Friedrich Schiller University Jena

Faculty of Arts
Institute of English and American Studies

The A *as* NP construction in English and Russian

Master thesis submitted to be awarded the academic degree
Master of Arts (MA)

Submitted by: Pianykh, Kristina
Matriculation no.: 176846

Born on 28.07.1995 in Voronezh, Russia

First reviewer: Prof. Dr. Gast, Volker
Second reviewer: Dr. Haas, Florian

11.12.2020
Jena

Zussamenfassung

Die jüngsten Fortschritte in der quantitativen Korpuslinguistik haben ein erneutes Interesse an kollokutionären Studien geweckt. Der Schlüsselfaktor für den Aufschwung der Kollokationsforschung in den letzten Jahrzehnten war die Entwicklung und Integration neuer Methoden und Technologien in der linguistischen Analyse. Bevor dieses Feld unter Einbeziehung von künstlicher Intelligenz und neuronalen Netzen zu einer Quelle neuer Lösungen für komplexe Probleme der Sprachverarbeitung wurde, war es Mitte der 1990er Jahre in der Lexikographie äußerst fruchtbar. Doch so paradox es klingen mag, trotz einer breiten Palette von kollokationsbasierten Technologien gibt es keinen Konsens über die Definition der lexikalischen Assoziation. Stattdessen variiert ihre Interpretation je nach wissenschaftlicher Disziplin und Anwendungszweck drastisch. In der vorliegenden Studie geht es um die A-*as*-NP-Konstruktion (z. B. *easy as pie*, *hard as a stone*, *poor as church mice*), die weniger für den eigentlichen Vergleich mit der Entität, die durch das Substantiv ausgedrückt wird, als vielmehr zur Betonung und subjektiven Intensivierung verwendet wird. Obwohl es ein perfekter Kandidat für eine kollostruktive Studie ist, wurde es bisher nur wenig untersucht. Trotz des mangelnden Interesses an der leicht zu traktierenden Konstruktion mit relativ fester Struktur hat sich gezeigt, dass das Verhalten der A-*as*-NP-Konstruktion wesentlich komplexer ist als bisher angenommen.

Table of Contents

1. Introduction

The recent advances in quantitative corpus linguistics have sparked renewed interest in collostructionist studies. The key factor in the boosting of collocational research in the last few decades has been the development and integration of new methods and technologies into linguistic analysis. Before this field became a source of new solutions to complex problems of language processing, incorporating artificial intelligence and neural networks, it was extremely fruitful in lexicography in the middle of the 1990s. For example, the development of methods for collocation recognition, resulted in the Collins COBUILD English Dictionary (Sinclair 1995) and a number of dictionaries of collocations (Kjellmer 1994, Lea 2002). The predominant application of advanced collocation-based techniques today is a vast variety of natural language processing tasks, e.g., natural language understanding and machine translation (Sag et al., 2002, Wehrli and Nerima 2015). Collocational analysis has also been extensively employed in second language teaching and computational linguistics for such purposes as the research on synonymity and contextual semantics (Leech 1974, Partington 1998), sense disambiguation, information extraction, sentiment analysis (Monti et al., 2018), as well as vector-space modeling in the field of distributional semantics (Evert 2009: 1217-1218).

However, paradoxical as it may sound, despite a wide range of collocation-based technologies, there is no consensus on the definition of the lexical association. Instead, its interpretation varies drastically depending on a scientific discipline and an application purpose. In the framework of Construction Grammar, which is adopted in the present study for methodological reasons, the notion of collocations is inextricably linked with grammatical context. The linguistic environment of a lexical sequence, viewed as a collocation, plays an important part in determining strength, direction, as well as productivity of this lexical association. The grammatical embedding of a collocation is conventionally expressed by a multi-slot construction (unless the concept of collocations is extended to encompass the relationship of attraction between one-slot constructions and the words, filling in the slot; cf. Stefanowitsch and Gries 2003). From the methodological point of view, such formations as multi-slot constructions represent a good starting point for investigating collocations, confined to a specific grammatical context with a certain function.

The subject for this purpose in the current study is the A *as* NP construction (e.g., *easy as pie, hard as a stone, poor as church mice*), used not so much for the actual comparison with the entity, expressed by the noun, as for emphasis and subjective intensification.

Although it represents a perfect candidate for a collostructional study, it has so far been poorly studied. Despite the lack of interest to the easily tractable construction with a relatively fixed structure, it has been shown that the behavior of the A *as* NP construction is substantially more complex than previously thought (cf. Moore 2004). Given conventionalized and highly idiomatic instances such as *right as rain* or *sick as a dog*, there are, nevertheless, numerous creative and innovative uses of sporadic nature, indicative of a certain productivity potential of the construction. This observation raises a number of questions: when are such extensions to novel lexical items allowed and considered perfectly natural while in other cases they are blocked? What causes some words to collocate with an unlimited variety of semantically different elements when others co-occur with just a handful of acceptable candidates? Could these processes be generalized over to account for the sporadic productivity of the construction? These and similar questions have provided inspiration for the current study, intended to investigate lexical association patterns of the A *as* NP construction in English and Russian to go beyond language-specific issues and draw a comparison on the cross-linguistic level. One should also point out that the present study draws significantly on Pianykh 2019.

Considering the aforementioned, the following paper is structured as follows. Chapter 2 outlines the controversy around the notion of collocations, as well as their general properties. The third chapter is dedicated to the overview of the previous research on similies as multi-slot constructions from the cross-linguistic perspective. Chapter 4 provides the characteristics of the A *as* NP construction in English and Russian and delineates its limitations in some construction grammar frameworks. The overview of the most common measures of symmetrical and asymmetrical lexical association is provided in Chapter 5 while Chapter 6 outlines the methodology, employed in the present corpus study. In the end, I summarize and compare the results of the analysis in both languages and demonstrate how the present study can be extended.

2. Collocations

The notion of *collocations* has been a subject to controversy in linguistics. Despite the general intuition that certain words have a tendency to appear next to each other, there is no general consensus in the linguistic literature on what specifically is to be understood under this term. The variation in the usage of this concept is attributed mainly to different interpretations of the semantic status of collocations, as will be shown below.

In the phraseological tradition, a collocation is defined as a lexicalized unit of words. For instance, the word *speech* collocates only with the verb *give* and not with *make*, *hold* or *do* (Benson et al. 1986, Mel’cuk 1998, Hausmann 2003). Other examples include to a certain extent formalized combinations with the verbs *make* and *do* (e.g., *make dinner/research/the bed* but *do a task/hair/the dishes*, etc.). Such an interpretation of collocations is referred to by Evert (2009) as lexical collation. In contrast, in computational linguistics, the concept is conventionally used in reference to word combinations with idiosyncratic semantic and syntactic properties, or, in other words, non-compositional configurations (Choueka 1988, Manning and Schütze 1999: 184). This approach goes even further than the phraseology-oriented research insofar as it reduces the denotational field of the term *collocations* by limiting it to the extreme end of idiomaticity. Sag et al. (2002) and Evert (2009) categorize this kind of collocations as multiword expressions. A comprehensive overview of all the different definitions of *collocations* can be found in Bartsch 2004: 27–64.

However, before the notion of collocations could be exploited in processing of large quantities of text, which revolutionized the field of lexicography (Sinclair 1991), it was already explored and recognized by Firth in the 50s of the last century (Firth 1957). In the Furthian tradition, collocations are examined from the purely statistical point of view and understood as statistically significant co-occurrences of words regardless of their compositional meaning (cf. Jones and Sinclair 1974, Sinclair 1991, 2004, Stubbs 1996, Hoey 2005, Michelbacher et al. 2007). They are, therefore, frequently identified with the use of statistical methods, aimed at measuring the strength of attraction between lexical items. The purpose of these calculations is to identify which word pairs co-occur with statistically significant regularity, given their raw frequencies in a particular corpus. Since word frequencies are known to be distributed in a text unevenly (Zipf 1949), these association measures are intended to capture these differences and take them into account in determining collocational strength. According to Evert (Evert 2009: 1243), to this day, there are over 50 lexical association measures, designed for various applications (see Chapter 5 for more details).

Traditionally, a collocation is considered a lexical bigram, a combination of two words (Jones and Sinclair 1974, Sinclair 1987), but some scholars embark upon the research on multiword expressions. They are also known in the literature as clusters (Kenny 2000: 99), multi-word strings (Mauranen 2000: 120) and lexical bundles (Biber and Conrad 1999) (see Biber 2009 and Greaves and Warren 2010 for general discussion). Similarly, the collocational window, i.e. the distance between the collocates, is determined in studies differently. Apart from the span of words within which items count as a collocation, collocations vary as to whether they are defined as within-sentence or cross-sentence combinations. As can been seen, the syntactic configuration of a collocation and its definition vary drastically, depending on the needs and specifications of a study.

The phenomenon of collocations has, in fact, a psychological basis. As claimed by Hoey (2005), a recurrent syntagmatic sequence of words strengthens the associative link between them in a speaker's mental lexicon. According to Hoey, every lexical element has its mental representation, stored in memory together with all the linguistic and extralinguistic contexts in which this item has been witnessed. It follows that a certain linguistic environment primes the associated with it lexical item which in this particular case is preferred over others. The context-dependency of collocations was also addressed by Sinclair, who highlighted the reciprocal relationship between words and their lexical, as well as grammatical surroundings (Sinclair 1991: 108). In fact, the distributional preferences of individual words shape the basis for the grammar knowledge, and being aware of them is what is called to know the language. One such information unit, stored in the mental representation of each word, is its semantic preference: the tendency of an item to co-occur with the items from a specific semantic category (Sinclair 1991, Hoey 2005, Dilts 2010). For example, as shown in the study by Stubbs (Stubbs 2001), the adjective *large* attracts nouns denoting quantity and size. Generalization over the semantic field of collocates (i.e. the components of a collocation) in regard to each individual word is, however, more than an ambitious task and in practice rarely straightforward.

In the last few decades, the context-dependency of collocations has been recognized and made subject to intense scrutiny. Specifically, it was found that certain words tend to be bound to or at least to be strongly influenced by the grammatical properties of the contexts they appear in. This view has become particularly popular in Construction Grammar (e.g. Fillmore 1988, Goldberg 2013), which reconciles lexicon and grammar – the widespread dichotomy, originated in Generative Grammar. In this framework, grammar is represented by the constructicon – a hierarchy of constructions, varying in abstraction – where each construction, apart from its syntactic properties, is associated with a specific meaning. As this approach lends itself to the purpose of the context-dependent research on collocations, it is not surprising that it gave rise to the so-called collostructional analysis, aimed at investigating the relationship between filler-slot constructions and the attracted to them lemmas. For example, it was found that the $[_{\text{Head}}$ N $[_{\text{Modifier}}$ waiting to happen]] construction (e.g., *Marriage to Mandy Smith was a disaster waiting to happen*) demonstrates a strong tendency to be used mostly with negatively connotated nouns such as *accident, disaster, earthquake, invasion, revolution*, etc. (Stefanowitsch and Gries 2003). The interest to the vast and promising field of the collostructional research resulted into the development of the family of methods, intended to measure various kinds of relationships between constructions and the words, most likely appearing in them, as well as between the lexical items of multiple-slot constructions (for an overview, see Stefanowitsch and Gries 2009).

The main metric of the collostructional analysis is collostructional strength, i.e. the strength of association between collocates. It is believed that high association scores are indicative of potential lexicalization of a specific instance of the construction in question. In the usage-based approach, such instances are considered conventionalized lexical prefabs, or simply idioms (cf. Erman and Warren 2000, Diessel 2019) due to their verbatim recurrence and, to a certain degree, entrenchment. The sequential link between the elements under investigation gets stronger, the more often they are processed together in a string (Bybee 2002, 2010: 33–37). This cognitive process is referred to as “chunking”, or the automatization of a frequently activated sequence of items, perceived as an unanalyzable unit (Langacker 2008: 60–73).

In this regard, Desagulier goes even further by proposing an inverse relationship between the collocational strength and the productivity of such instances: “the higher the association strength, the higher the level of autonomy, the lesser the productivity” (Desagulier 2016: 10). In the context of constructions, productivity refers to the potential of a specific syntactic configuration (an argument-structure construction par excellence) to be extended to novel lexical elements. This is specifically the case with constructions of a lower level of schematically where one slot of the construction is lexically filled while the other is subject to lexical variation. For instances, the ditransitive construction [Subj V Obj_1 Obj_2] with the verb *give* is deemed highly productive as there are hardly any lexical restrictions to the slot Obj_2. In contrast, the subschema of the ditransitive construction indexed on the verb *send* is far less productive as it limits the semantic field of the Obj_2 to the domain of postal service. It is for this reason, that constructions down the collostructional strength are less schematic but more productive: they are more likely to be the source of creative use, which means they can exhibit greater structural and semantic variability.

3. Similie as a Multi-Slot Construction

As evidenced by the previous research, similies represent a poorly investigated field of multi-slot constructions. Indeed, little has been written on the subject in the phraseological literature (although see Norrick 1986 for an exception), compared to other, more complex multi-word expressions such as phrasal verbs, idioms, and speech formulae. The lack of research interest to similies could be attributed to their cross-linguistically fixed syntactic structure (Haspelmath and Buchholz 1998, Omazić 2002, Mokienko 2016), rather limited structural variability and allegedly rigid lexicon (Moon 2008: 3). All the factors combined account for the apparent lack of motivation in corpus research, dismissive of 'tractable' patterns. Having said that, however, the study of the *as*-comparison in English by Moon (2008) demonstrated the underappreciated scientific potential of the quantitative analysis of similies, which revealed that there is more to the family of these constructions than previously thought. Later studies contributed to the field by analyzing the so-called A *as* NP construction in other languages, such as Croatian (Omazić 2002, Parizoska and Filipović Petrović 2017) or Italian (Giacinti 2019).

In the phraseological literature, similies are traditionally defined as institutionalized expressions that describe the relationship of comparison between two entities, usually connected by means of a preposition (*as* in English as in *cold as ice*, *как* in Russian as in *pobityj kak sobaka* (lit. *beaten as a dog*), *kao* in Croatian as in *ljut kao ris* (lit. *angry as a lynx*; Parizoska and Filipović Petrović 2017), *come* in Italian as in *liscio come l' olio* (lit. *smooth as silk*; Giacinti 2019). From the typological perspective, they could be categorized as equative and similative constructions while the preposition used to link the components of comparison is referred to as the standard marker (Haspelmath and Buchholz 1998). Similative constrictions express a comparison of equality, which is established between the entities, frequently denoted by a substantive and an attribute (1). The similative constructions, in contrast, are claimed to encode the relationship of similarity, typical of verbal structures (2).

1. a. Welsh

   Mae e cyn ddued â 'r frân.

   'He is black like a crow.' (Haspelmath and Buchholz 1998: 285)

   b. German

   Zürich ist so groß wie Wien.

   Zurich is as big as Vienna. (Haspelmath and Buchholz 1998: 278)

2. a. Italian

   Louis fume comme une cheminée.

   Louis smokes like a chimney. (Haspelmath and Buchholz 1998: 279)

   b. German

   Robert schmimmt wie eine Ente.

   Robert swims like a duck. (Haspelmath and Buchholz 1998: 278)

The A *as* NP construction, which is the subject of research in the current study, is an example of an equative expression, and it is for this reason that the further discussion will mostly be limited to this particular category of similies. Judging by examples (1a) and (1b), the underlying structure of the construction in question could be generalized with the help of the following frame: [ADJECTIVE preposition NOUN PHRASE]. Additionally, the noun phrase (NP) can optionally be followed by a prepositional phrase or expanded by a verbal structure. The entity or the concept, encoded by the NP, is commonly referred to as the vehicle of the comparison while the property of this object, expressed by an adjective, is identified as the topic (Ortony 1993, Glucksberg 2001, Chiappe and Kennedy 2001, Chiappe et al. 2003), or the tertium (Norrick 1986, Parizoska and Filipović Petrović 2017). In the given study, we will adopt Norick's convention of labeling these notions.

The lexical stability of high-frequency realizations of the construction has been addressed in the literature with a variety of terms: 'stock similes' (Norrick 1986), 'familiar similes' (Fernando 1996: 19), 'frozen similes' (McCarthy 1998: 131), 'idiomatic similes' (Carter 1998: 67). Other scholars such as Makkai pointed out that *as*-similies are nonidiomatic but nevertheless institutionalized (Makkai 1972: 338). Moon (2008) and Desagulier (2016) went even further by tentatively suggesting that the English A *as* NP construction is relatively productive although a large share of its instances are admittedly more idiomatic, i.e. automatized as non-compositional lexical sequences. Drawing on Gries's method of co-varying collexeme analysis of collocational strength between the slots of a multi-slot construction, Desagulier (2016) and Pianykh (2019) proposed that the productivity of the A *as* NP construction in English and Russian could be contingent on the strength of the mutual lexical association. To put it another way, the higher the attraction between co-occurring lexemes, the lower the instantiation of the construction in the hierarchy of schematicity because both slots have specific lexical realizations. It appears that the so-called 'stock similies' represent just one extreme end of the schematicity gradient of the *as*-constructions, characterized by a high degree of idiomaticity and a low degree of abstraction. However, as claimed by Desagulier and Pianykh, between the idiomatic expressions and the extremely scare instances on the other end of the continuum, the A *as* NP construction exhibits some lexical variation and productivity peaks.

