Hybrid Methods for the Extraction and Comparison of Multilingual Collocations in the Language for Specific Purposes of Marketing

Guadalupe Ruiz Yepes
Heilbronn University, Am Europaplatz 11, 74076 Heilbronn, Germany
Guadalupe.ruiz-yepes@hs-heilbronn.de

Abstract. This paper presents a status quo of an ongoing research with the main focus on the cross linguistic comparison of collocations. The main aim is to compare the collocations extracted from a multilingual comparable corpus – German, English, Spanish – in the language for specific purposes of marketing using hybrid methods. To achieve this objective the collocations are defined and clearly distinguished from other phrases, before describing the methods which have so far been used to extract collocations. Special attention will be paid to hybrid methods that combine linguistic and statistical information. After comparing these methods, one of them will be applied to the corpora. The findings will be analyzed and conclusions drawn in order to list differences in the structure of German, English and Spanish collocations in the field of marketing. The results and findings of this piece of research can be applied to improve the performance in the field of Second Language Acquisition, Translation and Translator training.

Keywords: collocations, languages for special purposes, marketing, corpus linguistics, phraseology of marketing, translation and translator training.

1 Introduction

According to Fluck (1996: 11), there has been much discussion in Linguistics on the exact nature of languages for specific purposes (LSP) and a widely-accepted definition does not exist. However, most linguists in this field share the opinion that LSP is a variety of common language, has developed from common language and uses the grammatical means of common language (Arntz/Picht/Schmitz 2014). The formation of collocations in LSP follows therefore similar rules as in common language. The formation of collocations is very often domain specific, since words which do not participate in a collocation in everyday language often do form part of a collocation in a LSP, i.e. the noun “file” collocates with the verbs “create”, “delete”, “save” in texts about computers, but not necessarily in other contexts (McKeown and Radev, 2000: 510). This conclusion opens the doors of research to the field of LSP phraseology which is still in its infancy compared to common language phraseology.

The main aim of this paper is to compare the collocations extracted from a multilingual comparable corpus, but also to compare different methods for the extraction of
collocations. The corpus we are working with is composed of marketing texts. It is a comparable corpus composed by three corpora in three languages (German, Spanish and English) which are comparable because they belong to the same text type; all three corpora consist of articles on marketing topics, published in scientific journals. Firstly collocations will be defined in order to distinguish them from other word combinations such as idioms or free word combinations. The next step will be to describe existing methods for the extraction of collocations. Finally one of these methods will be applied on the corpora, findings will be analyzed and the consequences of these findings for translation and foreign language training will be highlighted.

2 What is a “collocation”? 

There are numerous definitions for the phenomenon “collocation”. Originally the term “collocation” was used in a very broad sense to describe the “general event of recurrent word co-occurrence” (Seretan, 2011: 13). But this statistical view has been replaced later by a linguistically-motivated one, in which the items in a collocation are syntactically and semantically related. In recent studies it has been suggested to use the term “co-occurrence” for the recurrent co-appearance of two words, while the term “collocation” is reserved for the phraseological (linguistic) approach. This distinction between “co-occurrences” and “collocations” seems to be accepted and will be adopted in this paper.

Since the term collocation was introduced by Firth in 1957, researchers from all kind of disciplines have dealt with it. Consequently, there are definitions of “collocation” in phraseology, computational linguistics, corpus linguistics, etc. Each discipline tries to define “collocation” to meet it needs. As a consequence, the term is very vague and not clearly defined, but despite this lack of clarity, two traditions can be distinguished: one follows Firth’s empirical postulate within the British contextualism, the other has its origin in the German-French lexicography (Hausmann 1984). Unlike British contextualism where the main focus is on how frequently two words are combined, Hausmann’s focus is on the semantic interrelationship of these words.

According to Bartsch, a good definition of collocation has to consider both aspects:

The aim of devising a definition of collocation is twofold. In the first instance, it serves as the basis for the computer-aided extraction of collocation candidates, i.e. potential collocations, from the corpus. In a second step, a definition of collocations based on qualitative linguistic criteria must be devised to enable the systematic description of the structural and functional properties of collocations and the distinction between different collocational structures (Bartsch, 2004: 68).
2.1 Distinction between the Term “Collocation” and other Word combinations

In Phraseology, collocations are at the interface between free word combinations and idioms. According to McKeown and Radev (2000: 508) “an idiom, […], is a given rigid word combination to which no generalities apply; neither can its meaning be determined from the meaning of its parts nor can it participate in the usual word-order variations”. On the other hand, according to Cowie (1981: 223-235) a free word combination can be described using the general rules of grammar, for example, considering the semantic constraints on the words which appear in a certain syntactic relation with a given headword. “Collocations fall between these extremes and it can be difficult to draw the line between categories” (McKeown and Radev, 2000: 508).

2.2 Criteria for the identification of collocations

To determine if and to which extent a word co-occurrence is really a collocation, idiomaticity and stability, together with other aspects must be verified. With this purpose in mind, we present four criterions to identify collocations after the implementation of quantitative methods.

