Qf Lingüístics

A quantitative survey of N Prep N constructions in Romance languages and prepositional variability

Un estudio cuantitativo de las construcciones N Prep N en las lenguas románicas y variabilidad preposicional

Inga Hennecke[a] & Harald Baayen[b]

[a] Universität Tübingen. inga.hennecke@uni-tuebingen.de
[b] Universität Tübingen. harald.baayen@uni-tuebingen.de
Received: 24/04/2017. Accepted: 10/10/2017

Abstract: The distinction between syntagmatic compounds of the type N Prep N, such as Fr. *jouet d'enfant*, and nominal syntagms of the type N Prep N, such as the partially equivalent Fr. *jouet pour enfants*, remains unclear and vague. This is mainly because the lexical and syntactic status of syntagmatic compounds still is controversial. In some cases, as in *jouet d'enfant* and *jouet pour enfants*, partial equivalent syntagmatic compounds and nominal syntagms may coexist and underlie a specific variation and alternation. In other cases, such as Pt. *bracelete de aço* and *bracelete em aço*, two variants of a syntagmatic compound may alternate and coexist.
The first part of this paper provides an overview of the current discussion on these two types of constructions. The second part addresses the alternation and variation of syntagmatic compounds and nominal syntagms by means of analysis of large-scale corpus data, the French, Spanish and Portuguese corpus of the TenTen family. Here, the focus lies on the variation of the prepositional internal element of these constructions as well as on a comparison of different word formation patterns.

Keywords: Compounds; quantitative corpus linguistics; lexicon-syntax interface; Romance.

Resumen: La distinción entre los compuestos sintagmáticos del tipo N Prep N, como por ejemplo Fr. *jouet d'enfant*, y los sintagmas nominales del tipo N Prep N, como Fr. *jouet pour enfants*, sigue siendo confusa. Esto se debe, sobre todo, a que no existe consenso a propósito de la categorización léxica y sintáctica de los compuestos sintagmáticos. En algunos casos, como en *jouet d'enfant* y *jouet pour enfants*, se trata de equivalentes parciales que pueden coexistir y estar sujetos a una variación y alternancia

» Hennecke, Inga & Baayen, Harald. 2017. "A quantitative survey of N Prep N constructions in Romance languages and prepositional variability". *Quaderns de Filologia: Estudis Lingüístics* 22: 129-146. doi: 10.7203/qf.22.11305

específica. En otros, como en Pt. *bracelete de aço* y *bracelete em aço*, las posibles variaciones pueden alternar y coexistir en prácticamente todos los contextos.
La primera parte de esta contribución ofrece un breve resumen de la discusión reciente sobre estos dos tipos de construcciones. La segunda sección discute la alternancia y variación de los compuestos sintagmáticos y los sintagmas nominales mediante el análisis de diferentes corpus de gran tamaño: el corpus español, francés y portugués de los corpus TenTen. El análisis se centra especialmente en la variación del elemento preposicional interno de los compuestos y los sintagmas, y en la comparación entre los diferentes tipos de formación de palabras que tienen lugar en ellos.

Palabras clave: palabras compuestas; lingüística de corpus cuantitativa; interfaz léxico-sintaxis; lenguas románicas.

1. State of the Art

Terminological insecurity and inconsistent classifications dominate the scientific debate on syntagmatic compounds of the type N Prep N in Romance languages. Currently, possible denominations include terms such as phrasal compounds (Bisetto & Scalise, 2005), syntactic compounds (Rio-Torto & Ribeiro, 2009), improper compounds (Kornfeld, 2009), phrasal lexemes (Masini, 2007, 2009; Masini & Scalise, 2012), "frozen" multiword units (Guevara, 2012), lexicalized syntactic constructions (Villoing, 2012), lexicalized phrases (Fradin, 2009), syntactic words (DiSciullo & Williams, 1987) or even syntactic syntagms or prepositional syntagms. The heterogeneous terminology goes along with a diverse delimitation and integration of different types of lexical and syntagmatic units. In the same way, syntagmatic compounds of the type N Prep N may or may not – depending on the underlying terminology – be included in the group of compounds.