As has been alluded before, similies show little structural variability. The exceptions are listed below:

- insertion of optional words: e.g., Cro. *točan kao (švicarski) sat,* lit. *punctual as a (Swiss) watch*, i.e. 'very punctual' (Parizoska and Filipović Petrović 2017: 351);

- ellipsis of the adjective: e.g., *quick as lightning/as lightning* (Parizoska and Filipović Petrović 2017: 351);
- alternating forms in number: e.g., *poor as a church mouse/poor as church mice*;
- insertion or removal of an article: e.g., *hard as (a) stone* (especially relevant for English).

This raises a question of whether morphological alternates should be considered as individual instances or variant realizations of a single similie. The current study adopts the view that disregards the structural variation and focuses on the conceptual content, instead (however, see Moon 1998 and Moon 2008 for an overview of different approaches to variation).

In regard to the semantic aspects of similies, they are not uncommonly discussed separately for the tertium and the vehicle. The latter has been shown to generally denote the objects and phenomena, well known to the corresponding speech community. In the extensive overview of similies by Norrick (1986), the domain of animals is by far the main source for various lexical realizations of the vehicle, followed by natural products (e.g., water, flowers, daisy, horn), artifacts (e.g., tools, clothing, buildings, etc.), or elements of folklore. According to Norrick, the prevalence of animals in this list can be accounted for by the suggestion that they "provide serviceable cognitive models for perception and classification in superficially quite dissimilar contexts" (Norick 1986: 41). The high frequency of occurrence of the lemma *dog* as the vehicle is also indicative of the relevance of this animal in the long-term immediate contact with the human through playing and, originally, hunting. Hanks (2005) and Moon (2008) report similar observations, adding plants on the list of most typical semantic categories for the vehicle. In comparison, the largest share of tertia is claimed to refer to colors (Norrick 1986), general physical properties, dimensions, speed, and age (Moon 2008). In general, it is assumed that the semantic repertoire of the tertium is more limited, compared to that of the vehicle.

The meaning of the whole similie construction is believed to correspond to the meaning of the tertium. It is commonly expressed by an adjective, which prototypically refers to the most salient property of the entity, designated as the vehicle. For example, the adjectives in *red as blood, cold as ice, hot as hell* and *white as snow* encode the characteristics that are traditionally and, most likely universally, associated with the given phenomena, considering they are conceptually derived from sensory experiences or widely accepted beliefs. The reference to the most salient property of an object or a concept is used, therefore, for intensification of the meaning of the adjective. In other words, the meaning of the construction can be expressed with the help of the adverb *very*: *red as blood – very red*, *cold as ice – very cold, hot as hell – very hot, white as snow – very white*.

Despite the widespread, although intuitive, idea of the prototypicality link between the vehicle and the tertium, there is a large body of evidence that refutes this assumption. Following Norrick, "the vehicle need not be a prototype of the tertium evident in the nature of things, just so long as the society in question accepts the relation." (Norrick 1986: 40). According to Black (1962: 39 ff.), the property must be an associated commonplace of the vehicle. It is for this reason that we accept not only *black as a crow* but also *meek as a sheep, proud as a peacock* or *bold as a lion*, namely because these attributes are conventionally derived from our inculcation with a set of beliefs, valid in our culture, and not so much from the objective reality. The personification of animals and physical objects with typically human character traits and feelings, also known as anthropomorphism, roots back to the tradition of proverbs, tales and fables. It is an essential part of the linguistic heritage of each speech community, which makes folklore a valuable source for idiomatic expressions.

The link between the tertium and the vehicle can be literal or metaphoric. The latter is often achieved by incongruency between the semantic domains of the co-occurring lexical items. For instance, the adjective *black* in (3a) refers to Willy's hair in the same conceptual dimension as to coal, i.e. the color, resulting in the literal reading. In contrast, the word *sober* in (3b) refers to Judy in the sense 'not drunk' but 'solemn' – to the vehicle JUDGE. Sobriety, being an inherit component to the concept of a judge, is transferred here metaphorically to the domain of alcohol intoxication, implying Judy's abstinence at the party. Especially, the shift from the domain of sense perception to the domain of emotions is registered as regularly metaphoric (cf. Kövecses 2000, Vejdemo and Vandewinkel 2016).

3. a. Willy has hair black as coal.

   b. Judy alone stayed sober as a judge at the party. (Norrick 1986: 43)

Yet another reason for the non-literal reading of the A *as* NP construction is salience imbalance, caused by a high degree of vagueness of the tertium or simply its redundancy. Consider the examples below. The word *right* in (4a) is too ambiguous to denote a salient property of the vehicle *rain*; in other words, the adjective here possesses little salience for the vehicle. In comparison, the adjective *mute* in (4b) possesses minimal salience for *stone* as the latter is not capable of speech per se. The meaning of the tertium is, therefore, redundant to the meaning of the construction, but, on the other hand, this is exactly what endows the A *as* NP construction with the function of intensification.

4. a. He was very ill, but he's right as rain now.

b. I was as mute as any stone, I had no word to say.

In some cases, however, it could be argued that the lack of a motivating figurative link between the lexical components of the construction is indicative of grammaticalization. For instance, while *hot as hell* can be considered semantically motivated due to the conventional representation of hell as a place set aflame, the meaning of *cold as hell* is not related to the concept of hell as such anymore. In fact, there are reportedly a number of structural frames, that appear to be undergoing the process of semantic bleaching from the lexical meaning to the grammatical meaning of emphasis and ultimate intensification. The patterns listed below are examples of grammaticalized similie patterns (originally reported in Moon 2008):

(as) ADJECTIVE as anything

(as) ADJECTIVE as hell

(as) ADJECTIVE as you please

According to Diessel (1999, 2019), one of the factors, triggering grammaticalization, is a recurrent high-frequency occurrence of specific content words in a certain linguistic environment, or a context. As a result, the lexical element becomes so closely interrelated with the construction it frequently co-occurs with that its lexical meaning shifts to the grammatical meaning of the whole construction. The lexical material, following *as* in the examples above, has little to do with its actual semantic meaning but rather serves to make an expression more emphatic. Another factor that sets grammaticalization in motion is the co-occurrence of these patterns with a wide variety of semantically different words (here, adjectives), which leads to the generalization over these lexical sequences to abstract structures. In Langacker's words, such conceptually distant from each other instances as *ugly/beautiful/hard/necessary/boring/cool/funny/stubborn/pale as fuck* suggest that the A *as* NP subschema ____ *as fuck* has undergone subjectification (Langacker 2006). In other words, *as fuck* can used by a speaker irrespective of the semantic compatibility between the concept *fuck* and the predicated property; rather, the vehicle, being completely stripped off of its lexical content, merely adds emphasis to whatever meaning is expressed by a selected adjective.

In fact, subjectification has indeed been claimed to be closely related to grammaticalization. More specifically, a content word (or a sequence of content words) is understood as subjectified when particular elements of its conceptualization are construed with greater subjectivity (Langacker 2006). A case on point are English modal verbs, which have undergone the shift from content to function items with modal force, construed to mark the speaker's perspective. The same principle can be applied to similies and, specifically, to the A *as* NP construction, whose meaning has shifted in a number of occasions to the intensifying interpretation. According to Parizoska and Filipović Petrović (2017), who found similar patterns among Croation similies, such expressions provide evidence of language change. Cross-linguistic research is suggested to reveal more to the effects of subjectification of similies on their structure and meaning.

4. A *as* NP Construction in English and Russian

Although the scarce discussion of similies and the A *as* NP construction, in particular, is primarily focused on English, we have seen that there are studies that report similar structural and semantic characteristics, as well development pattern in other languages. Russian appears to be no exception.

The structural frame of the Russian *A as NP* construction can be depicted as follows (compare with the English frame below):

Rus.: [ADJ *kak/slovno/(kak) budto/toĉno* NP]

Eng.: [(*as*) ADJ *as/like* NP]

The greater variability of the standard markers in Russian, compared to English, could tentatively be accounted for by a long tradition of using this construction in the Russian folklore. It is specifically reflected in the archaic nature of the words *slovno, (kak) budto* and *toĉno*, whose usage is traditionally confined to folk narratives. Alternatively, they are employed for stylistic effect to intentionally evoke associations with this genre. The neutral and most common standard marker in the Russian A *as* NP construction is represented by the adverb *kak*, the exact equivalent of English *as*.

One peculiarity that differs the Russian construction from its English counterpart is the adjective ellipsis in highly conventionalized idiomatic expressions, allowing the recovery of the tertium from the context. This is particularly the case for the objects that highly idiomatic expressions most commonly co-occur with, which makes the use of the tertium redundant and easily ommitable. For instance, the instance *bol'shije kak u lokatora* (lit. *big as the radar antenna*) is most frequently observed in reference to the size of the ears. Hence, the co-occurrence of both (*ushi (bol'shije) kak u lokatora,* lit. *the ears (as big) as the radar antenna*) does not require the use of the adjective as the inference about the common ground for the comparison can be derived from the context.

Being the source for numerous idiomatic expressions in poetry and fairy tales, the Russian A *as* NP construction is studied relatively extensively, as evidenced by a number of dictionaries, listing well attested instances. Despite a high degree of lexicalization, the

Russian construction has been reported to vary along the gradient of idiomaticity and metaphoricity (Lebedeva 2017: 6). The idea of lexical variability, allowed in certain A *as* NP subschemas, is also reflected in the structure of Mokienko's dictionary (Mokienko 2016), where the expressions are categorized by the lexemes, encountered in the ADJ and NP slots. For example, the entry for adjective *glupyj* (lit. *stupid*) comprises the following NPs that this adjective has been attested to occur with: *baran* (*sheep*), *brevno (log), gus'/gusynya (male/female goose), pen' (tree stub), poleno (firewood log), sivyj merin (gray gelding*).

From the semantic perspective, the Russian construction matches the general semantic blueprint of similies in that it is used in reference to cognition, emotions, character traits, appearance, lifestyle, different events, natural phenomena, space, time, etc. Lebedeva's dictionary carves up the semantic space of the Russian A *as* NP construction into 22 categories although Lebedeva herself admits the relativity of such a classification, especially given the fact that a substantial number of instances are too ambiguous to be assigned to one category.

Similarly to English, some types of the Russian construction expose a lose motivational link between the tertium and the vehicle of comparison, suggestive of potential semantic bleaching of the latter. As was discussed in the previous chapter, such structural formations consist of one stable component that attracts a variety of semantically different lexical items (e.g., *zloj/gryaznyj/golodnyi/strashnyi/skushnyj kak chort*; lit. *angry/dirty/hungry/scary/boring as the devil*). Just as it has been established for English and Croatian, the Russian A *as* NP construction also shows signs of subjectification (although to a somewhat lesser extent) in those subschemas that are associated with lexemes from different semantic fields.

Given the characteristics that have been addressed up to this point, it becomes clear that the A *as* NP construction exhibits a number of semantic as well structural constraints across a number of languages. These limitations have played a crucial role in the discussion on the constructional status of the A as NP pattern in some constructional approaches (see Desagulier 2016 and Pianykh 2019 for a review). In the frameworks of the Berkeley Construction Grammar, represented by Kay, despite "many members", the construction can be nothing but merely a "pattern of coining". Therefore, following Kay, it should be considered at the periphery of grammar as its productivity is severely restricted (Kay 2013). For instance, the NP in *easy as pie* cannot be substituted by other synonymous words (e.g., $^{?}$*easy as cupcake*) despite the theoretical possibility of such an extension. The semantically restricted nature of the pattern, as claimed by Kay, does not let the speaker freely extend the schema to create new instances insofar as they “come into existence every now and then as analogical creations […] but […] die aborning” (Kay 2013: 38).

As he elaborates further, there are several other idiosyncrasies that hinder the pattern from being fully productive. First, despite the abundance of various formulaic expressions, fitting the A *as* NP formula, they range from perfectly motivated (e.g., *smooth as silk*) to metaphorically extended instances (e.g., *thick as mince*), and yet there are tokens with no semantic link at all (e.g., *cool as a cucumber* or *happy as a boa* (from Russian *dovol’yj kak udav*)). Moreover, some of them can encode both literal and non-literal interpretations, each activated in a specific context (e.g., *the ground was cold as ice* vs. *the killer was cold as ice*). Some instances can go even further in that they show signs of semantic bleaching and grammaticalization, as pointed out in the preceding chapter. Another idiosyncrasy of the A *as* NP construction is its constrained behavior with the *than*-phrase. While some expressions can be freely used in the comparative form (e.g., *hotter than hell, darker than night, redder than a crayfish* (from Russian *krasnee raka*)), others cannot (e.g., $^{?}$*righter than rain,* $^{?}$*more solid than good,* $^{?}$*more squeezed than a lemon* (from Russian *vyžat’eje chem limon*)).

Although the aforementioned arguments about the idiosyncratic nature of A *as* NP are clearly incontestable, the claims about its productivity should not be confined to the highest level of schematicity and idiomaticity, as has been argued above. In fact, as supposed by Diessel (2019) and shown by Barðdal (2008), Zeldes (2012) and Desagulier (2016), although productivity corresponds to maximal schematicity, productivity processes of different magnitude could be observed at all levels of the construction taxonomy. Moreover, the existence of the two lexical slots in the A *as* NP construction implies the possibility of their interaction, which indicates, once again, that the binary approach to productivity, as proposed by Kay, needs a more solid empirical grounding.

Furthermore, as argued by Goldberg, a proponent of the redundant construction grammar framework, constructions are understood to vary in their degree of abstraction and complexity as they represent a dynamic network of linguistic knowledge (“It’s constructions all the way down” (Goldberg 2006: 18)). As a matter of fact, even such highly schematic constructions as the transitive, causative and attributive modification constructions have their circumscriptions. Following Goldberg’s argumentation, “each construction has a restricted range of distribution, typically dependent on various semantic, pragmatic, and phonological properties of the exemplars that are witnessed.” (Goldberg 2016: 386). Productivity, therefore, represents merely the (dynamic) ability of a construction to expand to novel items and its potential to produce innovative instances. It follows that productivity varies in coverage at different levels of a construction hierarchy: from the highest at the top level of abstraction, to the lowest at the level of lexically filled subschemas.

That being said, the question about the constructional status of a specific pattern is a sheer matter of theoretical preference while the actual behavior of the A *as* NP construction is far more complex to be restricted to the dichotomy between fully productive and non-productive structural configurations (cf. Pianykh 2019). Regardless of productivity, the complexity of the construction could be attributed to the intricate interplay of several factors such as structural limitations, idiosyncratic selectivity of lexical items in specific contexts (e.g., with the *than*-comparison), as well as interaction between the ADJ and NP slots. The latter is the primary focus of the current study, aimed at analyzing lexical association strength and its direction in the English and Russian A *as* NP construction based on the real world data.

5. Association Measures

5.1. Symmetrical Association Measures

There is a wide variety of lexical association measures employed in corpus studies, ranging from collostructionist analyses to the development of neural network based architectures that learn the positioning of words and extract word associations (Kapetanios 2018). With the advances in artificial intelligence and, more recently, Deep Learning, the identification and extraction of meaningful chunks of lexical data has found its place in natural language understanding and machine translations (Sag et al. 2002, Wehrli and Nerima 2018). While some studies focused on small co-occurrence contexts, such as ngrams, others made use of syntactic dependencies (Lin 1999, Seretan 2003) or POS-tagging (Evert and Krenn 2001). Among other factors that influence the choice of a relevant association measure (AM) are data quality, sample size, and, last but not least, the interpretation of the notion of collocations (see Evert et al. 2017 for more details).

Admittedly, some AMs have been used more extensively in certain linguistic branches (e.g. log-likelihood in computational linguistics or t-score and mutual information (MI) in computational lexicography). Having said that, however, different AMs highlight different aspects of collocativity and should be selected in accordance with a certain task. For instance, significance measures provide a different amount of evidence from the sample data against the null hypothesis of independence and, hence, show the probability of observing the given data in the real world. Effect-size measures, on the other hand, are intended to capture the strength of association between the items. The choice between the two groups, though, is "purely philosophical" and cannot be made solely on mathematical grounds. As Evert puts it, "a conclusive answer can therefore only come from a comparative empirical evaluation of association measures, which plugs different measures into the intended application" (Evert 2005: 113). In the similar vein, Pecina and Schlesinger, who reviewed over 80 AMs, conclude that the performance of various AMs "depends heavily on data, language and notion of collocation itself" (Pecina and Schlesinger 2006: 658). Comparing numerous association measures in an attempt to determine the best fit to the available data is beyond the scope of the current study, but this chapter is intended just to briefly outline the most frequently used metrics.

In a nutshell, as Seretan (2018) put it, every AMs used by corpus linguists computes collocation strength based on the following information (here for a bigram):

1) the frequency of co-occurrence of $word_1$ and $word_2$;
2) the frequency of occurrence of $word_1$ with the other elements but $word_2$ in the given dataset;
3) the frequency of occurrence of $word_2$ with the other elements but $word_1$;
4) the total size of the lexical items, or the sample size.

These scores are usually represented in the ubiquitous contingency table with the observed co-occurrence frequencies and marginal totals (Table 1).

	$word_2$: present	$word_2$: absent	Totals
$word_1$: present	a	b	a + b
$word_1$: absent	c	d	c + d
Totals	a + c	b + d	a + b + c + d

Table 1. The co-occurrence frequency table submitted to the most lexical association analyses.

