Collocation Chains: How to Deal with Them?

Margarita Alonso Ramos(1), Leo Wanner (2)

(1) Universidade da Coruña
Campus da Zapateira s/n, 15071 Coruña
lxalonso@udc.es
(2) ICREA and Technology Department
Pompeu Fabra University, Passeig de Circumval·lació, 8 08003 Barcelona
leo.wanner@upf.edu

Abstract

Collocation chains are pairs of collocations in which one element participates in both collocations. Such chains have hardly been considered so far in MTT or in other theoretical frameworks. However, some of them raise interesting questions with respect to their representation in the dictionary and treatment in NLP – e.g., in automatic text generation. In this paper, we give an overview of the different types of collocation chains and provide a preliminary suggestion for their treatment. For illustration, we use examples from Spanish.

Keywords

Collocation, collocation chain, lexical function, Explanatory Combinatorial Dictionary, automatic text generation, semantics, syntax, lexicology

1 Introduction

Speaking about collocations, we usually think of two lexical units (LUs) between which a restricted lexical co-occurrence relation holds – as, for example [to] take and walk in John took a walk along the Hudson River. However, in real corpora, we quite often come across what we call collocation chains: a sequence of several collocations where one element is shared by more than one collocation. Consider, for instance, [to] commit a violation of the law. The noun violation is the base of the collocation commit [a] violation; at the same time, it is the collocate of the collocation violation [of the] law. Not all collocation chains are of the same kind. Some of them (as, e.g., [to] take drastic measures) can be treated as sequences of isolated collocations and thus do not need any new consideration with respect to their encoding in the dictionary and representation / mapping in a synthesis-oriented model (such as the Meaning-Text Model, MTM). Others require an extension of the standard collocation representation as well as of synthesis rules; cf., e.g., the chain [to] follow the rules rigidly, in
which the adverb *rigidly* modifies the verbal collocation *follow the rules* as a whole (rather than *follow* only).

In this paper, we take a closer look at chains which consist of two collocations (henceforth, *double collocation chains*), focusing on Spanish material and giving occasional evidence from German and English. To the best of our knowledge, collocation chains have not been studied in MTT yet, and only rather superficially outside MTT; cf., e.g., (Koike, 2001:147–149, 2004; Muñiz, 2004). Our goal is threefold. First, to provide a sketch of a collocation chain typology. Second, to determine how collocation chains are to be represented in a dictionary in the tradition of the Explanatory and Combinatorial Lexicology (Mel’čuk et al., 1995), using *lexical functions*, LFs (Wanner, 1996). Third, to propose a way to represent the different kinds of collocation chains in the semantic (Sem) and deep-syntactic (DSynt) structures of an MTM and to encode the Sem-DSynt transition rules in a generation framework.

### 2 Towards a Typology of Collocation Chains

From a somewhat more formal view, a double collocation chain is a sequence of three elements $E_1$, $E_2$, $E_3$, with collocation relations defined either between $E_1$ and $E_2$ and $E_3$ individually: $E_1$–$E_2$, $E_1$–$E_3$, $E_2$–$E_3$, or between $E_1$ and a complex unit consisting of $E_2$ and $E_3$: $E_1$–[$E_2$–$E_3$]. An element can be a base (B), a collocate (C), or both (|C/B|). The following five basic cases are theoretically possible:

1. **CCB**: two collocates share the same base (as, e.g., $\text{tomar}_C$ $\text{medidas}_B$ $\text{drásticas}_C$ ‘to take drastic measures’);
2. **CBB**: two bases share the same collocate (as, e.g., $\text{las abejas}_B$ $\text{liban}_C$ $\text{el néctar}_B$ ‘the bees suck the nectar’);
3. **C |B/C| B**: one element is simultaneously collocate and base of two different collocations (as, e.g., $\text{cometer}_C$ $\text{una violación}_{B/C}$ $\text{de la ley}_B$ ‘to commit a violation of the law’);
4. **C[B/C]**: a collocation is a base of another collocation (as, e.g., [llamar$_C$ la atención$_B$] $\text{poderosamente}_C$ ‘to draw the attention intensely’)
5. **B[CB]**: a collocation is a collocate of another collocation (as, e.g., [coger$_C$ el volante$_B$] $\text{del coche}_B$ ‘to take the wheel of the car’).