Moyna (2011) includes in her definition of syntagmatic compounds different combinations of substantives and adjectives, which may or may not show orthographic union:

[N PREP N]N	*dulce de leche*,	"caramel"
[N PREP Art N]N	*árbol de la cera*	"wax myrtle"
[N + A]N	*hierbabuena*	"mint"
[A + N]N	*malasombra*	"evil person"

(Moyna 2011: 38)

In contrast, Masini (2009) does not include orthographically unified combinations, such as *hierbabuena*, but she adds constructions of the type N Prep V_{INF}, such as *salle à manger* 'dining room'.

Traditional grammars and dictionaries generally classify nominal syntagmatic compounds of the type Sp. *bicicleta de montaña* 'mountain bike'*,* Fr. *brosse à dents* 'tooth brush' or Pt. *moinho de vento* 'windmill' as lexical units and therefore as compounds. But Kabatek & Pusch (2009) indicate that it is not always clear how to differentiate between lexical items of the type *perro de caza* and more syntactic items such as *libro para niños* (Kabatek & Pusch, 2009: 93f.). According to de Bustos Gisbert, syntagmatic compounds consist of at least two etymological words and are formally not distinguishable from nominal phrases (de Bustos Gisbert, 1986: 69). In the same line of argumentation, Masini

notes that syntagmatic compounds of the type N Prep N follow the normal syntactic patterns of head modification of the nominal phrase by the prepositional phrase (2009: 257). N Prep N constructions in Romance languages therefore tend to be left-headed and inflectional processes are performed at the head constituent (ibd.).

According to Val Àlvaro (1999), the main distinctive feature between syntagmatic compounds and free nominal syntagms is the absence of a compositional meaning in syntagmatic compounds (Val Àlvaro, 1999: 4827). Therefore, they can be interpreted as complex nominals and not as nominal phrases. In the same line of argumentation, Štekauer (2001b: 39) classifies 'syntax-based word formations' such as *son-in-law* or *stuff-leaver* as onomasiological naming units that dispose of an internal structure and resort to the same word formation processes as other naming units. Furthermore, syntagmatic compounds generally differ from nominal syntagms in that they form an accentual unit (de Bustos Gisbert, 1986).

Still, a main concern of past research on syntagmatic compounds was their delimitation, especially by introducing new delimitation tests (e.g. Bouvier, 2000; Buenafuentes de la Mata, 2006; Bisetto & Scalise, 2005; Lieber & Scalise, 2007; Masini, 2009; Masini & Scalise, 2012). These tests generally include criteria such as the modification of the constituents (e.g. modification of the constituent order, insertion or omission of elements) via topicalization, intensification or the insertion of modifying adjectives. For Portuguese, the last two tests can be exemplified by Rio-Torto and Ribeiro (2012: 125):

moinho de vento	"windmill"
*moinho *antigo de vento*	"*wind old mill"
*moinho de *muito vento*	"*wind much mill"

These delimitation tests are of major importance for studies taking a lexicological, semantic and morphological perspective. These studies generally follow Benveniste (1966) in his statement that syntagmatic compounds are the real word formation process in French. In this perspective, syntagmatic compounds are commonly perceived as lexical structures that may show signs of internal syntactic patterns (Z.B. Bisetto & Scalise, 1999, 2005; Rio-Torto & Ribeiro, 2012). In contrast, studies that focus on syntax, such as Kornfeld (2003) or Lieber (1992),

generally perceive syntagmatic compounding as a clearly syntactic process. Other studies again do not focus on the delimitation of lexicon and syntax. From a construction grammar, respectively a construction morphology perspective, syntagmatic compounds and (partially) equivalent nominal syntagms are both considered as constructions, lying on a continuum between lexicon and morphosyntax (e.g. Masini 2009). Still, these studies also target a description and classification of different constructions, such as syntagmatic compounds, phrases and other types of compounds (Masini 2009). In the present account, we argue that there is no clear line between syntagmatic compounds and syntactic constructions, but that they lie on a continuum between a lexicalized and syntactic pole.

A second major concern in research on syntagmatic compounds is the question of whether these constructions are lexicalized syntactic constructions or whether they emerge by productive word formation patterns. Rainer (2016) clearly opts for the classification of syntagmatic compounds as productive lexical patterns:

> Formations of this kind [syntagmatic compounds] are not, as often stated erroneously, the result of the lexicalization of regular syntactic sequences, but constitute very productive lexical patterns (…) (Rainer 2016: 2624).