The major shortcoming of the most association metrics, however, has been claimed to lie in their unreliability when applied to data with a great number of rare events (Pecina 2010: 26). For instances, the chi-squared and z-score tests, as well as information-theoretic mutual information (MI), have been denounced for their low-frequency bias that leads to overestimation of rare events and inflation of test scores (Evert 2009). Another point of criticism is violation of the normality assumption, underlying the z-score measure and the t-test, which is often the case for large samples. Following Church and Mercer (1993), normally distributed data can rarely be obtained in language use. The odds ratio has been proved relatively reliable from the mathematical point of view (Agresti 2002: 71) but its interpretation is far from intuitive. Since the present study does not pursue the goal of comparing different AMs, the following is the outline of the two measures employed in this particular study. For an extensive overview of various AMs in the domains of both lexical and lexicogrammatical co-occurrence see Evert 2005, Wiechmann 2008, Pecina 2010, or Hoang at al. 2009.

The recently popularized family of collostructional methods, developed by Gries and Stefanowitsch (Stefanowitsch and Gries 2003, Gries and Stefanowitsch 2004a, Gries and Stefanowitsch 2004b, Stefanowitsch and Gries 2005), makes extensive use of the Fisher's exact test – a non-parametric counterpart of the standard t-test – to gauge the association strength between elements and/or constructions. It is based on the comparison of observed and expected frequencies in the form of a 2-by-2 contingency table and is known to produce more robust results given low expected frequencies. The significance of this difference is reflected in the p-values, whose $\log_{10}$-transformation represents the strength of association. To put it another way, the higher the log score, the smaller the p-value and the smaller the chance of observing the co-occurring items independently.

Another well-established and widely-used AM is the log-likelihood ratio (LLR) (e.g., Dunning 1993), which has also found its application as the default association metric in the web-based interface to the British National Corpus (BNCweb) (Hoffmann and Evert 2006). It is known for its well-understood mathematical properties (Manning and Schütze 1999) and has been claimed "probably the best or second-best measure on mathematical grounds" (Gries 2013: 148). It represents a great approximation to the association scores, produced by the Fisher's exact test (Evert 2005), does not depend on the assumption of normality and allows for rare events. The latter is crucial when applied to natural language processing since, as stated by the Zipf's law, most distinct words in a given text will occur only a small number of times, if not once, no matter how large the sample is. The problem of rare events, also known as hapax-legomena if they occur in the sample just once (Baayen 1993), arises inevitably whenever we deal with individual words. On the other hand, the LLR does not exhibit oversensitivity to low frequency co-occurrences, a problem characteristic of other AMs (Daudaravičius and Marcinkevičienė 2004). Similarly to the Fisher's exact test, the LLR derives its ranking from the sampling distribution: the smaller the probability of a sample outcome under the null hypothesis, the more "surprising" it is and the more evidence against H_0 it provides (Evert 2005, Pecina 2010).

Figure 1 represents the formula for calculating the two-sided LLR without the distinction between $O > E$ and $O < E$ that assigns high values in both cases. However, multiplication of the association scores for all ngrams with $O < E$ by -1 solves this issue by introducing the explicit difference between attracted and repelled items. The significance of the score is nevertheless reflected by its absolute value. Since the LLR test statistic (known as

G^2, by analogy to χ^2) has an asymptotic chi-squared distribution with one degree of freedom (χ^2_1 for short), the cut-off threshold for the significance level of $\alpha = .05$ is $G^2 = 3.84$.

$$\log - likelihood = 2\sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

Figure 1: The two-sided log-likelihood association measure where O stands for observed frequencies and E - for expected frequencies.

In regard to the current study, both the Fisher’s exact test and the LLR are considered good candidates for measuring the lexical association strength between the slots of the A as NP construction on the following grounds:

1) given the relatively small size of the datasets (especially in Russian), the sampling distribution is not expected to conform to the assumption of normality;
2) the fair treatment and the precise evaluation of low-frequency co-occurences without inflating their association scores;
3) robust theoretical mathematical grounds that make both tests a good reference point in a general discussion on collocational strength;
4) relatively simple computation and uncomplicated interpretation of association scores.

The results of these tests will then be submitted to a correlation test to compare their performance in regard to the given English and Russian sample data.

5.2. Asymmetrical Association Measures

Human word associations have been shown to be largely asymmetric. This means that speakers have a strong intuition that one element of a collocation to a certain degree governs the choice of the other. Yet, the overwhelming majority of the AMs are traditionally symmetric, assuming that the association between word_1 and word_2 in a bigram, for example, is bidirectional. Indeed, word_1 does not necessarily evoke word_2 to the same extent as word_2 evokes word_1, meaning that one of the items is generally more likely to be a better cue for the other (Hausmann, 1989, Benson 1989, Mel’cuk 1998, Gries 2013, Desagulier 2016, inter alia). For instance, the word *ipso* necessarily evokes *facto* while *facto* alone could be a cue either for *de* or *post* (Desagulier 2016: 11). In the similar vein, Kjellmer (1991) distinguishes between 1) right and left predictive collocations (i.e. fully symmetric associations), 2) right predictive (e.g., *wellington boots*) and 3) left predictive collocations (e.g., *deadly nightshade*). The latter two are also frequently referred to as forward and backward associations (e.g., Michelbacher et al. 2007, Michelbacher et al. 2011). It is known that the asymmetrical nature of word associations is motivated by the mechanisms of human cognition, as evidenced by the results of free association tasks (see, for example, the University of South Florida Association Norms (Nelson et al. 2004)).

Despite the clear indications of directionality of lexical associations and a large variety of AMs used in collocation studies, they still do not distinguish between forward and backward associations. To put it another way, they are by and large symmetric even in the case of such theoretically and mathematically powerful metrics as $p_{\text{Fisher-Yates exact test}}$ or G^2: they do not allow differentiation in the direction of predictability. There are just a few directional AMs that have been recently suggested, and almost all of them were proved somewhat problematic (cf. Gries 2013, Wahl 2015 and Schneider 2018). Take, for example, the simple transitional probability in Figure 2a that is calculated as the probability of one word given another (cf. Gregory et al., 1999, Saffran et al. 1996 and Thiessen 2013). Yet what it does not take into account is the interfering associations that involve all other stimuli but the cue. Another measure, explored particularly in Michelbacher et al. 2007 and Michelbacher et al. 2011, is the conditional probability (Figure 2b). It has yielded some moderately satisfying results in their studies on detecting noun-adjective collocations in the British National Corpus but still exhibited a high error rate.

(a) $$fwdTP(word_2|word_1)=\frac{freq(a)}{freq(word_1)}$$ (b) $$p(word_2|word_1)=\frac{a}{a+b}$$

$$bckTP(word_1|word_2)=\frac{freq(a)}{freq(word_2)}$$ $$p(word_1|word_2)=\frac{a}{a+c}$$

Figure 2. The formulae for computing the transitional (a) and conditional (b) probabilities. The variables for these calculations are derived from Table 1.

Lastly, Gries (2013) proposed the ΔP measure (Figure 3) that overcomes the shortcomings of the previous asymmetric AMs and, despite its relative recency, has immediately gained in popularity in a wide range of applications (e.g., in Levshina 2015, Wahl 2015, Desagulier 2016, Seretan 2018, Schneider 2018, Todd 2019, Garcia et al. 2019, inter alia). This simple directional association measure was reintroduced by Gries to the constructionist field but, in fact, it emanated from the domain of associative learning and judgement tasks (Jenkins and Ward 1965, Ward and Jenkins 1965, Rescorla 1968, Allan 1980). It was originally used as a statistical measure in behavioral tasks to model directional associative learning from a to-be-conditioned stimulus (CS), temporally paired with an unconditioned stimulus (US), to a conditioned response (CR). As a result, it was concluded that it is contingency, not temporal pairing of the CS and the CR, that generated conditioned responding. It was then adopted in the sphere of language acquisition (MacWhinney et al. 1989, Ellis 2006, Ellis and Ferreira-

Junior 2009) to demonstrate that “speakers tend to infer a linguistic outcome from cues available in their linguistic environment in an asymmetric fashion” (Desagulier 2016). Before it was brought to a broader attention and applied in collostructionist studies by Stefanowitsch and Gries (2013), ΔP was also used as a one-way association measure in modeling L1 and L2 acquisition.

As can be seen from Figure 3, the ΔP values are also based on the co-occurrence frequencies from Table 1. The $\Delta P_{2|1}$ score in (1) computes the probability of observing word$_2$ given word$_1$, and the other way round for $\Delta P_{1|2}$ in (2). The results of these metrics are quite easy to interpret as they range from -1 (repulsion) to 1 (attraction). The ΔP in (3) is calculated as a difference between the $\Delta P_{2|1}$ and $\Delta P_{1|2}$ and represents the overall direction of association in the whole collocation. Negative ΔP suggests that word$_2$ is more informative of word$_1$ than vice versa whereas a positive value is indicative of higher predicative power of word$_1$ than word$_2$. It is important to note, however, that in case both lexical items are repulsed from each other, i.e. $\Delta P_{2|1}$ as well as $\Delta P_{1|2}$ scores are negative, they are withdrawn from the further analysis of association directionality.

$$(1)\quad \Delta P_2|_1 = p(word_2|\, word_1 = present) - p(word_2|word_1 = absent) = \frac{a}{a+b} - \frac{c}{c+d} = \frac{(ad-bc)}{(a+b)(c+d)}$$

$$(2)\quad \Delta P_1|_2 = p(word_1|\, word_2 = present) - p(word_1|word_2 = absent) = \frac{a}{a+c} - \frac{b}{b+d} = \frac{(ad-bc)}{(a+c)(b+d)}$$

$$(3)\quad \Delta P = \Delta P_2|_1 - \Delta P_1|_2$$

Figure 3. The right- and left-predictive ΔP values, based on the observed frequencies of the contingency table. The ΔP score in (3) is meant to capture the direction of association of the whole collocation.

To sum up, the ΔP measure exhibits a great potential in corpus linguistics to gauge unidirectional lexical association not only due to its computational ease but also because of the existing evidence that it reflects psychological and psycholinguistic reality (Gries 2013: 143–144).

Essentially, the present study builds on the methodology employed in Pianykh 2019 with the idea of gaining more insights into the nature of the A *as* NP construction in English and Russian. To achieve this purpose, the following objectives were set:

- investigate the lexical association between the slots and establish the characteristics of the construction instances along the continuum of collocational strength;
- identify A *as* NP subschemas for each language drawing on the concept of asymmetric association between collocates;

- analyze the construction observations in both languages for the signs of grammaticalization;
- inspect the relationship between symmetric and asymmetric association measures for each of the given data samples;
- give the construction a general semantic characteristic at the different ends of lexical association;
- delineate the most conspicuous semantic patterns;
- ascertain whether the semantic characteristics of the slots impact the collocational strength of the whole construction and the direction of association in a given pair of lexical items.

It is of note that the last point refers in the current study primarily to the English data as it provides a more comprehensive and reliable basis for an analysis, comprising a large number of variables and rare events.

6. Methods

6.1. Data

The data for the English sample come from the Corpus of Global Web-Based English (GloWbE) (Davies 2013), which contains 1.9 billion words from twenty different English varieties. Around 60% of the corpus is constituted by informal blogs while the rest of the data encompass a wide range of other genres. The Timestamped JSI web corpus 2014-2016 Russian with over 1 billion tokens (Bušta and Herman 2017) was chosen as a comparable data source for the Russian sample, similarly to the study of Pianykh 2019. The decisive factor in the selection of the corpora was a high degree of the present-day language representation, reflected in the way these two corpora were created. Both were compiled by extracting essentially random Google web pages for the GloWbE and RSS-enabled newsfeeds for the Timestamped corpus.

6.2. Data Extraction and Filtering

The text tokenization and POS-tagging of the Timestamped corpus, supplemented with the Corpus Query Language (CQL) as a specialized query engine, considerably simplified the search and extraction procedure for assembling the preliminary sample of the Russian instances of the A *as* NP construction. In order to maximize both precision and recall, one needs a highly accurate search query. The query below (5) not only finds sentences, matching the structural frame [ADJ *as* NP] but also filters out a number of unwanted lexical items, frequently occurring in similar expressions (e.g., *известный как, расценен как, популярен* *как*, etc., in English: *known as, considered as, popular as*). As can be seen from the query, the standard marker *kak* represents the core of the searched construction and, unlike the ADJ and the NP, is left without the tag specification. The main reason behind this decision is low accuracy of POS taggers in regard to function words, in general, as these items are highly context-dependent and grammatically versatile. The selected NP tags are meant to capture the whole range of grammatical variations, typical of Russian, including gender (feminine, masculine, neuter), number (Sg, Pl), case (here Nominative) and animacy (animate, inanimate). In spite of the detailed query, the search resulted in 7 749 observations in Russian, reduced to 2 586 instances after the thorough manual post-editing.

(5) [tag="A.*" & lemma!="известный|известен|известны|расценен|популярен|должен|нужен|необходима|скорее|причастны"] [lemma="как"] [tag="Ncmsnn.*" | tag="Ncmsny.*" | tag="Ncfsnn.*" | tag="Ncfsny.*" | tag="Ncnsnn.*" | tag="Ncnsny.*" | tag="Nccsny.*"]

The extraction and filtering of the English data, in contrast, required a more elaborate procedure. First, all the sentences, including the word *as*, were retrieved and annotated with the use of the Stanza package for the natural language processing, developed by the Stanford NLP Group (Qi et al. 2020). Apart from many other efficient tools, the package provides a variety of Universal Dependencies (UD) models, trained on the UD treebanks and available for 66 human languages. One of the shortcomings of the package is the fact that processing of large-sized text files is internally memory-expensive and, for this reason, requires running the neural pipeline on batches of documents in order to optimize speed performance. The first step was then to determine the optimal size of a data piece, susceptible of successful parsing and part-of-speech (POS) annotation. The file with the 8 684 000 retrieved sentences was then split into 4 342 text documents with 2 000 lines each. A Python script iterated through the text files, whose content was submitted to the pipeline with the specified processors, performing lemmatization, tokenization and POS tagging.

The resulted text object was filtered, based on structural and lexical criteria in the following way. On the one hand, the target instances had to conform to the [ADJ *as* … NP] pattern where the dots between ADJ and NP represent a potential occurrence of other items, especially the articles. On the other hand, a stoplist of specific lexical items filtered out unwanted instances, frequently found in the target construction (e.g., *same as, as well as, as much as, such as, famous as, necessary as, considered as, known as, as little as*, etc.), which also had the purpose of minimizing the post hoc manual filtering. Among other unwanted expressions were conventional temporal phrases (e.g., *as long as, as young as*), and patterns with indefinite pronouns, mistagged by the Stanford NLP processor (e.g., *as good as everyone else, as true as everything else (in his life), so irritating as somebody (with less intelligence and more sense than we have*)).

The follow-up manual inspection, which eventually decreased the size of both Russian and English samples to the total of 2 586 and 12 253 observations, respectively, had the main goal of contextual filtering of the preprocessed text output. All the co-occurring items were stripped of their contexts and structural variation elements. Such standartization enabled categorization of the instances such as *dark as night* together with *dark as the night*. The exceptions were the contents with the function to extend the ADJ or the NP. The longest NP in English includes 10 words (6a) while in Russian – 5 words (6b).

(6) *a. clear as a sharpened diamond cutting through a clear piece of glass*

b. прекрасный как бутон розы сверкающий каплями росы

Eng.: *beautiful as the rose bud shining with the morning dew*

Since Russian, being a highly synthetic language, is well-known for its abundance of inflectional morphemes to express syntactic relationships, the Russian adjectives in the raw data displayed a substantial degree of structural variability, depending on gender, case, number, and form, alternating between full and short. By convention, however, the base, or dictionary, form in such cases is normally full, nominative, masculine and singular so the raw Russian adjectives were transformed accordingly.

In the English sample, the plural NPs were likewise singularized by means of the web mining module Pattern for Python (De Smedt and Daelemans 2012) in order to avoid treating different number forms of one and the same lexeme in the semantic analysis as separate entities (e.g., *poor as a church mouse* and *poor as church mice*). The same goal was kept in mind when recovering the full forms of all the instances of scripted foul language, censored by a wordfilter implemented on Internet forums and chat rooms. One such representative example is the word *fuck*, which is partially substituted by content-control software with grawlix nonsense characters (e.g., *f@ck*) or asterisks (e.g., *f****). Yet another censoring strategy, frequently encountered in the sample and often employed by Internet users themselves, is the use of the so-called minced oaths, or "a form of euphemism for words or phrases that are considered profanity, swearing, or taking God's name in vain, such as *dang*, in place of *damn*, or *heck* in place of *hell*" (Hudson 2016). The last step in the English data normalization was removing the determiner *all*, functioning to modify the NP and intensify the meaning, conveyed by the ADJ. In the overwhelming majority of cases *all* co-occurs together with *hell* and *heck* (as in *annoying as all hell*, *intimidating as all heck*).

6.3. Coding System

The coding system in the current study resembles the one, employed in Pianykh 2019. Once the data were collected and filtered, all the observations had to be coded in line with the following criteria: 1) frequency of occurrence for each ADJ_NP unique pair, as well as for each lexical item separately, 2) association strength between the two lexical slots (i.e. symmetrical association) and 3) the direction and scale of this association (i.e. asymmetrical association, or ΔP). The first two measures were obtained in the covarying-collexeme analysis using the interactive program Coll.analysis 3.2 the R script, provided by Stefan Th. Gries (2007) and freely available online. The data, submitted to this analysis, had to be transformed into a data frame with two columns, each for the adjectives and the noun phrases, co-occurring together. The output listed the unique ADJ_NP values, or ADJ_NP types, with their raw frequencies in the presence of each other and separately. To measure asymmetric association between the slots, we computed ΔP for each ADJ_NP unique instance.