First criterion: verify that the word co-occurrence is not an idiom. As stated above an idiom is a word combination whose meaning cannot be determined from the meaning of its parts. Therefore the first criterion for identifying a collocation is: “word combination whose overall meaning can be derived from the meaning of each word”. But idioms are not only on the semantic level fixed associations of words. In contrast to collocations they also can display a fixed syntactic behavior not allowing modifiers, the passive voice, etc.

Second criterion: verify the psycholinguistic stability of the word co-occurrence. Stability can occur as formal, lexical, syntactic, psycholinguistic, pragmatic or semantic stability. Of particular significance for collocation is the psycholinguistic stability which says that collocations like other lexemes are firm parts of the mental lexicon and can be reproduced. Therefore the second criterion is: “collocations are stored and recalled as a unit from our mental lexicon”.

Third criterion: verify the pragmatic stability of the word co-occurrence. The pragmatic stability which depends on the recurrence of words being combined is of similar significance for collocations (Caro Cedillo, 2006: 41). The third criterion is: “the institutionalization of a collocation is determined by its frequent usage” (Seretan, 2011: 16).

Forth criterion: verify the word co-occurrence does not admit the substitution of one of its components by a synonym without altering the meaning. Research by Pearce (2002) into collocations extraction produced a method based on the substitutions for synonyms. His method is based on the assumption that in a free word combination, it is possible to substitute one of its components by a synonym without altering too much the meaning. “If a phrase does not permit such substitutions then it is a collocation” (Pearce, 2002: 1533). Therefore, the forth criterion is: “A collocation can only be
identified as such if the speaker has got several collocators available which can be combined with a base from the semantic point of view, but only one of these collocators is preferred in use.

3 Extraction of Collocations from Corpora

Just as there are different definitions of collocation, there are also different methods of extracting collocations from corpora. The following methods have been used so far:

3.1 Methods based on co-occurrence considerations

Depending on research interests and research purposes, different association measures were developed. So the corpus linguist has to decide in favour of certain association measures. According to Evert (2009: 1236), there is no “perfect” association measure. Therefore several ones should be used for a study in order to get different results which can be compared. There are two groups of association measures that pursue opposing goals: “effect-size measures” (mutual information, Dice, odds, ratio) and “significance measures” (z-score, t-score, simple II, chi-squared, log-likelihood). The linguist using “effect-size measures” is looking to find “How much does observed co-occurrence frequency exceed expected frequency” and when using “significance measures” “How unlikely is the null hypothesis that the words are independent?” (Evert, 2009: 1228).

3.2 Methods based on Collocation Patterns

Hausmann’s collocation typology (1989: 1010) distinguishes six types of collocations: noun + noun, adjective + noun (as an object), verb + noun (as a subject), adverb + verb and adverb + adjective. Weller and Heid (2010: 3195) call these types “collocation patterns”, once extracted from corpora they call them “collocation candidates”. In order to extract collocations from corpora using “collocation patterns” the corpora have to be annotated at least with POS-tags. However, a higher performance can be reached if the corpora are also syntactically analyzed (parsed).

3.3 Hybrid methods

Pamies and Pazos (2005: 317-329) compare different association measures and come to the conclusion that mathematical methods alone are not enough and prefer instead the use of “hybrid” methods (2005: 327). There are methods which use a combination of co-occurrence calculation and linguistic criteria in the form of collocation patterns. That is, a hybrid system combines statistical methods and multilingual parsing for detecting accurate collocational information. But, in which order are the filtering methods to be used? There are different approaches:

- First statistical methods are used and – for refining the filtering – additional collocation patterns are applied to the achieved results (Smadja 1993).
First collocation patterns are used and in the second step the statistical methods are applied to the achieved results (Krenn (2000), Evert (2005), Seretan and Wehrli (2006)).

Smadja (1993) is the most representative researcher for the first approach. He developed a system called Xtract that retrieved word pairs using a frequency-based metric in the first place. The metric computed the z-score of a pair of words. In addition to the metric, Xtract used three additional filters based on linguistic properties. As a final step, an evaluation of the retrieved collocations was carried out by a lexicographer in order to estimate the number of the true lexical collocations retrieved.

Another example for the first approach is ConcGram 1.0. This software was developed as an “inclusive” search engine for phraseological units and works on the basis of co-occurrence considerations. It is left to the linguist which of these co-occurrences are significant word combinations and which are chance word co-occurrences. After entering a command, the software compiles a list of “unique words” which are the basis for showing “ConcGrams”, whereas ConcGrams are both adjacent and non-adjacent word co-occurrences which can appear in any order in the corpus. As soon as the ConcGram lists are compiled, statistical methods can be applied. They allow a reduction of the lists and provide clear information on non-relevant word combinations which can be ignored. The applied statistical methods are t-score and MI tests whose formulas are explained in detail by Barnbrook (1996: 88-106). When doing computer-assisted corpus analysis though, automatically compiled frequency tables are – despite using association measures – not always directly usable, but need a human selection input. For this reason, when working with this kind of software the researcher has to apply manually collocation patterns in order to be able to extract collocations candidates. We also suggest to improve the process by adding work stages based on the criterions presented in section 2.2.