In contrast, Guevara (2012) excludes syntagmatic compounds of the type *fin de semana* 'weekend' from its description of Spanish compounds, along with cases such as *sabelotodo* 'know-it-all'. He explains his decision in that "they are clearly not formed by any rule of the language, they are "frozen" multiword units arising as the result of processes of lexicalization and fossilization and do not belong in the core of word-formation" (Guevara, 2012: 179). In a similar argumentation, Villoing excludes "lexicalized syntactic constructions that behave like lexical units" (Villoing, 2012: 35) such as *fil de fer* 'wire'*, brosse à dents* 'toothbrush' but also *sous verre* 'coaster'*, sans-papier* 'illegal immigrant' and *boit-sans-soif* 'boozehound' from his delimitation of compounds. By contrast, in the same volume on Romance compounds, Rio-Torto & Ribeiro (2012) propose a classification of phrasal compounds, such as *caminho de ferro* 'railway' in Portuguese, which are classified as involving "word sequences whose internal structure obeys the syntax rules typical of phrases" (Rio-Torto & Ribeiro, 2012: 7).

This short introduction to the current discussion demonstrates strikingly the terminological insecurity as well as the problematic delimitation and classification of syntagmatic compounds (for an overview see e.g. Bisetto & Scalise, 2005; Lieber & Scalise, 2007). The most prominent problem in this debate is by far the question of whether syntagmatic compounds should be considered as a part of the lexicon or a part of syntax. Furthermore, in most of the cases, the discussion comes down to the crucial question of whether syntagmatic compounding is a process of lexicalization or a process of productive word formation. In the present paper, we assume that syntagmatic compounding is a productive and rule-governed process of word formation in Romance languages. Furthermore, we assume that there is no clear boundary between lexicalized and syntactic constructions of the type N Prep N.

The aim of the present work is to have a closer look at syntagmatic compounding of the type N Prep N in corpora of written French, Spanish, and Portuguese, focusing on the internal variation of N Prep N constructions as well as on their frequency and productivity and potential differences across these three languages.

2. Internal alternation and variation in syntagmatic compounds

The above review of the theoretical status of syntagmatic compounds in Romance languages does not present a unified perspective. Nevertheless, syntagmatic compounds appear to be at least partially lexicalized constructions. The degree of their lexicalization may vary along with other factors such as semantic opacity/idiomaticity, entrenchment, fixedness of the internal constituents, frequency of occurrence, productivity etc. Despite their more or less strong degree of lexicalization, syntagmatic compounds still appear to preserve at least some of their syntactic characteristics. The at least partially syntactic character of syntagmatic compounds is apparent from the internal lexical and inflectional variation of these constructions. Rio-Torto and Ribeira (2012) consider the possibility of internal change in N Prep N – constructions as a test of compound status. From this perspective, examples of constructions in which the preposition can be replaced without changing meaning would imply the construction to be syntactic rather than lexical. Thus, the pair Pt. *forno a microondas* and *forno de microondas* ‘microwave oven’, where no clear semantic difference is discernable, would sug-

gest we are dealing with a syntactic construction, but conversely the French pair *flûte de champagne* 'glass of champagne' and *flûte à champagne* 'champagne glass', where there is a change of meaning, would indicate word formation is at issue. However, the phenomenon of internal prepositional alternation appears to be more complex than this. Internal alternation of the preposition appears to be not uncommon in Romance languages. The possibility of alternation depends to a large extent on factors such as the semantic function of the N2 as well as on the fixedness and idiomaticity of the whole construction. Consider the following examples:

1a. Sp. *esmalte de uñas – esmalte para uñas* (Pacagnini 2003)
"nail polish" "polish for nails"
b. Sp. *água de lavagem – água para lavagem* (ptTenTen)
"wash water" "water for washing"
c. Fr. *jouet d'enfant – jouet pour enfants* (frTenTen)
"toy" "toy for kids"

2a. Sp. *motor(es) de gasolina – motores a gasolina* (esTenTen)
"gas engine"
b. Fr. *épingle de nourrice – épingle à nourrice*
"safety pin"
c. Pt. *Fogão de lenha – Fogão a lenha* (ptTenTen)
"wood stove"

3a. Fr. *chemise de coton – chemise en coton* (frTenTen)
"cotton shirt" "shirt of cotton"
b. Pt. *bracelete de aço – bracelete em aço* (ptTenTen)
"steel bracelet" "bracelet of steel"
c. Sp. *ciclismo de pista – ciclismo en pista* (esTenTen)
"track cycling" "cycling on track"

In example 1, we see internal variation of the linking preposition *de/para* and *de/pour*. While the constructions containing *de* are clearly lexicalized, the combinations containing *para/pour* count as syntactic constructions. The use of *pour/para* intensifies the semantic relation of the two nominal items in the constructions, in this case 'function' (see Kornfeld 2009: 442 ff.). In 1a. and 1b., the N2 designates the object (1a.) or the process (1b.) of use of the N1, whereas in 1c. the user of N1 is specified.