The final step in the data coding procedure was semantic annotation of the ADJ and NP types. The data for this purpose was derived from the Concepticon database (List et al. 2020), which represents a large repository of concept lists with detailed concept characteristics and structured relations. Accessing and manipulating the Concepticon data is supported by the Pyconcepticon API tool, and the detailed user guide on its applications can be found in List 2018. The adjectives and noun phrases in the Russian sample were first translated into English with the help of the Google Translate API for Python and then likewise submitted to the semantic annotation procedure. The results are represented in Figure 4.

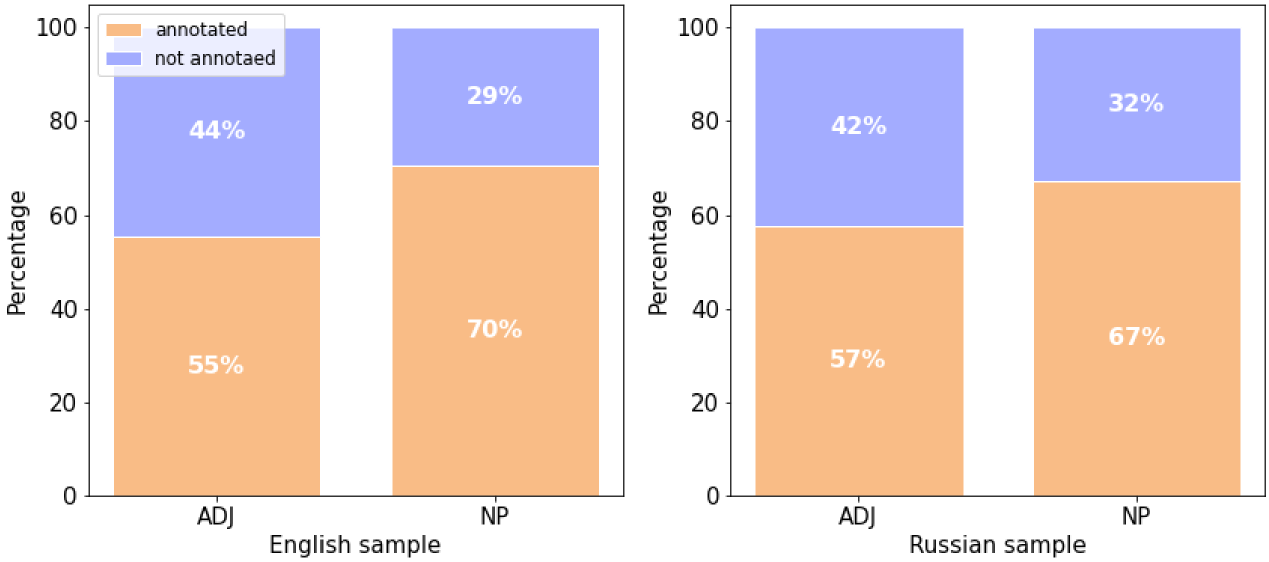


Figure 4. The results of the semantic annotation of the lexical items in the ADJ and NP slots in English and Russian.

Apparently, there are more matches for the elements in the NP slot, which is due to the drastic discrepancy between the number of substantive (2 244) and attributive concepts (323) in the Concepticon repository, available at the date of conducting the present study. It is for this reason, therefore, that only every second adjective in both samples is assigned a semantic tag whereas the successful annotation of the two thirds of the noun phrases provides a more solid ground for investigating the semantic relationships in the data.

In what follows, we will address the issues of symmetric and asymmetric associations between the slots of the A *as* NP construction in English and Russian by investigating the results of the covarying-collexeme analysis and the distribution of the computed ΔP values across the observations. In addition, the semantic analysis is assumed to unveil some correspondences between specific semantic aspects of the slots and their qunatitative characteristics..

7. Results and Discussion

7.1. Symmetric Association

As mentioned in Chapter 5.1., the data in the current study are subjected to two bidirectional AMs: the -$\log_{10}$ p-values of the Fisher's exact test and the LLR. The scores for the former were obtained in the covarying-collexeme analysis, performed on the datasets of 12 253 English and 2 586 Russian instances. The output of these computations entails a number of estimates. First of all, it is essential to note that the table, produced as a result, represents the data transformed into types. In other words, it lists unique instances of the construction with their respective frequencies. The following is the most informative measures in the frameworks of the current study:

1) the observed frequencies of each ADJ_NP co-occurrence, represented by *a* in Table 1;
2) the observed and expected frequencies for each attested instance of the construction, reflected by the marginal totals *a* + *b* and *a* + *c*;
3) attraction/repulsion relations, built upon the difference between the previous measures. Attracted ADJs and NPs are characterized by lower observed frequencies, compared to expected frequencies, and the other way round for repelled items.
4) collostructional strength, expressed by $p_{\text{Fisher-Yates exact test}}$. Since most p-values fluctuate between 1 and infinitely small scores, their transformation by means of the negative $\log_{10}$ resulted in a range of values from 0 to +*Inf*. As was pointed out in Chapter 5.1., collocation strength scores do not distinguish between attracted and repelled elements but merely show how unlikely it is to obtain the observed frequencies in a given

sample (irrespective of the direction of such a difference between observed and expected frequencies). To establish a threshold between the relations of attraction and repulsion and make this distinction explicit, the collocational strength scores for the repelled ADJ_NP pairs were multiplied by -1.

The LLR was computed using the R Rling package (Levshina 2014). Just as with the collostructional strength in (4) above, the LLR scores were, likewise, multiplied by -1 to reflect the relationship of repulsion if expected frequencies of the lexical items exceeded their actual frequencies. The results for the top observations in each language, ranked by the AMs, are listed in Table 2 for English and in Table 3 – for Russian. Note that the ADJ_NP pairs are singularized and stripped of their modifiers as pointed out in Chapter 6.2.

A as NP	Fish. test	LLR
Clear_day	Inf	1570
White_snow	Inf	1564
*Cheap_chip**	255	1175
*Easy_pie**	250	1146
Tough_nail	235	1078
*Clean_whistle**	206	944
Safe_house	169	779
*Pleased_punch**	167	769
Sweet_honey	166	762
Solid_rock	160	734
*Cool_cucumber**	156	716

Table 2. Top 11 A *as* NP instances in English, ranked by collocational strength.

A as NP		Fish. test	LLR
Russian	English transl.		
Staryj_mir	*Old _world*	Inf	1650
*Dovol'nyj_slon**	*Happy_elephant*	192	882
Neobhodimyj_vozduh	*Necessary_air*	92	418
Nuzhnyj_vozduh	*Essential_air*	90	409
Chyornyj_smol'	*Black_pitch*	84	385
Belyj_sneg	*White_snow*	83	381
Vyžatyj_limon	*Squeezed_lemon*	60	282
Holodnyj_lyod	*Cold_ice*	59	268
Ostryj_britva	*Sharp_razor*	42	193
Tvyordyj_kamen'	*Hard_rock*	40	183

Table 3. Top 10 A *as* NP instances in Russian, ranked by collocational strength and translated into English.

The observations marked with * have culture-specific meanings.

To begin with, there are two conspicuous observations that immediately catch the eye. Firstly, the scores of both the Fisher's exact test and the LLR appear to align perfectly, at least for the top associated items. A superficial exploratory analysis suggests a correlation between the two AMs along the whole range of the observations but, in order to make sure this correlation holds true for the entire data, the data sets were submitted to a one-tailed Kendall's rank correlation test. The reason underlying the choice of the non-parametric version of the classic Person's correlation test is heteroscedasticity of the data and a range of tied observations. Having said that, the linear and monotonic relationship between the p-values of the Fisher's exact test and the LLR scores is in line with the assumptions of the Kendall's test. The correlation between the variables is positive, extremely strong and statistically significant, τ = .86, df = 3181, $p_{\text{one-tailed}}$ < 2.2e-16 (the Russian sample demonstrated the comparable results: τ = .92, df = 936, $p_{\text{one-tailed}}$ < 2.2e-16). Moreover, they appear to rank the lexical association of the ADJ_NP pairs essentially in the same manner. These results demonstrate the comparability of both symmetrical AMs for the data with the given characteristics and appear to be in line with the findings reported by Moore (2004: 338). That being said, the G^2 scores will be used further in the current study as the reference for bidirectional lexical association between the slots solely because their computation did not produce infinity values, which would subsequently result in certain data loss.

The second prominent observation, following from Table 2 and Table 3, is the prima facie discrepancy between the association scores for the English and Russian observations that makes it tempting to compare them across the datasets. This would, however, be a misleading strategy since the difference in the sample sizes does not allow of a straightforward comparison. The $p_{\text{Fisher-Yates exact test}}$, returned by the Fisher's exact test, is known to depend heavily on a sample size: "the bigger the data set (corpus), the smaller the p-value [and the higher the collocational strength], even if the raw proportions of co-occurrences are the same" (Murmann 2019: 81-82). Since the p-values are commonly log-transformed in collostructionist studies (Levshina 2015: 232), a larger corpus leads to larger log-transformed scores, respectively. In the similar vein, the G^2 test statistic increase linearly with the size of the corpus (Moore 2004: 337) as it derives its values from the sampling distribution, which is an approximation of an asymptotic X^2 distribution. To sum up, the estimates, yielded by both AMs for the English and Russian datasets, will have to be examined separately although the semantic comparison of the most strongly associated elements between the samples is still possible.

As can be seen from the values in the tables above, the two top instances of A *as* NP in English and the top observation in Russian appear in the data so much more frequently than expected that their log-transformed $p_{\text{Fisher-Yates exact test}}$ values approach infinity. There are only two clear semantic correspondences between the two tables (*white_snow* and *solid/hard_rock*) but a further semantic analysis might reveal many more down the scale. Interestingly, a half of the most associated ADJ_NP pairs in English are culture-specific. In other words, a non-native speaker of English would be unable to infer the pragmatic meaning of these expressions as they are based on specific sociocultural experiences that originated in the British (e.g., *cheap as chips, pleased as punch, cool as a cucumber*), American (e.g., *clean as a whistle*) or even New Zealand (e.g., *easy as a pie*) culture. The expression *cheap as chips,* for instances, is framed by the conceptual system of the British gastronomy, where chips are such a ubiquitous and affordable meal that it has become a synonym of any low-priced offer, in general. The other half of the most associated items in English has the meaning of intensification of the most salient property of the NP. The link between the tertium and the vehicle in such instances is not arbitrary but semantically motivated owing to the semantic compatibility of the given ADJ and the NP. In comparison, the Russian subset in Table 3 is drastically different in that only one A *as* NP instance has a cultural background. The seemingly arbitrary choice of the collocates in *dovol'yj kak slon* (Eng.: *happy as an elephant*) has, in fact, its roots in the folklore Russian tradition. There are several theories about the etymology of this expression but the most common ones claim that it is either a shortened version of a Russian saying about being happy as an elephant after bathing in a pond or from one of the fables by Ivan Krylov, known for his extensive use of animal imagery in allegorical portrayals of basic human behavior. In fact, as will be seen further, the phenomenon of anthropomorphism is no exception in the given English and Russian corpora.

Following Pianykh 2019, the A *as* NP instances with high association scores can be considered lexical prefabs, or idioms, from the cognitive point of view. The observations in Table 2 and Table 3 are, therefore, accessed and activated in memory as indivisible and non-compositional lexicalized chunks. Moreover, some of these instances could be deemed unproductive as their components co-occur (almost) exclusively with each other (e.g., s*afe as houses*, *pleases as punch*, *cool as a cucumber*, Russ. *vyžatyj kak limon* (Eng.: *squeezed as a lemon*)).

To examine the opposite end of the association strength continuum, I first eliminated the observations with the collocational strength scores, for which the H_0 of no association cannot be rejected. Since the cut-off point at the significance level of 0.05 is $G^2 = 3.84$, by analogy with χ^2_1 (for the log-transformed $p_{\text{Fisher-Yates exact test}}$ the threshold equals approximately 1.3), the observations with absolute values $G^2 < 3.84$ had to be excluded from consideration. This measure resulted in the loss of 23% of the English data and 11% of the Russian data. The latter additionally lost its only 7 observations, characterized by the relation of repulsion, resulting into a heavily asymmetric data set with exclusively attracted items. The repelled observations in English were kept to prevent massive data loss and gain some insights into the lexical variability of the A *as* NP construction.

The analysis of the least associated (or even repelled) items included the 50 bottom observations, ranked by collocational strength. Characteristically, the number of NP types (n = 7) is six times the number of ADJ types (n = 43) that include a wide range of low-frequency items. The overwhelming majority of the ADJ_NP instances are indexed on the NP *hell*. The relationship between the adjectives and the NPs in this subset can be summed up with the help of the type-token ratio (TTR), aimed at evaluating lexical richness of a specific grammatical category: the $\mathrm{TTR}_{\mathrm{ADJ}}$ is 0.86 while the $\mathrm{TTR}_{\mathrm{NP}}$ equals just 0.14, accounted for by a considerably lower NP type count (Figure 5). In contrast, the 50 bottom observations in Russian, ranked by association strength, demonstrate a completely different pattern.

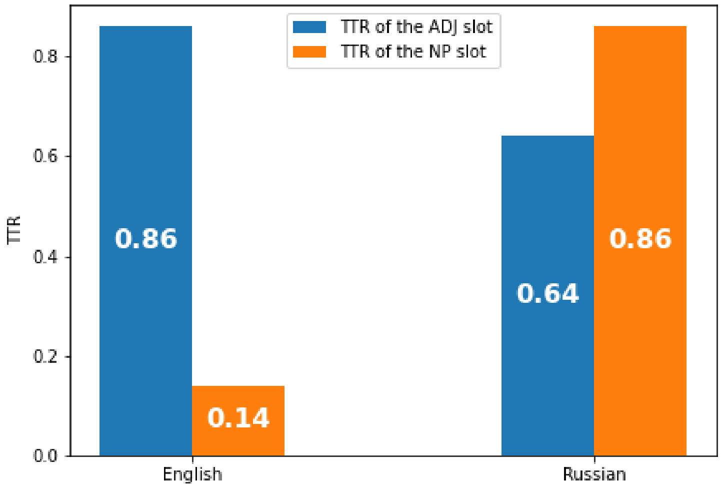


Figure 5. TTR of the ADJ and NP slots in English and Russian for the 50 least associated observations.

Unlike in English, the Russian ADJ and NP types in this subset are distributed far more evenly; in fact, they do not reveal such a drastic asymmetry in the lexical richness of the slots ($\mathrm{TTR}_{\mathrm{ADJ}}$ = 0.64 and $\mathrm{TTR}_{\mathrm{NP}}$ = 0.86). Interestingly, the NP slot in Russian is more lexically versatile than that of the NP, which stands in sharp contrast to English. The Russian data, therefore, appear more lexically diverse and the slots of the Russian A *as* NP construction – more lexically saturated. Any conclusions at this point, however, should rather be made with extreme caution as lexical richness is known to vary with increased corpus size (cf. Koplenig 2018). It is highly likely that the quantitative relationship between the slots in the Russian sample would change to a great extent, be the sample the size of the given English data set.

From what has been shown in this section, it is clear that the symmetric association scores reflect the hierarchy, or better, the gradient of lexical sequences, where the highest values reflect the highest level of automatization and conventionalization and the lowest values characterize the ADJ_NP pairs, allowing certain flexibility in the choice of lexical items. The lexically filled constructional frame such as those in Table 2 and Table 3 are located at the lowest level of abstraction of the A *as* NP construction; they are represented and activated in the mental lexicon as indivisible chunks. The higher up the abstraction scale, the more schematic a construction becomes. As stated by Barðdal, bearing on Clausner and Croft

(1997), productivity corresponds to maximal schematically: “a construction is more productive the higher its level of schematicity, which corresponds to more types being generated regularly and coherently (i.e. with consistent meaning) from the schema” (Barðdal 2008: 50).

It is for this reason that the observations down the scale of association strength are of most interest as they get more abstract, compared to the lexically predetermined slots of the idioms at the lowest level of abstraction. It is in this range of collocational strength that ADJ_NP pairs form partially-filled subschemas that are indexed on a specific lexical item (e.g., ADJ-subschemas as in (7) or NP-schemas as in (8)).

(7) Rus.: staryj kak ____ (drevnost’/brak/chelovechestvo/snouboard/žyzn’/mir/)
Eng.: old as ____ (antiquity/marriage/humanity/a snowboard/life/the world)

(8) Rus.: (chistyj/dovolnyj/iskrennij/scchastlivyj) ____ kak rebenok
Eng.: (pure/satisfied/sincere/happy) ____ as a child

To investigate the productivity of such ADJ- and NP-schemas in both English and Russian, we now turn to the analysis of asymmetric dependency between the two lexical slots.

7.2. Asymmetric Association

The direction of association was determined by computing $\Delta P_{forward}$ ($\Delta P_{NP|A}$), $\Delta P_{backward}$ ($\Delta P_{A|NP}$) and ΔP_{diff} for the whole construction ($\Delta P_{forward} - \Delta p_{backward}$), as shown in Chapter 5.2., for each ADJ_NP type. The last measure did only make sense when both slots are attracted to each other, i.e. both $\Delta P_{forward}$ and $\Delta P_{backward}$ are positive. The difference between them then is intended to reveal the overall direction of prediction in a given A *as* NP instance. On the contrary, if ΔP_{diff} values are negative, pointing to the relation of repulsion between the slots, a further analysis of the direction of association is not conceptually sensible. The subtraction of negative $\Delta P_{forward}$ and $\Delta P_{backward}$ values in this case leads to inflation of the ΔP_{diff} score, which is misleading as it cannot be used for investigating asymmetries between repulsed items. Therefore, the ADJ_NP pairs with their observed frequency of co-occurrence below the expected frequency were excluded from the further analysis of asymmetric lexical association, resulting into additional 0.4% of data loss in the English sample. In contrast, the Russian data set did not suffer any changes as the only 7 Russian observations, characterized by the relation of repulsion, were already removed previously.