Obviously a first rough distinction can be made between 1–3 and 4–5: in 1–3, the chain elements are individual lexical units; in 4–5, one of the elements of a chain is itself a collocation. In what follows, we call this element *complex element*. Each of the cases in 1–3 potentially contains several subcases, which vary with respect to the syntactic dependencies between the elements. In order to assess which cases are problematic, we need to examine these dependencies more closely.

---

1 The order in which the elements are given does not reflect any dependencies or default order within sentences. To make the role of each element in the examples explicit, we annotate it with the corresponding subscript.

2 In this example (and in all similar examples, the complex unit (*coger el volante*) is a quasi-synonym of *conducir* ‘drive’, i.e., it is the value of the LF Real$_1$(*coche*).
2.1 Syntactic Dependency between Individual Elements of Collocations

Roughly speaking, we can distinguish configurations in which a dependency relation holds between collocate(s) and base(s) and those in which it does not.

2.1.1 Syntactically Related Base and Collocate

This is the most frequent case for the patterns CCB and CBB. For CCB, the possible syntactic configurations are $C_1 \rightarrow B \rightarrow C_2$ and $C_1 \leftarrow B \rightarrow C_2$. The elements can be of different parts of speech; cf., e.g.,

Base: N; Collocates: Verb, Adj, as in *tomar medidas drásticas* ‘take drastic measures’; *einen ausgedehnten Spaziergang machen* lit. ‘take a stretched-out walk’

Base: N; Collocates: Verb, Verb, as in *callar el miedo sentido* lit. ‘make silent the felt fear’

Base: N; Collocates: Adj, Adj, as in *medidas drásticas pero justas* ‘drastic but fair measures’

For the CBB pattern, we find among possible syntactic configurations $B_1 \leftarrow C \rightarrow B_2$, which is exemplified, e.g., by *las abejas liban el néctar; el conductor conduce el coche* ‘the driver drives the car’, etc., where ABEJAS and CONDUCTOR are $B_1$, LIBAR and CONDUCIR are $C$ and NÉCTAR and COCHE are $B_2$.

2.1.2 Syntactically Unrelated Collocate and Base

The cases with no syntactic relation between the base and the collocate are less frequent, but, as we will see later, much more interesting from the viewpoint of their treatment. In what follows, we restrict ourselves to the analysis of two configurations of the pattern CCB: $C_1 \rightarrow C_2 \rightarrow B$, $C_1 \leftarrow C_2 \rightarrow B$, and one configuration of pattern $C \mid B/C \mid B$: $C_1 \rightarrow B_1/C_2 \rightarrow B_2$ with these characteristics.

The configuration $C_1 \rightarrow C_2 \rightarrow B$ is instantiated by, e.g., *sufrir [una] oleada de atentados* ‘[to] suffer a spate of terrorist attacks’, *[el] estado de alarma cunde* ‘[the] state of alarm spreads’, as well as by *tener una sensación de alivio* ‘[to] have a feeling of relief’ and *tener un ataque de sueño* ‘[to] have a fit of sleep’ (with SUFRIR, CUNDIR, and TENER as $C_1$, OLEADA, ESTADO, SENSACIÓN and ATaque as $C_2$, and ATENTADO, ALARMA, ALIVIO and SUEÑO as $B$). Note that SUFRIR and OLEADA and CUNDIR and ESTADO, respectively, do not form a collocation. The case of *tener una sensación de alivio* and *tener un ataque de sueño* is somewhat different. TENER and SENSACIÓN, and TENER and ATaque, respectively, form a collocation. That is, this case can be also interpreted as the $C \mid B/C \mid B$-pattern (see below).