Example 2 illustrates the alternation between the prepositions *de* and à (a). Here, both variants have lexical status that does not trigger a change from lexical to syntactic status. The same applies to the examples in 3, where we cannot identify a change in the lexical status, but clearly a certain discrepancy in the degree of lexicalization and the semantic relation between N1 and N2. That is to say that the constructions as shown in example 1.-3. are only considered partial equivalents, as they may also differ from each other in their actual usage frequency, their productivity and their opacity.

Some authors, such as Kampers-Manhé (2001), argue that the internal preposition has purely connecting properties ("opérateurs de couplage") (Kampers-Manhé 2001: 107) and "ne sont pas porteuses de sens" (ibd.). The above examples suggest that the preposition is not semantically completely inert, even though, as we shall see below, some noun pairs show considerable variation with respect to the choice of the internal preposition. Furthermore, the possibility of internal variation in the above examples indicates that these constructions may not be completely lexicalized. They still allow internal modification that appears to be syntactically motivated.

The following quantitative corpus survey aims to give further evidence for the productivity and frequency of the internal prepositional variation in syntagmatic compounds in Romance languages.

3. Corpus survey

3.1. *Data*

The present corpus linguistic investigation is based on three web corpora from the TenTen corpus family from Sketchengine[1], more precisely on the corpora frTenTen12 (French), esTenTen11 (Spanish) and ptTenTen11 (Portuguese). Their type counts range from 4 to 10 billion and their token count ranges from 5 to 11 billion (see General Corpus Information on sketchengine.co.uk):

[1] <https://www.sketchengine.co.uk>.

	frTenTen	*esTenTen*	*ptTenTen*
Tokens	11,444,973,582	10,994,616,207	4,626,584,246
Words	9,889,689,889	9,497,402,122	3,900,501,097
Sentences	456,065,104	407,205,587	190,221,913
Paragraphs	188,079,362	213,364,685	91,248,976
Documents	20,400,411	22,287,566	10,216,060

Table 1. Corpus Info of the TenTen corpora for French, Spanish and Portuguese (https://the.sketchengine.co.uk)

The corpora ptTenTen and esTenTen can furthermore be divided into an American and a European part, whereby the majority of the data represent American varieties of Spanish (79% of the esTenTen data) and Portuguese (76% of the ptTenTen data). We made use of normalized samples of 100 million tokens each, provided to us by Sketchengine.

Language	*Types*	*Tokens*
French	284.432	1.301.850
Spanish	385.162	1.949.941
Portuguese	642.022	3.204.462

Table 2. Type and token counts of N Prep N sequences in the TenTen corpora for French, Spanish, and Portuguese

Table 1 lists type and token counts for all N Prep N sequences in the three corpora. In Portuguese, the construction seems to appear on a particularly frequent basis when compared to French and Spanish, which show relatively similar frequencies. The frequent occurrence of the N Prep N construction is in part due to the existence in Portuguese of hybrid forms of the type Prep + Art (*do(s), da(s), na(s), no(s)*) as well as Prep + Pron (*daquela(s)/e(s), naquela(s)/e(s); deste(s)/a(s), neste(s)/a(s)*). The equivalent constructions in French and Spanish would be of the form N Prep Article N. In order to dispose of a syntactically homogenous dataset, these constructions were not included for the present analysis. In what follows, we refer to the complete set of N Prep N sequences extracted from the corpora as dataset 1. This dataset

is noisy and contains instances in which the N Prep N sequence is not a syntactic or onomasiological unit, that is to say a naming unit (Štekauer (2001b). Removal of these irrelevant cases from a list of more than 6 million examples was beyond the scope of the present study. Despite this noise, dataset 1 was included in the quantitative survey in order to obtain an overview of the occurrence and productivity of the construction type N Prep N in the languages under investigation. Furthermore, the results from the analysis of dataset 1 offer a first point of comparison of the analysis of dataset 2.