The distribution of the overall ΔP_{diff} values across the English sample, sorted in ascending order, is illustrated in Figure 6. The plot demonstrates the heavily skewed data and thereby confirms the assumption that association between the ADJ and the NP slots is directional. As can be seen, in the vast majority of cases it is the existence of the NP that increases the likelihood of a lexical item in the ADJ slot as evidenced by the 67% of the observations, characterized by negative ΔP_{diff}. In contrast, the ADJ is a better predictor in the whole A *as* NP instance in just 29% of cases. The discrepancy between the predictive power scores of the ADJ and the NP is also confirmed by the measures of central tendency: the $mean_{\Delta P}$ in the sample equals -0.32 and the $median_{\Delta P}$ is -0.48. For the rest 4% of the data, ΔP_{diff} = 0, which means that $\Delta P_{forward} = \Delta P_{backward}$: the likelihood of ADJ given the NP is identical to the likelihood of the NP given the ADJ. To put it differently, the observations, located at y = 0 in Figure 6, are associated with each other in a perfectly symmetrical fashion.

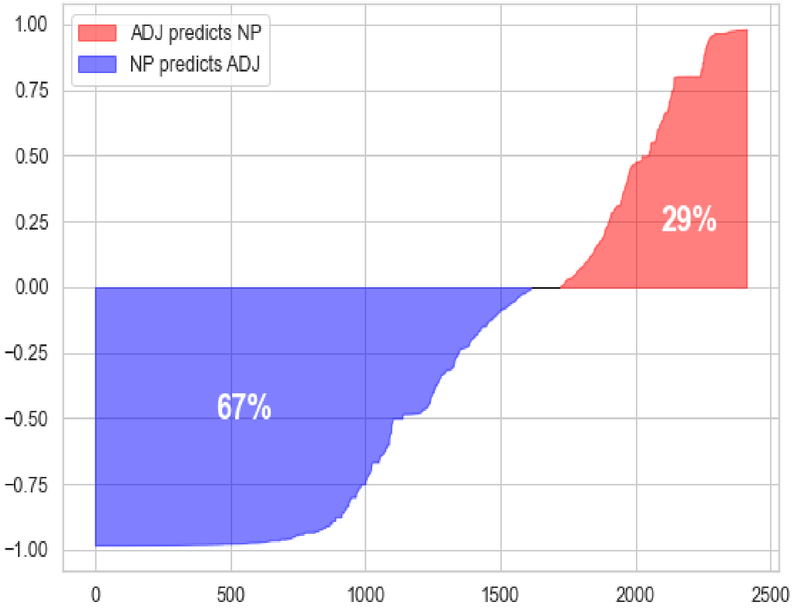


Figure 6. Distribution of ΔP_{diff} across the ADJ_NP types in the English sample, sorted in ascending order.

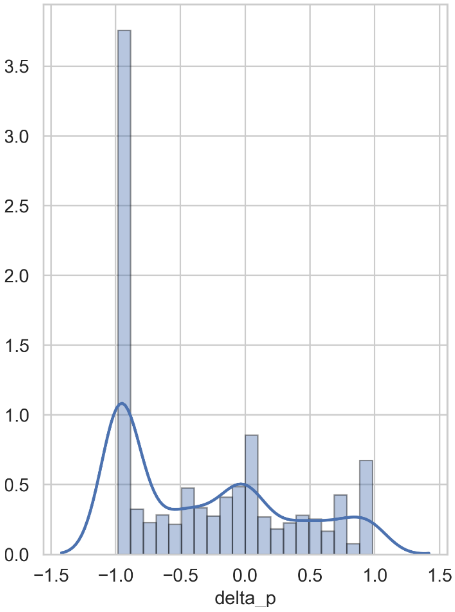


Figure 7. Density of ΔP_{diff} distribution in the English sample.

To investigate the meaningfulness of the discrepancy between the predictive power values of the ADJ and NP slots, their respective $\Delta P_{forward}$ and $\Delta P_{backward}$ scores were submitted to the one-tailed Wilcoxon test for independent samples. The non-parametric t-test was chosen based on a number of factors: 1) the difference in variance of the two variables and 2) non-normal distribution of the data, confirmed by the extremely small p-values of the Shapiro–Wilk test in each case (p < 2.2e-16). On average, the ability of the ADJ slot to increase the likelihood of lexical contents in the NP slot (M = 0.06, SE = 0.01) is significantly lower than that of the NP (M = 0.79, SE = 0.01), df = 4643.3, W > 1.7e+6, p < 2.2e-16. Put differently, the NPs are by and large better cues for the adjectives in the given sample than vice versa.

To take a closer look at Figure 6, let us zoom in on the margins of the x-axis that are characterized by a high degree of density of observations with similar scores. In fact, if we look at the density plot with the ΔP_{diff} values (Figure 7), it becomes clear that there are three conspicuous peaks that center around the extreme ends of the range, as well as around 0. The concentration of the data on the left end of the x-axis ($-1 < \Delta P > -0.95$) is defined by a small number of high-frequency ADJ types (e.g., *flat, cold, sweet, dark*, etc.) and a large variety of complex low-frequency NPs (e.g., *ticking of a clock, a diamond-studded brick, a scrap of parchment*, etc.), modified by numerous determiners, attributes and other nouns. These NPs (and not uncommonly their co-occurrence with adjectives) are unique in the given sample and often lend themselves to the category of hapax-legomena. Theoretically speaking, if ΔP reveals that the lexical item in the NP slot is a better cue to infer the item in the ADJ slot, it means that the latter co-occurs with a vast number of NP types. The more NP types are attracted to an ADJ-schema as in (9) or (10), the less informative this adjective becomes. Note that s*trong as a wet straw* in (10) represents a special use of the A *as* NP construction which encodes intensification by means of a pair of oppositional concepts, creating an ironic effect.

(9) ***smooth as*** *a sheet of glass / young baby's skin / weathered pebble on the sea shore / mirrored glass / chocolate milk* / etc.

(10) ***strong as*** *a champion / a grinding stone / hurricane / an old oak tree / a wet straw / life* / *a leopard* / etc.

The observations on the other side of the ΔP_{diff} range, although being more than 5 times as scarce as those with extreme negative scores, reveal a number of NP-schemas where the presence of an NP considerably increases the likelihood of an item in the ADJ slot. Similar to the ADJ-schemas, described above, high-frequency nouns in the NP-schemas co-occur with adjectives with extremely low frequencies. Another characteristic of the negative margin of the ΔP_{diff} range is the pervasive presence of the NP-schemas, indexed on the traditionally taboo words (Table 4).

Due to their extremely high frequencies in the given sample, they are the least informative of the items they co-occur with. This is also reflected in their ΔP_{diff} distribution scores, which are strictly confined to the positive range of the scope. The column with the respective ΔP_{diff} values for each ADJ_NP instance appears to be positively correlated with the raw frequencies of adjectives in each pair, i.e. the less frequent an adjective is, the more it increases the probability of the NP. This is best reflected in the extremely high scores for *autistic as fuck* and *uneducated as shit* simply because the adjectives *autistic* and *uneducated* are identified in the sample as hapax-legomena. Other, considerably less salient, NP-schemas include ____ *as a rock* (e.g., *steady/ dumb/bare/solid/rigid*/etc.), ____ *as a lamb* (e.g., *quiet/innocent/docile/patient/peaceful*/etc.), ____ *as a snake* (e.g., *wise/poor/slippery/crooked/clever*/etc.) and ____ *as a dog* (e.g., *sick/tired/humble/lustful/loyal*/etc.), among others.

ADJ frequency	ADJ	NP	NP frequency	Co-occurrence frequency	ΔP_{diff}
121	scary	hell	2977	102	0.573
14	irritating			12	0.611
2	humorous			2	0.756
131	hot	fuck	384	13	0.045
5	brutal			2	0.364
1	autistic			1	0.966
153	boring	shit	180	7	0.005
2	paranoid			1	0.48
1	uneducated			1	0.98

Table 4. Examples of the NP-schemas indexed on the most frequent NPs, sorted by the frequency of the adjectives and ΔP_{diff}.

From the semantic perspective, the NP-schemas with the taboo words are also defined by the lack of a motivational link between the tertium and the vehicle. In Goldberg's words, their configurations are completely non-compositional although the NP still functions as an intensifier of the adjective. These subschemas have become so widely used that they are assumed to be undergoing grammaticalization (Desagulier 2016: 4).

The Russian data display the similar picture, as shown by Figure 8 and Figure 9. The proportion of the negative and positive ΔP_{diff} scores is almost identical to that in the English sample (63% and 29%, respectively). However, the percentage of the observations with $\Delta P_{forward} = \Delta P_{backward}$ ($\Delta P_{diff} = 0$) is slightly higher and amounts to 8%, as represented by a longer black line on the x-axis. The high density of these ADJ_NP pairs is reflected in Figure 9 where the data points, centered around $\Delta P_{diff} = 0$, constitute the second highest density peak. In essence, however, the ΔP_{diff} distribution follows the same pattern as in English, peaking at the left side ($-1 < \Delta P_{diff} > -0.9$), the right side ($0.9 < \Delta P_{diff} > 1$) and in the center ($-0.1 < \Delta P_{diff} > 0$).

Identical to English, the Russian data reveal a number of subschemas. For example, in *prostoj kak* ____ (Eng.: *simple as* ___), the adjective *prostoj* co-occurs in the data with 15 different NPs (e.g., *prostoj kak prazdnik/poleno/kirpič*/etc. (Eng.: *simple as a holiday/a log/a brick*/etc.). It seems that, with an increasing number of NP types, the adjective loses its predictive power as it stops influencing the realization of the NP slot. On the contrary, every NP type used in this subschema will be associated exclusively with *prostoj*, thus, having greater predictive power.

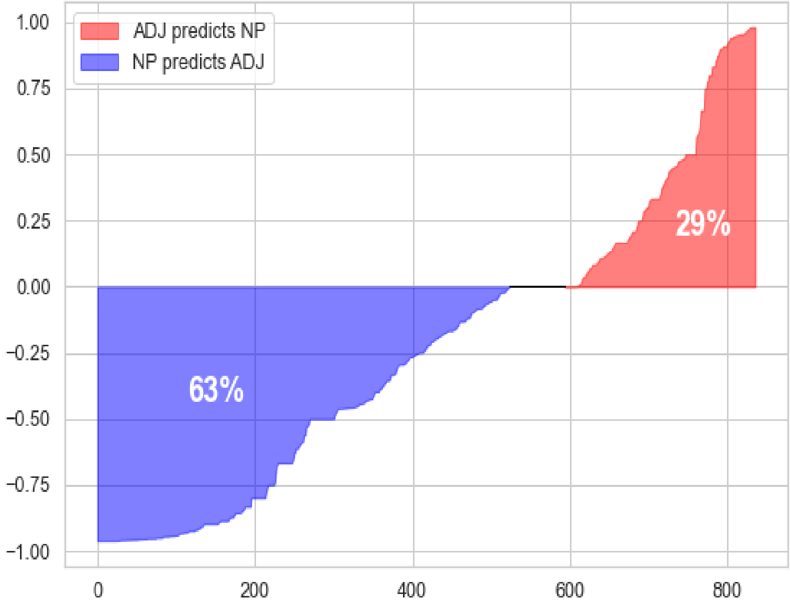


Figure 8: Distribution of ΔP_{diff} across the ADJ_NP types in the Russian sample, sorted in ascending order.

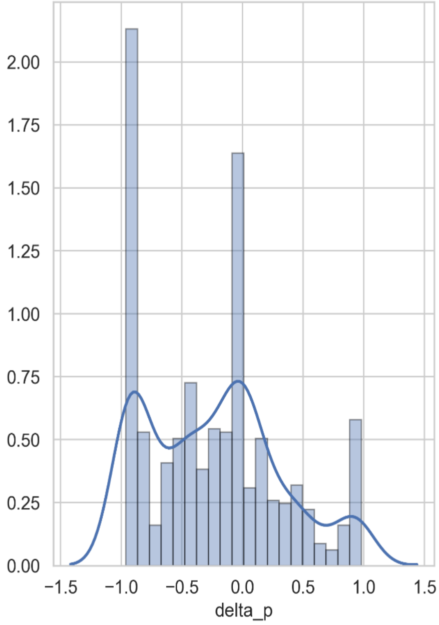


Figure 9. Density of ΔP_{diff} distribution in the Russian sample.

Just as it was the case with English, the difference between the right- and left-predictive ΔP values for each of the ADJ_NP pairs in the Russian data was tested for significance. The data distribution shape in Russian closely resembles that in English and does not meet the assumptions of normality and homogeneous variance, which calls for a non-parametric Wilcoxon test. The result of the test is statistically significant ($W > 30.1e+4$, $df = 1857.3$), with the p-value lying well below the threshold of 0.05 ($p < 2.2e-16$). The probability of encountering the observed data is, thus, exceedingly small, which gives us a reason to reject the null hypothesis of no difference between the $\Delta P_{forward}$ ($M = 0.17$, $SE = 0.01$) and $\Delta P_{backward}$ ($M = 0.49$, $SE = 0.01$). Bearing on these findings, one can conclude that the NP slot in both English and Russian, in general, serves as a better cue to infer a lexical item in the ADJ slot.

This is congruent with the conclusion on the direction of lexical association in Russian in the recent study by Watson Todd (2019). He explored the direction of collocations in eight genealogically unrelated languages and classified Russian as a language with stable leftward informativeness. This is, however, not the case for English, which, according to Watson Todd, shows a preference for right-predictive collocations (including adpositional phrases). In contrast, an earlier study by Onnis and Thiessen (2013) reports that English speakers readily exploit leftward statistical associations in analyzing the grammatical structures of utterances (cf. Wahl 2015). The reason is claimed to be the typical adjacency of function words (mainly prepositions) to open-class content words. Such grammatical configurations are frequently analyzed ‘retrospectively’, drawing on the content word. The findings of Onnis and Thiessen, however, cannot be extrapolated to the current study, which is strictly restricted to a very specific construction and the subject of the investigation is confined to the slots with content words.

Lastly, from the semantic point of view, most frequent ADJ- and NP-schemas in Russian (as in (11) and (12)) refer to entities of the physical world or their properties. Due to the semantic compatibility of the domains of the tertium and the vehicle, these instances can be deemed to represent the prototypical use of the A *as* NP construction.

(11) **prostoj kak ____** (jaičnitsa na zavtrak/poleno/prazdnik/kirpič/valenok)
Eng.: **simple as _____** (an omelette for breakfast/a log/a holiday/a brick/a valenok)
čistyj kak ____ (rosa/angel/l'ubov'/kristall/mladenec)
Eng.: **pure as _____** (dew/an angel/love/a crystal/an inflant/etc.)

(12) (iskrennij/nevinnyj/doverčivyj/l'ubopytnyj) **____ kak rebenok**
Eng.: (sincere/pure/naive/curious) **____ as a child**
(zloj/gryaznyj/strashnyj/obayatel'nyj/skučnyj) **____ kak chort**
Eng.: (angry/dirty/scary/charming/boring) **____ as the devil**

To briefly summarize, in terms of asymmetrical association, the A *as* NP construction in English exhibits the same pattern as in Russian, notwithstanding the differences in the sample size. The NP slot is significantly more informative for inference of an element in the ADJ slot, than vice versa. The ΔP association measure unveiled a number of ADJ- and NP-schemas, with the node word attracting a great variety of items to the lexically dependent slot. All things considered, the English A *as* NP construction appears to be more idiosyncratic in that it manifests incipient grammaticalization of certain NP-schemas, indexed on the "socially inappropriate" taboo terms. On the contrary, in Russian, this construction appears to be semantically closer to the prototype because it displays a stronger semantic link between the tertium and the vehicle while the ground is the most salient property of the vehicle.

7.3. Relationship between Symmetric and Asymmetric AMs

Lastly, the relationship between the two AMs – the LLR of symmetrical lexical association and the ΔP measure of directed association – were examined with the help of two-dimensional relational plots from the Python package seaborn (Waskom 2017). The results are illustrated in Figure 10 for English and Figure 11 – for Russian. The LLR scores were $\log_e$-transformed to overcome the heavy positive skewness of the original data, centered around the lowest LLR values.

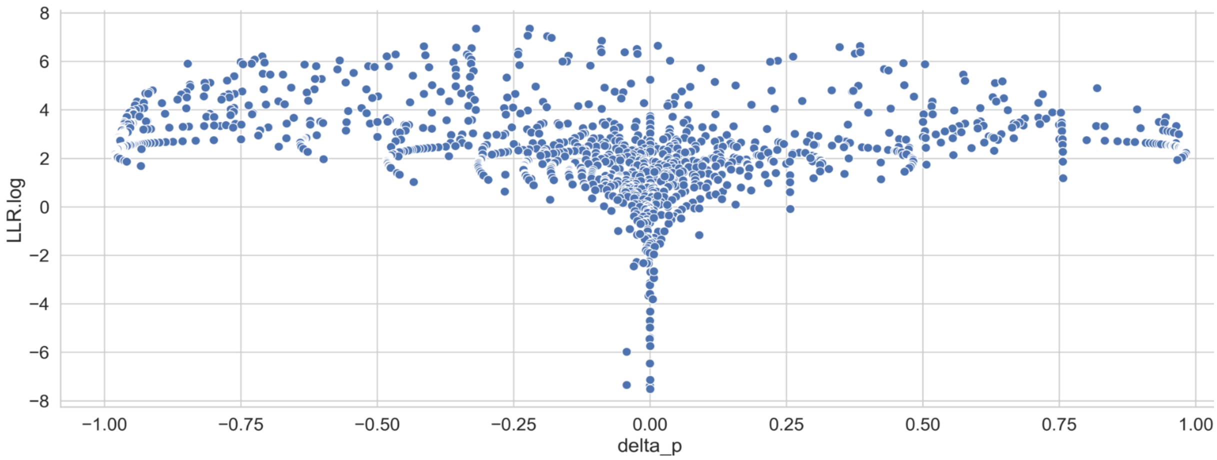


Figure 10. The relationship between the log-transformed LLR and ΔP_{diff} in English.