The nouns OLEADA, ATaque and ESTADO have the semantic feature called by Apresjan et al. (1989:264) crystal. They are also known as light (Bosque, 2001; Koike, 2003) or transparent nouns (Fillmore et al. 2002, 2003). As Fillmore et al. (2003:244) put it: “a transparent noun is one which can appear as the first noun in N1-of-N2 constructions in contexts where the governing verb actually selects N2 rather than N1, the syntactic head”. They suggest that transparent nouns can denote types, parts, portions, aggregates, borders, classifiers, and quantifiers. In the LF-terminology, the transparent noun $C_2$ in a $C_1 \rightarrow C_2 \rightarrow B$ dependency chain is most often Mult($B$), Sing($B$), Figur($B$) or Gener($B$) – if the generic noun is abstract enough, as, e.g., ESTADO ‘state’. Henceforth, we refer to these LFs as transparent LFs.
The configuration $C_1 \leftarrow C_2 \rightarrow B$ as, e.g. *hacer solemnemente una promesa* ‘[to] solemnly make a promise’ is another example in which $C_1$ and $B$ are syntactically not directly related. It is thus to be handled along the same lines as $C_1 \rightarrow C_2 \rightarrow B$.

The $C_1 \rightarrow B_1 / C_2 \rightarrow B_2$ pattern (called *chained collocation*) as manifested in *cometer una violación de la ley*, *die Urheberschaft einer Tat leugnen* ‘deny the authorship of an act’, *die Lösung eines Problems finden* ‘find the solution of a problem’, etc. is very productive (and thus very common). Since $B_1$ (VIOLACIÓN, URHEBERSCHAFT, etc.) functions at the same time as collocate $C_2$, we have to label it in the DSyntS as an LF (rather than as a regular LU).

### 2.2 Syntactic Dependency between Complex and Individual Elements of a Collocation

The last two patterns $C[CB]$ and $B[CB]$ in our list of collocation chain types share the property that one of their elements is complex, i.e., a collocation. In the pattern $C[CB]$, the complex element is the base, and in the pattern $B[CB]$, the complex element is the collocate. The pattern $C[CB]$ as illustrated above by *llamar la atención poderosamente* is rather rare; it corresponds to the phenomenon Muñiz (2004:24) calls integration of collocations. Two syntactic configurations can be distinguished: $C_1 \rightarrow [C_2B]$ and $C_1 \leftarrow [C_2B]$. The above example illustrates the pattern $C_1 \leftarrow [C_2B]$ (with *llamar la atención* as $[C_2B]$ and PODEROSAMENTE as $C_1$). As *llamar poderosamente* and *atención poderosamente* show, $C_1$ indeed refers to $C_2B$ as a combination rather than to $C_2$. Further examples for this pattern are: *izar la bandera a media asta* ‘[to] raise the flag at half-mast’ and *abrir la puerta de par en par* ‘[to] open wide the door’.

The representation of the pattern $B[CB]$ does not require any extraordinary strategy. Again, there are two dependency configurations: $B_1 \rightarrow [CB_2]$ and $B_1 \leftarrow [CB_2]$; let us restrict ourselves to the second of them. $B_1 \leftarrow [CB_2]$ is illustrated by *[coger el volante$_B$] del coche$_B_1$, [darc riendas$_{B_2}$ suelta] al llanto$_B_1* ‘to give free rein to crying’, *[tomar las riendas$_{B_2}$] del negocio$_B_1 lit.* ‘to take the reins of the business’, etc. This configuration is more productive for symptomatic expressions of emotions, such as *quedarse sin habla de miedo* lit. ‘stay speechless of fear’, *los dientes castañetean de miedo* lit. ‘the teeth are chattering of fear’, *quedarse helado del susto* lit. ‘stay frozen of the shock’.