From dataset 1, a second dataset was derived from which word triplets that did not instantiate the N Prep N construction were manually removed. This second dataset, henceforth dataset 2, focused on the internal preposition of the constructions. In a first step, all constructions overlapping in their N1 and N2 and diverging in their preposition were selected (e.g. *livre pour/d'enfants*). In a second step, the data was manually inspected and the following constructions were excluded: grammaticalized constructions (*frente a, jusqu'à, en dehors*), partitive constructions or spatial, temporal or mass nouns (*kilo de, lunes a viernes, visita a Roma, journées par semaine*), binominal pairs (*dia a dia, instant après instant*), antonyms (*chien sans/avec laisse, personnes avec/sans emploi*), preposition phrases (*N1* à base de, par hasard de), verb phrases (*mettre N1 en danger, donner N1 à N2*), and hybrid forms of the above.

Language	*Types*	*Tokens*
French	1062	6991
Spanish	547	10219
Portuguese	6795	58932

Table 3. Type and token counts for dataset 2, which includes all pairs of nouns that are attested with at least two different internal prepositions

Table 3 lists type and token counts for dataset 2. As for dataset 1, the counts for Portuguese outnumber those for French and Spanish.

Both datasets were further analysed by considering, in addition to the counts of tokens (N) and types (V), the counts of hapax legomena (V1, the formations occurring once only), the productivity measure P =

V1/N, which assesses the probability that an additional N Prep N token represents a novel, previously unobserved type, and an estimate S of the potential number of formations in use in the text type sampled by the corpus. Note that S = V + V0, where V0 is the count of formations that do not appear in the sample. S can be estimated given the numbers of word types Vk that occur once, twice, three times etc., when these counts Vk decrease in a regular way. If so, V0 can be estimated and given V0, an estimate of S = V + V0 follows immediately. For further mathematical detail on these measures, see Baayen (2009) and for the estimation of S, Baayen (2001, 2008).

Thus, we have three estimates, each highlighting a different aspect of productivity: The number of types V for the extent to which a head or modifier position is used in the corpus, the probability P that when the corpus is increased, new types will be sampled, and the limiting number of types that one might sample if the corpus size were increased to infinity.

3.2. *Analysis dataset 1*

Table 4 summarizes the frequency and productivity statistics for dataset 1, focusing on the productivity of the nominal slots in the N Prep N construction.

The upper subtable documents the counts when types are defined by the first noun of the construction. The lower subtable concerns the corresponding counts for the second noun. On the basis of the numbers of tokens N, types V, potential types S, and hapax legomena V1, the N Prep N construction appears least productive in French, of medium productivity in Spanish, and most productive in Portuguese. This ordering holds for both the first and the second noun.

The ranking of the three languages by P is different, with Portuguese having the lowest productivity measure. It should be kept in mind, however, that P is itself a function of N, and that it decreases as N (and V) increase. (As we read through a text, the rate at which new words are encountered decreases steadily.) Given that N is very much larger for Portuguese, the value of P is actually surprisingly large. Comparing Spanish and French, the similar values of P are surprising given that N is substantially larger for Spanish than for French. Therefore, the P values provide further support for the ranking based on the other statistics.

Noun1	*French*	*Spanish*	*Portuguese*
P	0.0023	0.0023	0.0017
S	20147	28755	36624
V	13719	18407	23409
N	1301850	1949941	3204462
V_1	2994	4485	5448
Noun2	*French*	*Spanish*	*Portuguese*
P	0.0028	0.0031	0.0023
S	24688	39037	49079
V	16174	23245	28545
N	1301850	1949941	3204462
V_1	3645	6045	7370

Table 4. Frequency and productivity statistics for dataset 1. The upper part of the table defines types on the basis of the first noun, the lower part bases types on the second noun

Table 4 also indicates that the second noun position of the construction is used more productively than the first noun position: all measures assume larger values in the second part of the table. The greater productivity of the modifier position makes sense from an onomasiological perspective, as the second noun slot is typically used to differentiate between subcategories of the head noun, which in Romance languages generally occupies the first noun slot.

The large numbers of hapax legomena, as well as the fact that S >> V all support – within the limits of dataset 1 – that the N Prep N construction is solidly productive in the three Romance languages under consideration here.