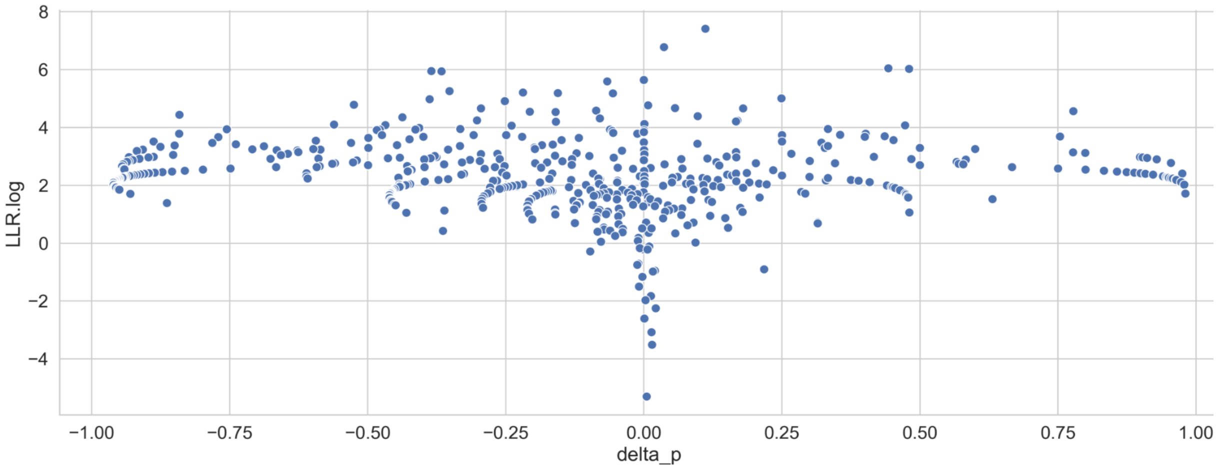


Figure 11. The relationship between the log-transformed LLR and ΔP_{diff} in Russian.

Due to the smaller sample size, the Russian data is distributed noticeably less densely but nevertheless exhibit a similar pattern as in English. Besides, it might also be the reason why the data points centering around $\Delta P_{diff} \approx 0$ in Figure 11 stretch out to the upper side of the plane, making the plot look more symmetrical than that in Figure 10. In contrast, the data points in Figure 10 are mostly concentrated in the upper part of the plot with a significantly long tail down the y-axis where LLR.log < -1. Furthermore, Figure 10 demonstrates a densely populated center whereas the data points in Russian are more evenly distributed. However, as was already mentioned above, these discrepancies could be attributed to the difference in the sample sizes.

As the margins of ΔP_{diff} were already examined in Chapter 7.2., they will not be included in the discussion here. The long tail of the data points with ΔP_{diff} in the range [-1, 1] and LLR.log < -1 in Figure 10 is characterized by the ADJ_NP pairs with little or no attraction. The negative log of LLR here corresponds with the original LLR scores ranging between 0 and 1 and nearing the attraction/repulsion threshold. These observations are represented by the adjectives and the NPs that typically do not co-occur together. Example (13) is one such case in point.

(13) harp as a crystal ($f_{\text{co-occurrence}} = 1$)

The raw frequency of *sharp* in the given sample is 137 and that of *crystal* – 88. Being the most commonly used words in the English A *as* NP construction, they co-occur together just once. On the other hand, the NPs that are most frequently encountered with *sharp* are *tack* ($f_{\text{co-occurrence}} = 57$), *razor* ($f_{\text{co-occurrence}} = 16$) and *knife* ($f_{\text{co-occurrence}} = 13$) while the more typical adjectives to co-occur with *crystal* are *clear* ($f_{\text{co-occurrence}} = 78$), *pure* ($f_{\text{co-occurrence}} = 3$) and *transparent* ($f_{\text{co-occurrence}} = 3$). This is quite similar to the Russian observations on the bottom margin of the plot in Figure 11. For example, the instance in (14) can be counted as a hapax-legonema in the given data set because its co-occurrence frequency equals 1: each word in this expression is not the most apparent lexical choice for the other. The top three candidates to be associated with the adjective *černyj* (f = 106) are *smol’* (lit. pitch, $f_{\text{co-occurrence}} = 58$), *ugol’* (lit. coal, $f_{\text{co-occurrence}} = 13$), and *noč* (lit. night, $f_{\text{co-occurrence}} = 13$).

(14) černyj kak čert ($f_{\text{co-occurrence}} = 1$)
Eng.: black as the devil

These findings provide evidence in support of the idea of semantic preference, according to which lexical and grammatical preferences of individual words are stored together with the mental representation of words. It seems, other factors being equal, that the frequency of occurrence of an item in a specific collocation/construction plays a significant role in strengthening the associative link between the two components and facilitates the faster activation of this item in memory, given the collocation/construction. On this account, statistical learning mechanisms enable speakers to make lexical decision based on the ongoing distributional analysis. It also explains why some words are dispreferred in the A *as* NP construction despite semantic similarity or synonymity (e.g., *white as a sheet* is preferred over *pale as a sheet* or *vast as the ocean* sounds more natural than *wide as the ocean*).

The interpretation of the top margin of the relational plots is far less straightforward. To begin with, the English data exhibit more observations with extreme positive LLR scores as shown by a densely populated cloud of data points in the upper part of Figure 10, compared to Figure 11. What is represented by a substantial amount of data in English with LLR.log > 6 is just two outliers in the Russian sample (see Table 3 above). As was elaborated in Chapter 5.1., these observation can be referred to as lexical prefabs, chunks, or idioms, owing to their strong mutual attraction. Interestingly, they are also inclined to be rather left-predictive in both English and Russian (the mean ΔP_{diff} equals -0.22 and -0.12, respectively).

If we compare the spread of the data along the x-axis in Figure 10 and Figure 11, it becomes apparent that the largest variance of the ΔP_{diff} data points is reached when LLR.log is in the range (4, 6) for English and (2, 4) for Russian, respectively. This could be interpreted the following way: the most salient A *as* NP subschemas with the highest asymmetrical association scores are located in the center of the data, ranked by the collocational strength LLR. Another observation worth pointing out is the LLR data distribution, which is identical in both samples, notwithstanding the difference in the sample size. The measures of central tendency – the mean and the median – yield extremely similar values (M = 2.13, med = 2.23 for English and M = 2.34, med = 2.37 for Russian), signaling potential symmetry in the data distribution. This assumption is additionally confirmed by density plots and the results of the Shapiro–Wilk test (p-value = 1 in each case). Apparently, irrespective of the data volume, the most observations group in the range $2 < \text{LLR.log} < 3$. Having said that, however, an increase of the sample size is positively correlated with the increase of dispersion of association strength scores. Compare: $SD_{log\text{-}LLR} = 1.37$ in English and $SD_{log\text{-}LLR} = 1.03$ in Russian.

Up to this point, we have examined the symmetric and asymmetric components of the collocativity in the A *as* NP construction and attempted to characterize it based on the association scores at the extreme ends. The following chapter, however, is intended to shed light on the semantic relationship between the two slots in both languages and determine the semantic blueprint of the construction.

7.4. Semantic Relationship between ADJ and NP

The further semantic analysis is based on the annotation tags, derived from the Concepticon database, as mentioned in Chapter 6.3. Before we move on to the results, one should be aware of the incomplete annotation of the data, resulted from the partial overlap between the lexical items in the sample and the Concepticon concepts. It is for this reason that the annotated share of the given sample might turn out to be insufficient to reflect the real world language use. A thorough manual annotation, however, with special attention to complex noun phrases and extended context in ambiguous cases could potentially lead to improved results.

The preliminary findings concerning the distribution of the semantic categories in both ADJ and NP slots in English are illustrated by Figure 12 and Figure 13, respectively. The plots in the first column (Figure 12a and Figure 13a) represent the bar plots where each bar stands for a particular semantic category and its size reflects the frequencies of the lexical items tagged with the respective semantic tag. The numbers on top of each bar reproduce the counts once again. The blue line on top of the bar plots represents the mean LLR scores for the observations withing each semantic category, and it is plotted against the right-hand y-axis. Lastly, the bar charts in the second columns (Figure 12b and Figure 13b) reflect the mean ΔP_{diff} scores for the lexical elements in each semantic category.

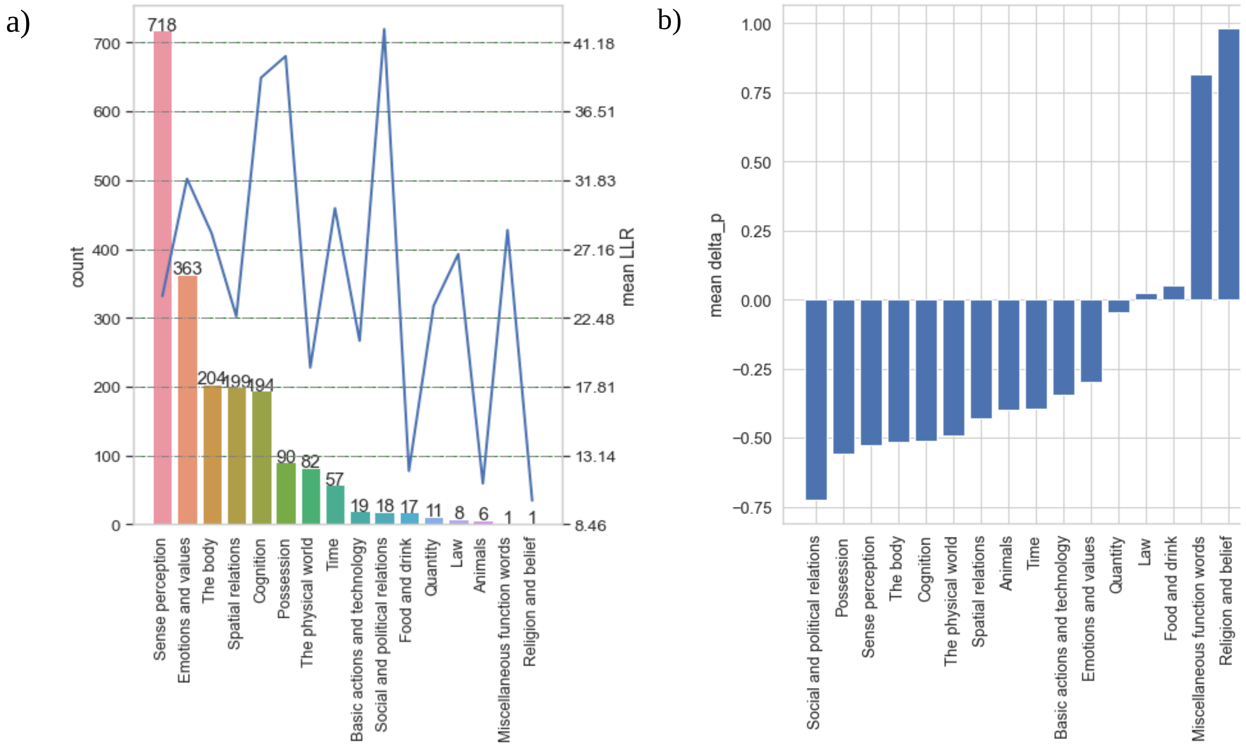


Figure 12. Semantic categories in the ADJ slot by their frequency (left), mean LLR (left) and mean ΔP_{diff} (right) in English.

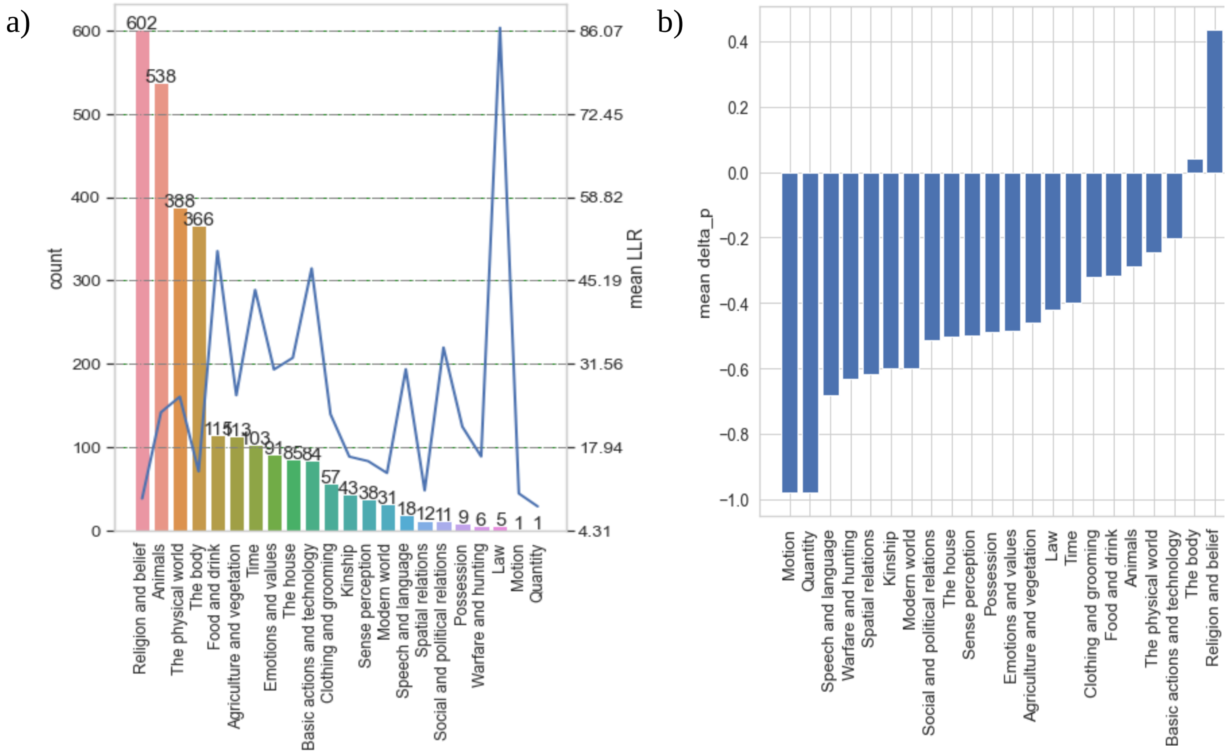


Figure 13. Semantic categories in the NP slot by their frequency (left), mean LLR (left) and mean ΔP_{diff} (right) in English.

Overall, the most lexical items in the ADJ slot (Figure 12a) refer to a bigger or lesser degree to the domain of emotions and sensory perceptions. Another, although certainly less significant, conceptual source for the ADJ slot is THE BODY, SPATIAL RELATIONS and COGNITION. The most frequent items in the NP slot, on the other hand, (Figure 13a) denote the concepts relating to RELIGION AND BELIEF (inflated by a large number of *hell*-occurrences), ANIMALS, THE PHYSICAL WORLD and, to a lesser extent, THE BODY (made up mostly by the high raw frequencies of *shit* and *fuck*).

7.5. Semantic Relationship with Symmetric Association

The comparison of the semantic categories with the mean LLR scores, as shown by the line chart in Figure 12a for the ADJ slot and Figure 13a for the NP slot, yields the following semantic configurations of the most attracted lexical items:

Semantic pattern	Example	Mean LLR
[COGNITION as TIME]	sure as the sunrise	807
[POSSESSION as ANIMALS]	free as a bird	677
[POSSESSION AS THE BODY]	rare as hen’s teeth	541
[COGNITION as FOOD AND DRINK]	easy as a pie	406
[COGNITION as BASIC ACTIONS AND TECHNOLOGY]	clear as bell	396
[SPATIAL RELATIONS as LAW]	thick as thieves	362

Table 5. The top semantic patterns of the A *as* NP construction in English, sorted by the symmetric association LLR scores.

With the exception of the last configuration, whose members in the given sample are rather scarce, the ADJ slot at the top of the symmetric association scale appears to be filled by the elements encoding either a spectrum of mental and cognitive states or different degrees of possession. By contrast, the NP slot on this margin (the uppermost part of Figure 10), manifests a significantly greater semantic variability. As a matter of fact, a greater semantic versatility of the vehicle becomes especially apparent in the quick overview of the semantic repositories, represented in each slot: whereas there are 22 semantically different NPs, the semantic space of adjectives is carved up by just 16 categories.

The quantitative analysis of the relationship between the semantics of the slots and the mutual collocational strength was performed with the use of a multiple linear regression analysis. The purpose of this analysis is to predict the symmetric association (i.e. the LLR) given the semantic categories of both slots. Before feeding the data to the linear regression model, the levels of the categorical parameters with less than 20 observations were dropped.