### 3 Representation of Collocation Chains in the Dictionary

We presuppose that the reader is familiar with the notion of LF and the representation of collocations in terms of LFs in the dictionary. The different types of collocation chains call for the distinction of three approaches to their representation: (a) the standard way used in ECL-dictionaries; (b) in terms of separate lexical entries for collocates, (c) in terms of embedded lexical zones for collocates in the entries for the bases.

For the patterns CCB, BBC and B[CB], the lexicographic description is standard. Each collocation is described in isolation in terms of LFs in the entry for its base (= keyword of the LF), no matter whether the value is an individual lexeme, an idiom, a collocation or a free phrase. For extensive illustrative material, see the ECL-dictionaries.
Collocation chains

In the case of the pattern \( C_1|B_1/C_2 |B_2 \), the description of \( B_1/ C_2 | B_2 \) (in our example, violación de la ley) is standard. The description of \( C_1 B_1/ C_2 \) (cometer violación) depends on whether \( B_1/C_2 \) receives its own lexical entry or not. Until recently, in ECL a collocate did not systematically receive an entry. The decision on defining a lexical entry for a collocate depends on several factors (Alonso Ramos, 2003). One of the factors is the capacity of the LU in question to select its own collocates. This is the case for the noun VIOLACIÓN. Apart from COMETER, it co-occurs, e.g., with GRAVE ‘gross’, SERIA ‘serious’, etc. – also when VIOLACIÓN serves as collocate for such bases as NORMA ‘norm’, REGLA ‘rule’, etc. Therefore, there is no better place than its lexical entry to state that it selects COMETER as a value of the LF Oper_1, GRAVE as a value of the LF Magn, etc.

The case of the pattern \( C_2[C_1 B] \) is similar to the previous one. It is to be decided whether an entry for \( C_1 \) (llamar in our example) is to be introduced or its collocate (PODEROSAMENTE) should be described in the entry for the base \( B \) (ATENCIÓN). Thus, we could introduce an entry for LLAMAR with the particular meaning it has in co-occurrence with ATENCIÓN. In this entry, the adverb Poderosamente would be the value of Magn. However, if Poderosamente can modify LLAMAR only in combination with ATENCIÓN, it seems more appropriate to describe the relation between the adverb and the collocation in the entry for the noun. Thus, for the entry ATENCIÓN, we would open a lexical zone for the collocation in such way that we can assign information to some of the values of the LF, but not to all.

\[
\text{Caus Func:}
\begin{align*}
\text{llamar [ART} & \sim a X] \\
\text{Magn = poderosamente} \\
\text{suscitar [ART} & \sim a X]
\end{align*}
\]

4 Collocation Chains in Generation

In this section, we take a more systematic look at the encoding of collocation chains at the Sem- and DSynt-levels and their treatment during the transition between these two levels.

4.1 Representation of Collocation Chains

We present the SemSs and DSyntSs of collocation chains from some of the examples cited above. The examples show that SemS of the chains we look at in the paper is fairly simple; it is the DSyntS (and thus also the Sem-DSynt transition) that deserves some attention.

1. tomar medidas drásticas
   ‘take drastic measures’

2. sufrir una oleada de atentados
   ‘suffer a spate of attacks’
3. cometer una violación de la ley  
‘commit a violation of the law’

4. llamar atención poderosamente  
‘call powerfuly for attention’

These structures call for the following remarks:

(i) In accordance with Wanner & Alonso (2005), we interpret LF-labels at the DSynt-level as deep LUs, i.e., we do not use in the structures the functional notation f(L). This does not mean that the information concerning the functional interpretation of an LF is lost. Each syntagmatic LF receives an entry where we specify its government pattern (GP), with the GP allowing for the recovery of the keyword.

(ii) The case illustrated in (1) is simple since the representation of the chain at the DSynt-level is equivalent to the representation of two LFs in isolation.

(iii) At the DSynt-level, where we consider LFs as deep LUs, cases (2) and (3) are analogous: in both cases, an LF has as its DSyntA another LF. The difference lies in their functional interpretation. The standard Oper-representation in DSyntSs requires the keyword to be actant II of the Oper₁-node. However, in (2), actant II is the Mult-node, which is not the keyword of Oper₂: the keyword is ATENTADO. In (3), although actant II of the Oper-node is also an LF, the LF-node acts as its keyword.