Further informal surveys of the prepositions *de*, *en-em*, à-a, *pour-para* as well as *avec-con-com,* again using dataset 1, indicated that French N Prep N constructions containing the prepositions *avec* and *pour* are less frequent and productive than equivalent constructions in Portuguese and Spanish containing the prepositions *con-com* and *para*. French appears to resort to other types of word formation such as NN or NA constructions instead of using N Prep N constructions containing *avec*, as in:

5a) Fr. *personne handicapée* "handicapped person"
b) Sp. *persona con discapacidad física/mental* "handicapped person"
c) Pt. *pessoa com necessidades especiais* "handicapped person"

French also shows a preference for constructions with *de* instead of *pour*. At the same time, constructions with the preposition à-a appear to be more productive and frequent in French than in Spanish and Portuguese. Semantic relations that are expressed via à in French tend to require other prepositions, such as *de* or *para*, in Spanish or Portuguese:

7a) Fr. *Verre à vin* "wine glass"
b) Sp. *Copo de vino/ Copo para vino* "wine glass"
c) Pt. *Copo de vinho* "wine glass"

3.3. *Analysis dataset 2*

Table 5 summarizes the frequency and productivity measures for data set 2, which includes only those (manually verified) examples of N Prep N constructions in which the first and second noun co-occur with at least two different prepositions. For this analysis, each combination of first and second noun and preposition counted as a separate type.

	French	*Spanish*	*Portuguese*
P	0.0594	0	0.0464
S	1748.232	-	13378.57
V	1062	547	6795
N	6991	10219	58932
V_1	415	0	2733

Table 5. Frequency and productivity statistics for dataset 2, which comprises all instances of noun pairs that occur with at least two different prepositions

As in the analysis of dataset 1, Portuguese again shows the highest type (V) and token (N) frequencies, the largest number of hapax legomena (V1), the highest estimate of possible types (S), and given the large numbers of tokens, a surprisingly large degree of productivity P. Although numbers are reduced for French, the construction – as evaluated on the basis of dataset 2 – remains solidly productive, as evidenced

by the large number of types missed in the sample (S – V = 1748-1062 = 686 = V0).

Spanish, by contrast, shows a very different pattern. There are no hapax legomena in dataset 2 for Spanish, and hence P is zero, and S cannot even be estimated (it is expected to be only slightly larger than V, if at all). The number of types (547) is roughly half of that observed for French, and less than 10% of that observed for Portuguese. In other words, internal variation of the preposition for fixed head and modifier nouns is not productive in Spanish, whereas it is productive in French and especially Portuguese. In Portuguese, we find examples of noun pairs occurring with 5 different prepositions, in French, this reduces to 4, and in Spanish, the maximum is 3.

Thus, when we consider the productivity of internal variation of the preposition, the ranking of the languages places French above Spanish. Inspection of the Spanish examples suggests a strong tendency to make use of the high frequent preposition *de* and to restrict variation in prepositions to a relatively small set of lexicalized compounds.

4. Discussion

The present study sheds new light on the vexed question of the status of N Prep N construction in Romance languages. First, the survey of N Prep N sequences in the TenTen corpora of French, Spanish, and Portuguese clearly shows that this construction contributes substantially to the lexicon (in the onomasiological sense) of these languages. In all three languages, the construction is realized in tens of thousands of examples (dataset 1). Admittedly, dataset 1 includes many instances that do not conform to the N Prep N construction. Nevertheless, even if half of the tokens and types were to be discarded, the counts of legitimate constructions still would portray this construction as the most productive onomasiological process in Romance – mirroring the evidence from Germanic languages suggests that derivational word formation is less productive than compounding by several orders of magnitude. It is therefore unlikely that N Prep N constructions in Romance languages are merely lexicalized or fossilized syntactic constructions without support of a productive process of word formation (pace Guevara 2012 and Villoing 2012). To the contrary, for all three languages, large numbers of novel types are expected to be observable in larger samples of

language use, as indicated by the (tentative) estimates of the population numbers of types (S).