Despite the general rule of thumb of minimum 10 events per variables, recent studies (e.g., van der Ploeg et al. 2014) have highlighted the importance of the sufficient data size by introducing the concept of data hungriness – “the sample size needed for a modeling technique to generate a prediction model with a good predictive accuracy” (van der Ploeg et al. 2014). According to the scientists, the optimal number of observations per variable varies from 20 to 50. Besides, apart from a more optimal model performance, the data reduction for the regression analysis had a goal of avoiding overfitting in the light of “heavy” categorical parameters with multiple levels. However, even after the removal of the categories with fewer than 10 observations, the variables included 13 values for the ADJ slot and 17 values for the NP slot.

The model with the formula below includes the interaction between the semantic categories of the ADJ and NP slots as there is a reason to believe that one slot is influenced by the other. According to the ANOVA analysis, the models with and without the interaction term are significantly different.

formula = LLR.log ~ Semantic_category_NP + Semantic_category_ADJ +
Semantic_category_NP * Semantic_category_NP

Before we proceed to the discussion of the results, let us check the regression assumptions. First of all, there is no reason to believe that the observations in the given data are dependent as they were all drawn from a thoroughly compiled and well balanced corpus without any further biased selection. Secondly, the response variable is continuous, which allows its modeling in the linear regression. Thirdly, the assumption of the linear relationship between the dependent and independent variables is not applicable to the given sample as both predictors are categorical and not ordinal, which makes the concept of linearity irrelevant here. Fourthly, the assumption of constant error variance is violated in the original data (the left plot in Figure 14), which is confirmed by the results of the non-constant variance test from the package *car* (Fox and Weisberg 2019). The returned p-value is above the significance level, allowing one to reject the null hypothesis of homoscedasticity of the response variable ($p < 2.22e-16$). To bypass this issue, the LLR scores were added 7.509 in order to overcome the problem of negative values and find the optimal power transformation of the data with the help of the Box-Cox plot from the car package. Raising the LLR variable to the power of 1.7, suggested by the plot, has eliminated the problem of heteroscedastic error variance ($p = 0.196$, see the right plot in Figure 14).

The assumption of no multicollinearity was tested with the use of the vif()-function from the car package. The VIF-scores in the output did not exceed 5, considered the general cut-off point, so multicollinearity is not a matter of concern. The Durbin-Watson test was used to check the data for the presence of autocorrelation between the residuals. It is known for a fact, that sorting values can have a serious impact on the result of this test and, since our data were initially sorted by the collocational strength in descending order, we reordered the data in an arbitrary way to eschew the interpretation of the given order as meaningful. The p-value, returned by the test, is greater than 0.05 (p = 0.15) so we can conclude that there is no autocorrelation.

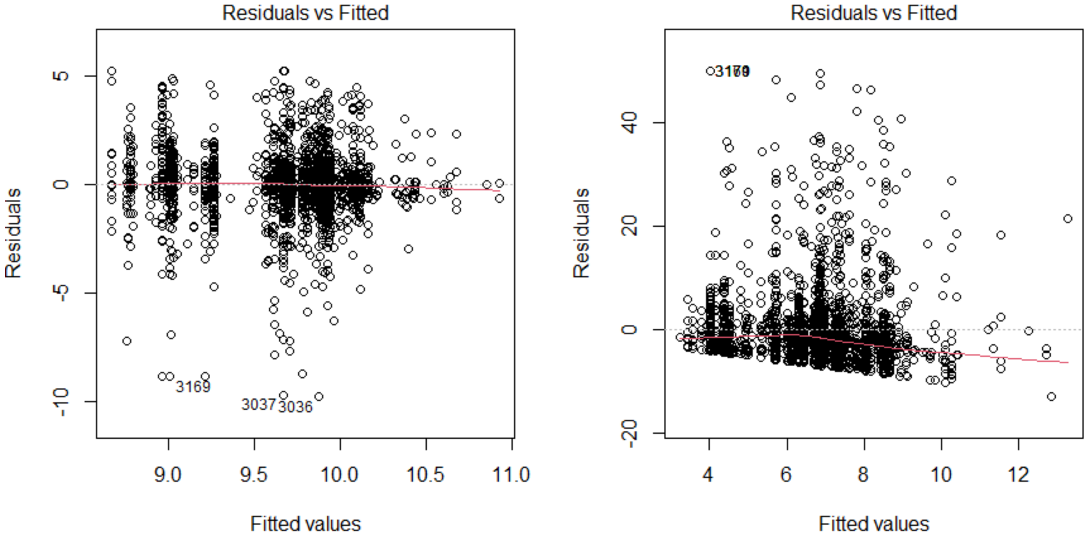


Figure 14. The error variance before the power transformation of the response variable (left) and after (right).

Additionally, the data were checked for outliers and overly influential observations with large differences between the observed and predicted estimates, indicative of the lack of fit. Such problematic values were detected with the help of the interactive plot, called by the influencePlot()-function in the car package. Subsequently, a number of observations with inflated hat-values above 1 and high Cook's distances were removed from the data (the left plot in Figure 15) so that the largest hat-values, kept in the sample, did not exceed 0.6 (the right plot in Figure 15).

The model with the interaction term is significant (F(2800) = 3.94, p < 2.2e-16). The goodness-of-fit is rather poor ($R^2_{adjusted}$ = 0.12), meaning that the model accounts for roughly 12% of the data variance. Initially, the table of coefficients contained over 220 items so, for the sake of brevity, Table 6 displays only the coefficients with the p-values above 0.05. The reference level for the semantic category of the ADJ slot is set to POSSESSION and for the semantic category of the NP slot – to RELIGION AND BELIEF. The estimates were raised to the power of 1/1.7 to reverse the log-transformation, performed at the stage of correcting the error variance. Thus, the estimates in Table 6 represent the simple log-LLR, introduced originally for the sake of investigating the data with the lowest scores of the raw LLR. The estimates in red are negative values that indicate a decrease in the log-LLR, compared to the LLR score at the reference level, whereas the black estimates imply an increase thereof.

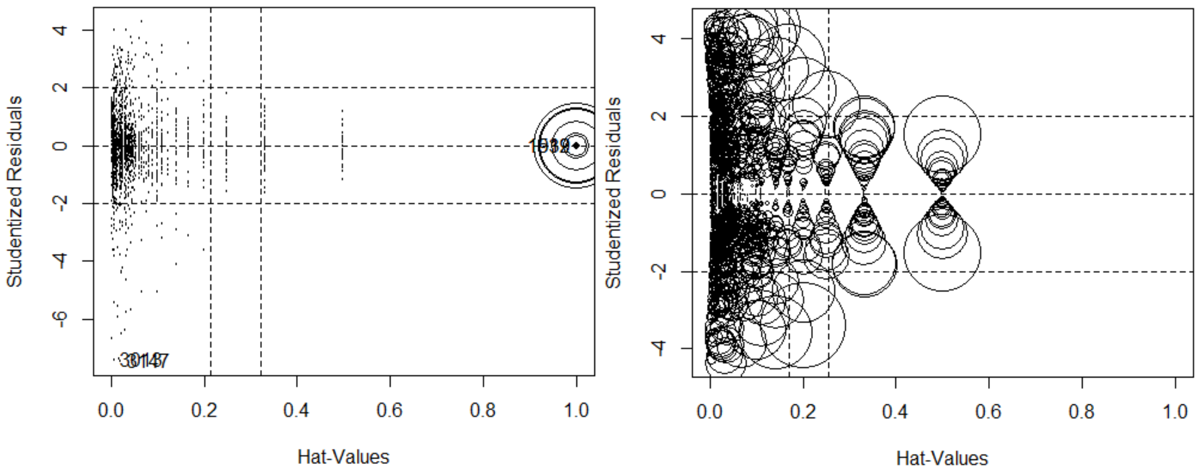
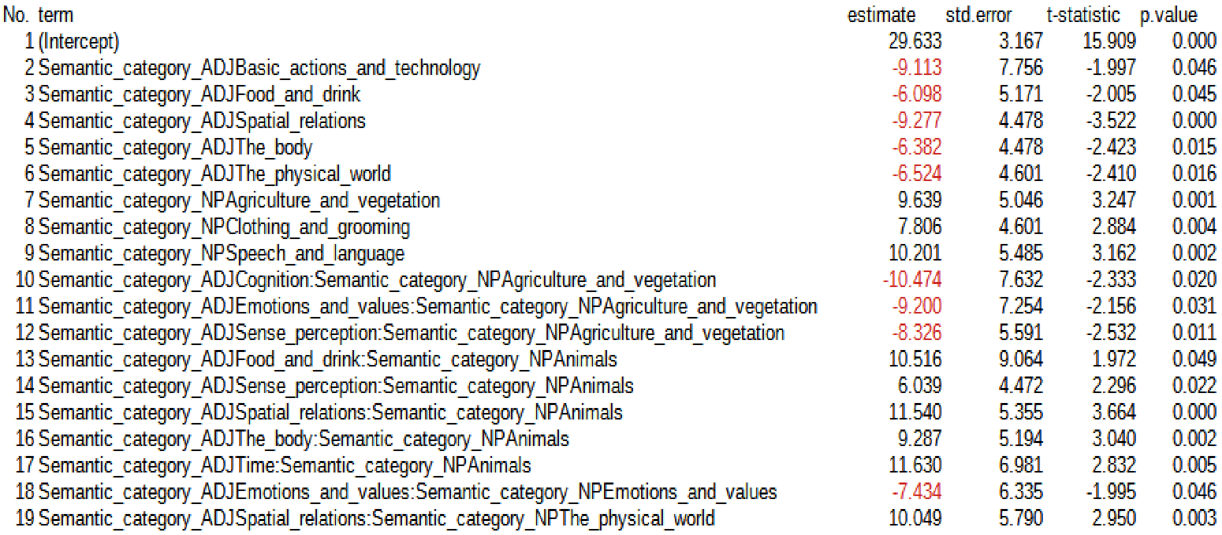


Figure 15. Outliers and overly influential observations with discrepancies in the observed and fitted values before (left) and after the data cleaning (right).



No.	term	estimate	std.error	t-statistic	p.value
1	(Intercept)	29.633	3.167	15.909	0.000
2	Semantic_category_ADJBasic_actions_and_technology	-9.113	7.756	-1.997	0.046
3	Semantic_category_ADJFood_and_drink	-6.098	5.171	-2.005	0.045
4	Semantic_category_ADJSpatial_relations	-9.277	4.478	-3.522	0.000
5	Semantic_category_ADJThe_body	-6.382	4.478	-2.423	0.015
6	Semantic_category_ADJThe_physical_world	-6.524	4.601	-2.410	0.016
7	Semantic_category_NPAgriculture_and_vegetation	9.639	5.046	3.247	0.001
8	Semantic_category_NPClothing_and_grooming	7.806	4.601	2.884	0.004
9	Semantic_category_NPSpeech_and_language	10.201	5.485	3.162	0.002
10	Semantic_category_ADJCognition:Semantic_category_NPAgriculture_and_vegetation	-10.474	7.632	-2.333	0.020
11	Semantic_category_ADJEmotions_and_values:Semantic_category_NPAgriculture_and_vegetation	-9.200	7.254	-2.156	0.031
12	Semantic_category_ADJSense_perception:Semantic_category_NPAgriculture_and_vegetation	-8.326	5.591	-2.532	0.011
13	Semantic_category_ADJFood_and_drink:Semantic_category_NPAnimals	10.516	9.064	1.972	0.049
14	Semantic_category_ADJSense_perception:Semantic_category_NPAnimals	6.039	4.472	2.296	0.022
15	Semantic_category_ADJSpatial_relations:Semantic_category_NPAnimals	11.540	5.355	3.664	0.000
16	Semantic_category_ADJThe_body:Semantic_category_NPAnimals	9.287	5.194	3.040	0.002
17	Semantic_category_ADJTime:Semantic_category_NPAnimals	11.630	6.981	2.832	0.005
18	Semantic_category_ADJEmotions_and_values:Semantic_category_NPEmotions_and_values	-7.434	6.335	-1.995	0.046
19	Semantic_category_ADJSpatial_relations:Semantic_category_NPThe_physical_world	10.049	5.790	2.950	0.003

Table 6. The significant coefficients of the linear regression model with the LLR as the response variable.

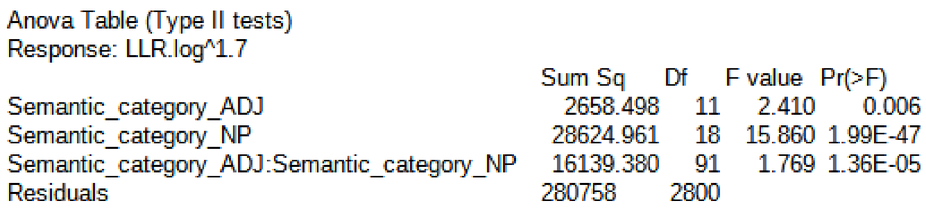
First of all, the log-LLR is equal 29.6 when the interaction terms are at their reference levels: the ADJ is associated with the domain of possession while the NP – with the domain of religion and belief. Interpretation of the other coefficients in the model with an interaction term should be done with particular caution as the coefficients of the terms in rows 2-9 do not correspond to the effect of the respective variable for all the data, but only for the observations with the reference level of the interacting term. For example, the collocational strength is decreased by 9.1 when the NP reference level co-occurs with an adjective from the domain FOOD AND DRINK. The adjectives, associated with BASIC ACTIONS AND TECHNOLOGY, SPATIAL RELATIONS, THE BODY, or THE PHYSICAL WORLD, likewise, bring down the LLR score of the

[____ *as* RELIGION AND BELIEF] configuration. POSSESSION was chosen as the reference level for the ADJ slot because it has the highest mean LLR score and, hence, represents a convenient comparand. It is for this reason, that all the ADJ semantic categories in rows 2-6 exhibit negative estimates.

In the NP slot, the picture is completely the opposite because the category RELIGION AND BELIEF was chosen as the reference level with the lowest mean LLR score. Thus the positive estimates for the significant semantic categories in the NP slot. The change of the NP slot from RELIGION AND BELIEF to AGRICULTURE AND VEGETATION, CLOTHING AND GROOMING, and, especially, SPEECH AND LANGUAGE, in the [POSSESSION *as* ____] pattern triggers an increase in the collocations strength.

The estimates of the interaction term, marked with a ":", should be interpreted as follows: any change in the given semantic category in the ADJ slot from the reference level POSSESSION and in the NP slot – from the reference level RELIGION AND BELIEF – to any other category either raises the lexical association strength if the estimate is positive or decreases it if the estimate is negative. Essentially, an occurrence of a lexical item from the domain AGRICULTURE AND VEGETATION, compared to that of RELIGION AND BELIEF, causes a decline in the mutual attraction score of the construction, on the whole. The NPs, associated with animals, on the contrary, are evident of stronger association between the lexical components.

All in all, a large number of significant coefficients prevents one from drawing more meaningful comparisons between the numerous levels of the predictors. To measure the overall significance of the predictors in the given model, however, one can resort to the ANOVA analysis, which yields the following results:



```
Anova Table (Type II tests)
Response: LLR.log^1.7
                                               Sum Sq      Df    F value  Pr(>F)
Semantic_category_ADJ                         2658.498     11     2.410     0.006
Semantic_category_NP                         28624.961     18    15.860  1.99E-47
Semantic_category_ADJ:Semantic_category_NP   16139.380     91     1.769  1.36E-05
Residuals                                    280758      2800
```

Figure 16 . The results of the ANOVA variance analysis for linear regression model to assess the statistical significance of the predictors.

Apparently, the main contribution to the explanatory power of the model is made by the semantic category of the vehicle, as evidenced by the largest F-value in the table F(18) = 15.86, p = 1.99e-47). Although the interaction between the predictors is significant (p < 1.36e-05), its effect is relatively moderate, compared to the other coefficients in the model. Bearing on the findings, one can conclude that that symmetric lexical association is heavily contingent on the semantic affiliation of both slots but especially the NP slot. Furthermore, the distribution of semantic categories between the slots, as expected, is not arbitrary; instead it is indicative of the mutual impact of the slots on each other's semantics.

7.6. Semantic Relationship with Asymmetric Association

Figure 12b and Figure 13b are intended to illustrate how semantically different lexical items are distributed in relation to ΔP_{diff}. To this end, the mean ΔP_{diff} was calculated for each category. First of all, the results shown by the bar plots exhibit extreme skewness of the data, resulted from a large number of syntactically complex items in both slots, which, in its turn, hindered the automatic semantic annotation. For example, the elements filling in the ADJ slot such as *dry and unforgiving* (*as a desert*) or *dark and impenetrable* (*as the night*) were not assigned a semantic tag because of their structural and conceptual complexity. Yet another reason for the skewness of the data is the higher density of data points with mean $\Delta P_{diff} < 0$ than with positive scores. These two factors account for the higher concentration of semantic categories in the negative range of ΔP_{diff}. We will focus on those with the positive ΔP_{diff} means.

As shown by Figure 12b, the adjectives in the right-predictive instances of the A *as* NP construction are tagged as RELIGION AND BELIEF and MISCELLANEOUS FUNCTION WORDS. However, these categories can be disregarded as both are represented by hapax-legomena: the former category is expressed by *holy as a dove* while the latter – by *opposite as day and night*. The other two candidates for the ADJ slot in the right-predictive instances are LAW and FOOD AND DRINK. The low adjective type count in the LAW category strips it of any potential value for a quantitative analysis, which makes the category FOOD AND DRINK the main source of adjectives in the right-predictive construction instances. Among them are such observations as *ripe as a cherry, easy as a pie, tasty as omelets without egg, thirsty as a camel*, *raw as fuck*, and others.

