* Genève-Paris, Droz.

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

GLEDHILL C. (2005): 'Problems of Adverbial Placement in Learner English and the British National Corpus.' In D.J. Allerton, C. Tschirhold, & J. Wieser, (eds.) *Linguistics, Language Learning and Language Teaching*. (ICSELL 10.) Basel, Schwabe, 85-104.

GLEDHILL, C. (2007): "La portée : seul dénominateur commun dans les constructions verbo-nominales", in Frath, P., Pauchard, J. & Gledhill, C. (éds) Actes du 1er colloque *Res per nomen,* Reims 24-36 mai 2007, Université de Reims, Champagne, 113-124.

GLEDHILL C. & FRATH P. (2007): "Collocation, phrasème, dénomination : vers une théorie de la créativité phraséologique", in *La Linguistique,* 43/1, 65-90.

GLEDHILL C. HEID U. MIHĂILĂ C. ROUSSELOT F. ŞTEFĂNESCU D. TODIRAŞCU A. TUFIŞ D. & WELLER M. (2007): *Collocations en contexte: extraction et analyse contrastive,* Project Report for the *Agence Universitaire pour la Francophonie* 'Réseau Lexicologie, Terminologie, Traduction', Paris :1-38.

GRIMSHAW, JANE & MESTER, ARMIN, (1988): "Light Verbs and θ-Marking", in *Linguistic Inquiry,* 19, 205-232.

GROSS, G. (1989): *Les constructions converses du français,* Genève-Paris, Droz.

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

HALLIDAY M.A.K. & MATTHIESSEN M. (2004): *An Introduction to Functional Grammar: 3rd Edition*, London, Arnold.

HAUSMANN F. J. (1989): Le dictionnaire de collocations. In Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L.(eds), *Wörterbücher : ein internationales Handbuch zur*

Lexicographie. Dictionaries. Dictionnaires. Berlin/New-York : De Gruyter. 1010-1019.

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

HEID U. & RITZ J. (2005): Extracting collocations and their contexts from corpora, Actes de COMPLEX-2005, Budapest.

HUNSTON S. & FRANCIS G. (2000): *Pattern Grammar- A Corpus-Driven Approach to the Lexical Grammar of English,* Amsterdam, John Benjamins.

ION R. (2006): *TTL: A portable framework for tokenization, tagging and lemmatization of large corpora*. Research Institute for Artificial Intelligence, Romanian Academy, Bucharest (in Romanian).

KERMES H. (2003): *Off-line (and On-line) Text Analysis for Computational Lexicography,* Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), vol. 9, no. 3.

KEARNS K. (1989): "Predicate Nominals in Complex Predicates", in *MIT Working Papers in Linguistics*, 10, 123-134.

KIM S-W. (1994): "A Study on the Light Verb Construction in English and Korean", in *Language Research,* 30:1, 197-159.

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

MOON R. (1998): *Fixed Expressions and Idioms, a Corpus-Based Approach.* Oxford, Oxford University Press.

POTTELBERGE J. VAN (2000): "Light Verb Constructions: What they Are and What they are Not", in *Logos and Language*, 1:2, 17-33.

RITZ J. & HEID U. (2006): "Extraction tools for collocations and their morphosyntactic specificities", in: *Proceedings of LREC-2006,* Genova, Italy.

ROUSSELOT F, & MONTESSUIT N. (2004): LIKES un environnement d'ingénierie linguistique et d'ingénierie des connaissances. *Workshop INTEX* Sofia.

SMADJA F. A & MCKEOWN, K. R. (1990): "Automatically extracting and representing collocations for language generation", *Proceedings of ACL'90*, 252-259, Pittsburgh, Pennsylvania.

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

SINCLAIR J. (1996): "The Search for Units of Meaning", in *Textus,* IX, 75-106.

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", in *Proceedings of the 5th LREC Conference*, pp.2142-2147.

STEFANESCU D, TUFIS, D, IRIMIA E. (2006): "Extragerea colocatiilor dintr-un text", in *Atelierul ' Resurse lingvistice si instrumente pentru prelucrarea limbii române',* Universitatea Al.I.Cuza Iasi, pp. 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.

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, Ann Arbor, Michigan, pp. 107-110.

VIVÈS R., (1984): "L'Aspect dans les constructions nominales prédicatives: avoir, prendre, verbe support et extension aspectuelle», in *Linguisticae Investigationes,* 3:1, 161-185.

WIERZBICKA A. (1982): "Why can you Have a Drink when you can't Have an Eat?", in *Language*, 58, 753-799.