An analysis of a hand-curated subset of dataset 1, comprising all attestations of N1 Prep N2 constructions in which N1 and N2 co-occur with at least two different prepositions (dataset 2), brought to light an unexpected difference between Portuguese and French on one hand, and Spanish on the other hand. Portuguese, and to a lesser extent French, exhibit productive internal variation of the preposition. Spanish, by contrast, appears not to allow its speakers the same flexibility in the choice of preposition. In the absence of hapax legomena for Spanish noun pairs, Spanish emerges as a language that avoids both "free" variation of the preposition for approximately the same meaning, as well as using different prepositions for differentiating between shades of meaning given a modifier and head noun (as instantiated for instance for French by the pair '*verre à vin*' and '*verre de vin*').

An informal survey of which prepositions are favored revealed French as showing a stronger preference for constructions containing the preposition *à* compared to Spanish or Portuguese, which use *de* or *para* more productively. The absence of *avec* in French N Prep N constructions is likely to be due to NA-constructions being preferred. In French, *pour* emerged as slightly more productive than *de* (e.g. *livre d'enfant* and *livre pour enfants*).

5. Conclusions

The present quantitative survey of N Prep N constructions in Spanish, French and Portuguese offers new empirical evidence for the discussion on Romance word formation. The two main points addressed in this study concern the lexical or syntactic status of syntagmatic compounds as well as their productivity and degree of lexicalization or fossilization.

The analysis indicates that these constructions indeed are realized according to productive processes of Romance word formation. That is to say, syntagmatic compounds are naming units that form part of the lexicon. N Prep N constructions are not merely fossilized syntactic constructions, rather, the construction type N Prep N is an important and frequently used mechanism of word formation. Still, it is important to highlight that it is neither possible nor necessary to draw a clear line between lexical onomasiological units of the type N Prep N and syn-

tactic constructions of the type N Prep N. Here, different criteria, such as the degree of fixedness, idiomaticity and compositionality play an important role.

Furthermore, the present quantitative analysis points out that internal prepositional variation is possible in N Prep N constructions in Romance languages, but that this variation displays different characteristics in the three Romance languages under investigation. Portuguese shows the highest frequency and productivity of internal prepositional variation in a large number of different semantic contexts. In contrast, the Spanish data do not allow any productivity in the internal variation of N Prep N constructions. In the same line, Spanish has the strongest tendency of employing the preposition *de* as internal prepositions in N Prep N constructions.

In conclusion, it can be stated that syntagmatic compounds of the type N Prep N form a productive and frequent part of Romance word formation. Still, their frequency and productivity as a word formation type vary in the three Romance languages, as well as their disposition for internal prepositional variation. Further studies on this subject need to consider the qualitative characteristics of internal prepositional variation, notably the semantic relation between the N1 and the N2.

5. References

Anshen, Frank & Aronoff, Mark. 1997. Morphology in real time. In Geert, E. Booij & van Marle, Jaap (eds.) *Yearbook of Morphology 1996*. Dordrecht: Kluwer Academic Publishers, 9-12.

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Baayen, Harald & Lieber, Rochelle. 1991. Productivity and English derivation: a corpus-based study. *Linguistics* 29(5): 801-844.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge.

Baayen, R. H. 2009. Corpus linguistics in morphology: morphological productivity. In Lüdeling, A. & Kyto, M. (eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton De Gruyter, 900-919.

Baayen, R. H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. H. 2001. *Word Frequency Distributions*. Kluwer.

Benveniste, Émile (ed.). 1966. *Problèmes de linguistique générale* (Bibliothèque des sciences humaines 1). Paris: Gallimard.

Bisetto, Antonietta & Scalise, Sergio. 1999. Compounding: morphology and/or syntax? In Mereu, Lunella (ed.) *Boundaries of Morphology and Syntax* (Amsterdam Studies in the Theory and History of Linguistic Science 4). Amsterdam/Philadelphia: Benjamins, 31-49.

Bisetto, Antonietta & Scalise, Sergio. 2005. The classification of compounds. *Lingue e Linguaggio* 4(2): 319-332.

Bouvier, Yves F. 2000. Definir les composes par opposition aux syntagmes. In Haeberli, Eric & Laenzlinger, Christopher (eds.) *Generative Grammar in Geneva*, 165-187.

Buenafuentes de la Malta, Cristina. 2006/04. *Entre la morfología, la sintaxis y el léxico: la delimitaciòn de la composición sintagmática en espanol* (VII Congrés de Lingüística General). Barcelona.

Di Sciullo, Anne-Marie & Williams, Edwin. 1987. *On the definition of word* (Linguistic inquiry. Monographs 14). Cambridge, MA.