(iv) Case (3) is not trivial because it contains an LF-label (namely, S₀AntiReal₃) as keyword of the Oper-node. This means that we need to know the LU that functions as value of S₀AntiReal₃ in order to ensure that this LU possesses an Oper.

(v) (4) is the same as (3) with respect to the use of an LF-label (CausFunc₁) as keyword of another LF-labelled node (Magn). However, (4) requires a different description in the dictionary and thus a different treatment during the DSynt-SSynt transition.

4.2 Transition between Sem- and DSynt-Levels of Representation

The Sem-DSynt transition between the structures of the type shown in (1) has already been discussed, e.g., in (Iordanskaja et al., 1996; Kahane & Mel’čuk, 1999; Alonso Ramos, 2007). Therefore, we focus on the transitions between structures illustrated in (2)–(4). Each transition rule is elementary in the sense that it covers the smallest fragment possible of a SemS that is being “transduced” into its DSynt-equivalent: single nodes and binary relations between nodes. As a result, each transition grammar consists of a set of nodal and sagittal rules (Kahane & Mel’čuk, 1999), with any of the rules being applicable to a multitude of structures. Let us introduce the types of rules shared by all our cases.
4.2.1 Regular Nodal and Sagittal Rules

Two types of regular nodal rules are pertinent for our task: lexical-semantic rules and lexical-functional rules. Furthermore, we need a number of sagittal rules of which we illustrate one. The lexical semantic rule cited below maps a communicatively dominant predicative semanteme (denoted here by the variable \( ?X_s \)) onto the corresponding deep LU that is realized as a noun in an LF-constructin (assigned to the variable \( ?X_{ds} \)). We assume that the choice of an Oper-construction is motivated essentially by the communicative structure, CommS, captured in the rule in question by the feature ‘aspectual focus-event’.  

\[
?X_s \leftrightarrow ?X_{ds} \\
\text{lex} = \text{nominal LU of 's'} \\
\text{?Xs has the comm. feature 'aspectual focus'}
\]

The following rule illustrates a “regular” lexical-functional rule which maps the semanteme ‘serie’ ‘series’ onto the LF-label Mult. The keyword of Mult is the lexeme corresponding to the first argument of ‘serie’ (?Xs−1→?Ys). Both the lexeme and the relation are introduced by other elementary rules. Therefore, both appear in the context (shown in grey). Obviously, the condition to be fulfilled is that this LU has a value of Mult.

\[
?X_s \leftrightarrow ?X_{ds} \\
\text{The lexical equivalent of } ?Ys.sem \text{ has a Mult-LF}
\]

The sagittal rule below maps the semantic relation ‘1’ onto the DSynt-relation ‘I’. The relation ‘1’ holds between a quantifier semanteme (such as ‘series’) and its argument. The relation ‘I’ holds between the LF-label of a transparent LFs such as Mult, Sing, etc. and the corresponding keyword.

\[
?X_s \leftrightarrow ?X_{ds} \\
\text{?Xds is the transparent Mult node introduced before} \\
\text{?Xds has the feature "keyword", which is instantiated with the LU assigned to ?Yds}
\]

In order to carry out the mapping between the SemSs and DSyntSs in (2)–(4), Section 4.1, we need, in addition to the regular rules sketched above, some idiosyncratic rules that take care of collocation chains.

---

3 This feature is intended to capture the different communicative aspectual perspectives on a semantic configuration. For instance, a support verb construction such as take a walk expresses a bound event, while the verb to walk expresses a process.