On the other hand, Seretan and Wehrli are the most representative researchers for the second approach. They consider that “syntactic analysis of source corpora is an inescapable precondition for collocations extraction” (2006: 1). The hybrid method Seretan and Wehrli (2006) developed relies on a deep parser called Fips (Wehrli, 2004) and can be seen as a two-stage process. Firstly the collocation candidates are identified by the parser while POS-tagging and parsing the text corpora. Secondly the candidates are scored and ranked using specific association measures (Seretan and Wehrli, 2006: 2). In this approach the parser is used in the first stage of the extraction in order to identify the collocation candidates and the criterion they employ firstly for the selection of the collocation candidates is the syntactic proximity. As Seretan and Wehrli explain, “as the parsing goes on, the syntactic word pairs are extracted from the parse structures created, from each head-specifier or head-complement relation. The pairs obtained are then partitioned according to their syntactic configuration” (2006: 2). An advantage of this method is that the pairs obtained are then partitioned according to the collocations patterns presented by Hausmann (1984). A major disadvantage, however, is the dependence on a specific linguistic theory. Finally, the log-likelihood test is applied.
4 Extraction and comparison of word co-occurrences

The terms “satisfaction” and “loyalty” occur extraordinarily frequently in the English corpus. The occurrence of the Spanish and German equivalents is similarly frequent. Therefore, certain co-occurrences of the words “satisfaction” and “loyalty” were extracted and compared with their German and Spanish analog co-occurrences using the software ConcGram. The words “satisfaction” and “loyalty” – in German “Zufriedenheit” and “Loyalität” and in Spanish “satisfacción” and “lealtad” – belong to the category of business psychology in the LSP of marketing. They occur particularly frequently with the words “consumer/customer”, “consumidor/cliente” and “Konsument/Kunde”. To verify if this word combinations were collocations at all the four criterions explained in section 2.2 were applied. Once checked that they met the criterions, a cross linguistic comparison was carried out.

<table>
<thead>
<tr>
<th>Table 1. Collocations of “satisfaction” and “loyalty” in the English corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
</tr>
<tr>
<td>loyalty</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>satisfaction</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Collocations of “Zufriedenheit”/“Loyalität/Bindung” in the German corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
</tr>
<tr>
<td>Zufriedenheit</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Bindung</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Loyalität</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Collocations of “satisfacción” and “lealtad” in the Spanish corpus.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
</tr>
<tr>
<td>lealtad</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>satisfacción</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
The comparison of the three languages English, German and Spanish has led to the conclusion that the collocations in one language are often expressed by other types of word combinations in the other languages. In the analyzed cases English collocations with the structure noun + noun are often expressed by compounds or possessive markers in the German language, and by prepositional phrases – either with or without article – in the Spanish language.

These findings provide also evidence of the fact that collocations do not preserve their meaning across languages. While the compound “Kundenloyalität” occurs in the corpus only 19 times, the compound “Kundenbindung” occurs 286 times in the same contexts as “customer loyalty” occurs in the English corpus. Therefore we can assume that “Kundenbindung” is the German equivalent for “customer loyalty” rather than “Kundenloyalität” which is the literal translation. Literal translations lead very often to unnatural and awkward sounding formulations.

Last but not least, these findings have also shown that while in the Spanish language the terms “consumidor” and “cliente” are used in the same way, in German and English the terms “Kunde” and “customer” are preferred. This is related to socio-linguistic and pragmatic factors – the words with the root “Konsum-”/“consum-” do have a negative semantic prosody or connotation in both English and German, but not in the Spanish language.

5 Consequences for translation and foreign language training

Among the numerous fields of applications of this study, translation training and foreign language training must be particularly highlighted. For translators, especially when translating into the foreign language, collocations are a frequent source of errors, all the more when it comes to detecting false friends. For the term “customer loyalty” the most obvious German translation would be “Kundenloyalität”. However, after a detailed research in the corpus, we have found out that the term “Kundenbindung” is preferred. These findings are of major importance for translators, especially when translating into a foreign language because collocations are unpredictable for non-native speakers of a language. The conducted corpus search has also shown that while in the Spanish language the terms “consumidor” and “cliente” are used in the same way, in German and English the terms “Kunde” and “customer” are preferred due to the semantic prosody of these words. This leads to a clear consequence for the translator: “consumidor” is not necessarily translated as “Konsument” or “consumer”, which would be the literal translation, but rather as “Kunde” and “customer”.

For foreign language didactics, these findings are equally valuable. The (foreign) language teacher has a base for explaining to the students that not only grammatical aspects are important for the use of a language, but that there are also uses which can neither be explained by grammatical nor semantical rules. Aspects which are related to the use of a language, which are pragmatic and can only be taught and learned in connection with the culture and the values of a society.
References