Faria, André. 2010. Formação de compostos nominais de base livre do PB. In Almeida, Maria L.; Ferreira, Rosangela & Pinheiro, Diogo (eds.) *Linguística cognitiva em foco: morfologia e semântica do português*. Rio de Janeiro: Soluções Editoriais.

Fradin, Bernhard. 2009. IE, Romance: French. In Lieber, Rochelle & Štekauer, Pavol (eds.) *The Oxford Handbook of compounding*. Oxford University Press, 417-435.

Guevara, Emiliano R. 2012. Spanish compounds. *Probus. International Journal of Latin and Romance Linguistics* 24(1): 175-195.

Kabatek, Johannes & Pusch, Claus D. 2009. *Spanische Sprachwissenschaft*: *Eine Einführung*. Tübingen: Narr Franke Attempto Verlag.

Kampers-Manhe, Brigitte. 2001. Le statut de la préposition dans les mots composés. *Travaux de Linguistique* 42-43(1), 83-95.

Kornfeld, Laura M. 2003. Compounds N+N as formally lexicalized appositions in Spanish. In Booij, Geert; De Cesaris, Janet; Ralli, Angela & Scalise, Sergio (eds.) *Topics in Morphology*: *Selected Papers from the Third Meditteranean Morphology Meeting*. Barcelona: Universitat de Pompeu Fabra, 211-225.

Kornfeld, Laura M. 2009. IE, Romance: Spanish. In Rochelle Lieber & Pavol Štekauer (eds.) *The Oxford Handbook of Compounding*. Oxford University Press , 436-453.

Lieber, Rochelle. 1992. *Deconstruction Morphology*: *Word Formation in Syntactic Theory*. Chicago/London: University of Chicago Press.

Lieber, Rochelle & Scalise, Sergio. 2007. The lexical integrity hypothesis in a new theoretical universe. In Booij, Geert; Ducceschi, Luca; Fradin, Bernhard; Guevara, Emiliano R.; Ralli, Angela & Scalise, Sergio (eds.) *On-line Proceedings of the Fifth Mediterranean Morphology Meeting*, 1-25.

Masini, Francesca. 2009. Phrasal lexemes, compounds and phrases: A construcionist perspective. *Word Structure* 2(2): 254-271.

Masini, Francesca & Scalise, Sergio. 2012. Italian compounds. *Probus. International Journal of Latin and Romance Linguistics* 24(1): 61-91.

Masini, Francesca & Thornton, Anna. 2007. Italian VEV lexical constructions. In Booij, Geert; Ralli, Angela & Scalise, Sergio (eds.) *Morphology and Dialectology* 6: 148-189.

Pacagnini, Ana M. J. 2003. Compuestos sintagmáticos y alternancia preposicional. *Moenia* 9: 159-172.

Rainer, Franz. 2016. Italian. In Müller, O. P.; Ohnheiser, Ingeborg; Olsen, Susan & Rainer, Franz (eds.) *Word Formation*: *An International Handbook of the Languages of Europe*. Berlin/Boston: Mouton de Gruyter, 2712-2731.

Rainer, Franz. 2016. Spanish. In Müller, O. P.; Ohnheiser, Ingeborg; Olsen, Susan & Rainer, Franz (eds.) *Word Formation*: *An International Handbook of the Languages of Europe*. Berlin/Boston: Mouton de Gruyter, 2620-2640.

Rio-Torto, Graça & Ribeiro, Sílvia. 2009. Compounds in portuguese. *Lingue e Linguaggio* 8(2): 271-291.

Rio-Torto, Graça & Ribeiro, Sílvia. 2012. Portuguese compounds. *Probus. International Journal of Latin and Romance Linguistics* 24(1): 119-145.

Schlechtweg, Marcel & Härtl, Holden. 2015. Compound versus phrase: Evidence from a learning study (10th Mediterranean Morphology Meeting). Haifa.

Štekauer, Pavol. 2001b. Fundamental principles of an onomasiological theory of English word-formation. *Onomasiology Online* 2: 1-42.

van Goethem, Kristel. 2009. Choosing between A+N compounds and lexicalized A+N phrases: The position of French in comparison to Germanic languages. *Word Structure* 2(2): 241-253.

Villoing, Florence. 2012. French compounds. *Probus. International Journal of Latin and Romance Linguistics* 24(1): 29-60.