4 To facilitate the readability, we have chosen a very specific rule. In practice, this rule is substituted by a more generic rule that handles the mapping of semantic configurations onto LF-labels.
4.2.2 The Case of Transparent LFs

This case corresponds to the mapping of the structures in (2). To realize the collocation chain, we need a special Oper2-rule. Any Oper-rule needs to consider the CommS. However, in this case, it must also ensure the availability of a transparent LF already introduced into the DSynS. The rule specifies that the second actant of the newly introduced Oper2-node is a transparent LF and both LFs share the same keyword (?Zds.lex). ?Xs→?Ys is the relation between the main node and its second actant.

• ?Zds is the keyword introduced before and carries an Oper2-label
• ?Xs is the communicatively dominant node
• ?Xs has the communicative feature ‘aspectual focus_event’
• ?Yds is the transparent Mult node introduced before
• ?Yds has a feature ‘keyword’, which is instantiated by ?Zds

The left-hand side of this rule consists of a context only since ?Xs, ?Ys and the relation between them are mapped by other rules.

4.2.3 The Case of Chained LFs

In the case of chained LFs of the type cometer violación de la ley (cf. (3) in Section 4.1), two rules similar to those already introduced above are essential: one for the introduction of the S0AntiReal3-node and one for the introduction of the Oper1-node. A preliminary S0AntiReal3-rule looks as follows:

The semanteme ‘act inappropriately’ corresponds at the DSyntS-side to AntiReal3. The nominalization enforced by the communicative features is encoded by S0. The Oper1-rule is somewhat different from the Oper-rule in Section 4.2.2:

5 The rule is fairly idiosyncratic and implies that each verbal LF-rule must be duplicated in order to account for nominalization of the value of the verbal LF. A dynamic composition of LF-labels (in the spirit of Kahane & Polguère, 2001) during the transition would be more appropriate. However, this strategy has the disadvantage that it implies a modification of already introduced DSynt-nodes.
Collocation chains

The conditions require in the dictionary an Oper1-value (here, COMETER) for the value of the keyword-LF assigned to ?Yds (i.e., VIOLACIÓN).

4.2.4 The Case of Complex Elements

Chains containing complex elements of the kind *llamar la atención poderosamente* can be dealt with during the Sem-DSynt transition along the same lines as discussed in 4.2.3. However, they call for attention during the DSynt-SSynt transition because the collocate (such as PODEROSAMENTE) does not co-occur with all values of the LF applied to ATENCIÓN (in our case, Caus2Func1). In order to ensure the realization of Magn, we have to select a Caus2Func1-value which has a Magn-collocate. This is the price that we must pay to account for having a collocation as a base. The following rule handles the Caus2Func1-lexicalization:

\[
\begin{align*}
\text{ATTR} & \quad \Downarrow \quad \text{?Xds} \quad \Leftrightarrow \quad \text{?Xss} \\
\text{lex} &= \text{<value of the LF assigned to ?Xds>} \\
\text{?Xds carries an LF-label} \\
\text{?Zds carries an LF-label} \\
\text{?Xss has the LF carried by ?Zds}
\end{align*}
\]

The rule for the lexicalization of Magn ensures the choice of a value that co-occurs with the already determined surface lexeme of Caus2Func1. We omit the discussion of the Magn-rule.

5 Summary and Conclusions

We presented some examples of dual collocation chains, showing that they cannot always be dealt with as isolated collocations. Dependency relations between the bases and collocates play a particular role in the complexity of a collocation chain. Furthermore, the double role of LFs, as deep LUs and as functions, which comes to bear in collocation chains raises again the question on the nature of LFs in an MTM and thus, subsequently, the nature of collocates as proper LUs. Finally, despite the preliminary nature of our SemDSynt-transition rules and of the discussion of the representation of collocation chains in the dictionary, we hope to have shown that they need to be taken into account in a comprehensive synthesis model.

Acknowledgements

Many thanks to I. Mel’čuk, J. Milićević and the three anonymous reviewers whose valuable comments significantly improved the final version of this paper. As always, the remaining errors are ours. The work on the paper was supported by a research grant (HUM2005-08052-C02-02, Feder) and travel grants from the Spanish Ministry for Science and Education (MEC).

Bibliography