The majority of the vehicles to occur in the right-predictive A *as* NP instances, as shown by Figure 13b, encode concepts from the domain RELIGION AND BELIEF. Although the overwhelming majority of lexical items in this category consists of *hell* tokens, other members are represented by *god* (as in *powerful as god*), *heaven* (as in *clear as heavens*), *devil* (as in *sure as the devil*), *ghost* (as in *sneaky as a ghost*), *angel* (as in *innocent as an angel*), *church bell* (as in *clear as a church bell*), etc. The second, although considerably less probable, source domain for the NP slot in the right-predictive A *as* NP construction is THE BODY. This is partly due to the massive presence of the high-frequency words *shit* and *fuck* that construct the productive subschemas, attracting a semantically unlimited variety of tertia. That being said, THE BODY also incorporates the concepts denoting different parts of the human and animal body that can be partitioned as follows: the NPs referring to the face (e.g., *hen’s teeth*, the *nose on* [*his/her/my*/etc.] *face*, *woodpecker's lips*, *a baby's cheek*, a *bag of hair*), appendages and their parts (e.g., *a virgin's thigh, duck's instep, a crow's wing, a dog's hind leg*), organs (e.g., *womb*), bodily fluids (e.g., *blood*), and the like.

The third common lexeme on this list is *death* and its variations, such as the addition of the reflexive pronoun to amplify the intensification or its typical co-occurrence with the word *taxes* as in *certain as death and taxes*. Despite the fact that this NP is found among the top most frequent NP types within the category THE BODY, the expressions with the pattern ____ as *death* show no signs of semantic bleaching, in sharp contrast to *fuck*- and *shit*-schemas. Quite the opposite, the lexical association between the adjectives co-occurring with the NP *death* in this case is perfectly semantically motivated. For instance, the tertia in the examples in (15) are unambiguously related to the negative connotations of the imminence and gravity, universally associated with the concept of mortality.

(15) pale as death
silent as death
inescapable as death
quiet as death
dark as death
sure as death
ugly as death
fierce as death

The relationship between the directional association and the semantics of the slots of the A *as* NP construction was analyzed used a beta regression analysis from the betareg package (Simas and Rocha 2006), providing a sub-type of the generalized linear regression models. The main reason behind the choice of this particular model is the limited distribution of ΔP_{diff}, tied to the interval [-1, 1]. The data of this type (typically restricted to the range between 0 and 1) are allegedly characterized by a beta distribution. Before performing the analysis, the Δp_{diff} values were first transformed to approximate to the required interval of (0, 1) by virtue of the min-max normalization (16).

(16) $$z=\frac{x-min(x)}{max(x)-min(x)}$$

It is important to mention one essential peculiarity of the beta regression model from the betareg package, which is the requirement to adhere to the open interval (0, 1). In other words, the endpoints of the interval with the response values are not to be included (Ferrari and Cribari-Neto 2004). In case the predicted variable has a substantial number of observation at the limit points of the interval, Smithson and Verkuilen (2006) recommend the following data transformation: (y * (n − 1) + 0.5) / n where n is the sample size.

It is of note that the beta regression analysis is a relatively new method for modeling the data with a limited response distribution and has recently been a topic of discussion on appropriate diagnostic tools (for details, see Espinheira et al. 2008, Rocha and Simas 2011, Ferrari et al. 2011). Among other caveats of the betareg package is, for example, the lack of infrastructure for constructing confidence intervals the way it would be possible to do in R for a linear regression. This issue is attributed to certain differences between the beta and linear regression models: the variance of the predicted values in the former refers to the variance of the response while in the latter – to the variance of the predicted mean. Another fundamental difference is the precision coefficient, computed for a beta regression model. It can be viewed as a dispersion measure: the higher its score, the better the fit of a model and the smaller its error variance. Therefore, the best model improvement strategy would be to detect and minimize the residual dispersion by removing influential outliers.

As it was the case for the linear regression in the preceding chapter, the levels of the predictors containing fewer than 20 observations were kept out of the data. The model was intended to predict the direction of association and its magnitude in the A *as* NP construction, given the semantic categories of the slots. It was built used the following formula:

formula = norm_delta.p ~ Semantic_category_ADJ + Semantic_category_NP

The interaction term between the predictors was not included as the betareg model does not allow for one. The graphical tools for detecting high-leverage data points produced a Cook's distance plot (left-side in Figure 17) and a leverage-predicted values plot (right-side in Figure 17).

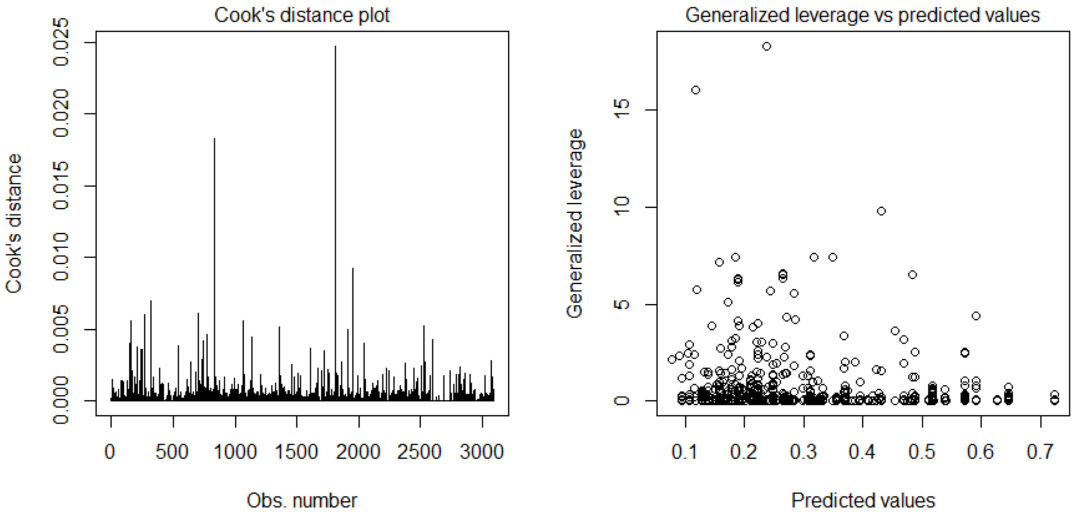
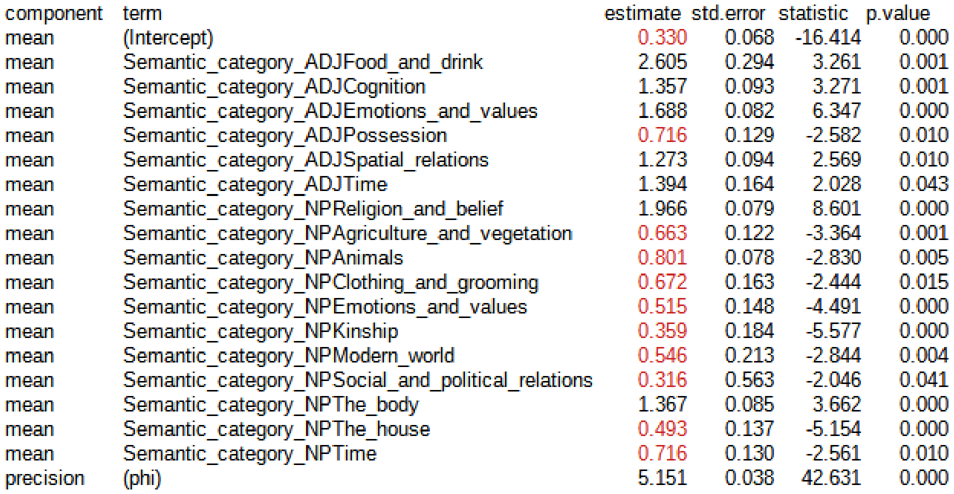


Figure 17. A Cook's distance plot (left) and a leverage/predicted values plot (right) for beta regression model.

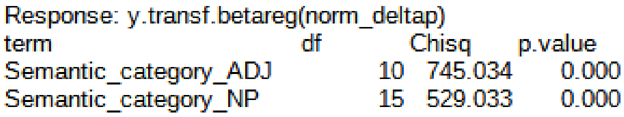
The observations with Cook's distance larger than 0.0065 and leverage greater than 10 were removed from the data. The model was then tested on homoscedasticity of the error variance with the help of the studentized version of the Breusch–Pagan test (Breusch and Pagan 1979), proposed by Koenker (1981) and available in the lmtest R package (Zeileis and Hothorn 2002). The extremely low p-value ($p < 2.2e-16$) lets us infer that the model exhibits certain heteroscedasticity. However, one of the advantages of the beta regression is its ability to attenuate non-constant residual variance: “one alternative would be to consider a logit-transformed response in a traditional OLS [ordinary least squares] regression but this would make the residuals asymmetric. However, both issues – heteroskedasticity and skewness – can be alleviated when a beta regression model with a logit link for the mean” (Cribari-Neto and Zeileis 2010: 11).

The results of the model can be found in Figure 18, which contains only statistically significant coefficients. The Figure 19 below illustrates the output of the ANOVA analysis, estimating the contribution of each regressor to the predictive power of the model. It can be easily seen that both variables are equally relevant in the model and cannot be removed from it without a substantial loss in the explanatory power.



```
component  term                                          estimate  std.error  statistic  p.value
mean       (Intercept)                                      0.330      0.068    -16.414    0.000
mean       Semantic_category_ADJFood_and_drink              2.605      0.294      3.261    0.001
mean       Semantic_category_ADJCognition                   1.357      0.093      3.271    0.001
mean       Semantic_category_ADJEmotions_and_values         1.688      0.082      6.347    0.000
mean       Semantic_category_ADJPossession                  0.716      0.129     -2.582    0.010
mean       Semantic_category_ADJSpatial_relations           1.273      0.094      2.569    0.010
mean       Semantic_category_ADJTime                        1.394      0.164      2.028    0.043
mean       Semantic_category_NPReligion_and_belief          1.966      0.079      8.601    0.000
mean       Semantic_category_NPAgriculture_and_vegetation   0.663      0.122     -3.364    0.001
mean       Semantic_category_NPAnimals                      0.801      0.078     -2.830    0.005
mean       Semantic_category_NPClothing_and_grooming        0.672      0.163     -2.444    0.015
mean       Semantic_category_NPEmotions_and_values          0.515      0.148     -4.491    0.000
mean       Semantic_category_NPKinship                      0.359      0.184     -5.577    0.000
mean       Semantic_category_NPModern_world                 0.546      0.213     -2.844    0.004
mean       Semantic_category_NPSocial_and_political_relations 0.316    0.563     -2.046    0.041
mean       Semantic_category_NPThe_body                     1.367      0.085      3.662    0.000
mean       Semantic_category_NPThe_house                    0.493      0.137     -5.154    0.000
mean       Semantic_category_NPTime                         0.716      0.130     -2.561    0.010
precision  (phi)                                            5.151      0.038     42.631    0.000
```

Figure 18. The results of the beta regression model with statistically significant coefficients. The estimates, originally log-transformed, are exponentiated.



```
Response: y.transf.betareg(norm_deltap)
term                       df      Chisq    p.value
Semantic_category_ADJ      10    745.034      0.000
Semantic_category_NP       15    529.033      0.000
```

Figure 19. The significance of the predictors in the beta regression model, produced by the ANOVA test.

According to the likelihood-ratio test, performed with the help of lrtest()-function from the lmtest package, the model is overall significant ($X^2(2207) = 1447.8$, $p < 2.2e-16$), the probability of obtaining the given data by chance is extremely small. To interpret the table of coefficients correctly, it is important to remember that the mean response is automatically transformed by the logit link function, which means that the estimates in Figure 18 represent log odds. The precision equation φ, however, is calculated using the identity link.

Interpretation of the table of coefficients requires one to know the reference levels of the predictors. For the ADJ predictor, it is represented by SENSE PERCEPTION and for the NP predictor – by THE PHYSICAL WORLD because the co-occurrence of these two categories is the most frequent in the data. The output estimates were automatically log-transformed by virtue of the logit link but, in Figure 18, they are exponentiated for better readability. The estimates below 1 are marked with red: they stand for the categories whose co-occurrence with the reference level of the other slot brings down the odds ratio of ΔP_{diff}. This kind of a relationship implies an increase in associative strength of the given NP category or a decrease in that of the given ADJ category. The estimates greater than 1, on the contrary, indicate that the corresponding categories increase the odds ratio of ΔP_{diff}, which is indicative of greater predictive power of the ADJ slot. For instance, the likelihood of observing an adjective from the domain of FOOD AND DRINK together with a NP from THE PHYSICAL WORLD increases the odds ratio of ΔP_{diff} by 2.6 times, compared to the reference level SENSE PERCEPTION. Put differently, the [____ as THE PHYSICAL WORLD] semantic configuration is more likely to be right-predictive. In contrast, an adjective from the domain POSSESSION in the very same semantic pattern decreases the odds ratio of ΔP_{diff} of the construction by 0.7, which then has a higher probability to be considered left-predictive.

It is obvious that most regressors exhibit the estimates smaller than 1, which implies that the construction is leaning towards the backward kind of lexical association. As far as the effect size is concerned, it is expressed for each of the regressor by the odds ratios themselves. Based on the estimates, there are more NP categories influencing the mean ΔP_{diff}, but it is a handful of ADJ categories that have a greater effect.

There are two NP coefficients that deserve attention: RELIGION AND BELIEF and THE BODY. The estimate of the former is not surprising, given a high frequency of *hell*-instances that deprive the NP slot of its predictive power. An increase of ΔP_{diff}, triggered by the vehicle in the [____ as THE PHYSICAL WORLD] semantic configuration, also suggests the connection between the basic human anatomy and embodied experiences, shaped in interaction with the world.

7.7. General Semantic Patterns

To this point, we have examined the semantic characteristics of the lexical contents in each slot separately. The most typical semantic configurations for the A *as* NP construction, with each slot mapped to a certain semantic domain, were obtained from the co-occurrence table of the semantic annotation tags. The most frequent patterns with examples are listed in (17).

(17) a)	[SENSE PERCEPTION as THE PHYSICAL WORLD]	white as snow black as tar light as the air dumb as a stick sharp as a thorn
b)	[EMOTIONS AND VALUES as ANIMALS]	stubborn as a mule gentle as lamb strong as a bull greedy as pig slim and smooth as a fish
c)	[SENSE PERCEPTION as THE BODY]	red as blood useful as tits on a bull clear as a nose on one’s face dark as shit
d)	[SENSE PERCEPTION as ANIMALS]	quiet as a mouse weak as kitten proud as peacock dumb as a fox happy as pig

Other most common co-occurrence patterns include [SENSE PERCEPTION as RELIGION AND BELIEF], [THE BODY as ANIMALS], [SENSE PERCEPTION as BASIC ACTIONS AND TECHNOLOGY], etc. According to Evans and Green, all these semantic categories (with the exception of animals) rest upon the notion of experiential grounding, or embodiment: they “derive from the pre-conceptual experience, such as sensory-perceptual experience, which forms the basis of more complex knowledge” (Evans and Green 2006: 232). Apart from SPACE and EMOTION, the basic domains proposed by Evans and Green also include COLORS and TEMPERATURE, which are also present in the data, although under the umbrella category SENSE PERCEPTION:

(18) a)	[SENSE PERCEPTION: COLORS as ____ (THE PHYSICAL WORLD / THE BODY / RELIGION AND BELIEF / etc.)]	green as grass black as ink blue as the sky red as a tomato pink as a kitten's tongue
b)	[SENSE PERCEPTION: TEMPERATURE as ____ (THE PHYSICAL WORLD / AGRICULTURE AND VEGETATION / RELIGION AND BELIEF / etc.)]	hot as hell hot as fire hot as blazes cold as ice cold as steel

The stock of semantic patterns is anything else but surprising. In fact, it accords with the prototypical semantic orientation of the A *as* NP construction where the tertium encodes the most salient propensity of the vehicle. It is a known fact that humans are good not only at making judgments about material classes or material properties of the objects surrounding them but also at categorizing them based on these qualities (Hiramatsu and Fujita 2015, Morgenstern et al. 2019). Considering that categories are commonly defined in current cognitive psychology in terms of prototypes and exemplars (Murphy 2002: §3–§4, Diessel 2019: 40), it becomes clear that the human disposition to assign physical objects to categories based on subjective judgments about their visual/auditory/material qualities is a domain-general process (cf. Lakoff 1987).

A large variety of literal meanings are extended to abstract concepts, for instance, human behavior or state of matters. In this case, we deal with the non-literal, or metaphorical, use of the construction. A typical example is the PEOPLE ARE ANIMALS conceptual metaphor, underlying the [EMOTIONS AND VALUES as ANIMALS] pattern in our data and used as a common way to link human behavior to the environment. Since animals have been part of the immediate environment of the human for centuries, the cognitive motivation of representing animals according to different roles we assign to them is not surprising. In fact, the phenomenon of anthropomorphism allows us to construe animals as embodiements of a specific character trait or emotion and enables the conceptual mapping between humans and animals. The PEOPLE ARE ANIMALS metaphor, therefore, involves a two-step process: attribution of human emotions or intentions to animals and the subsequent representation of human behavior in terms of the typical roles, assigned to animals (e.g., *peaceful as a dove, busy as a bee, fierce as a wolf, greedy as pig, courageous as a lion, angry as a bear*, etc.).

Another common non-literal use of the A *as* NP construction is the extension of the properties of physical objects or body-related experiences to feelings and emotions. A case in point is the conceptual mapping of the temperature scale onto the proximity of social relations (e.g., *the eyes as cold as ice*) and emotions (e.g., *love as warm as the summer rain*). Hard materials in English are endowed with coldness (e.g., *cold as steel/a stone/a rock*) and stupidity (e.g., *dumb/stupid as a rock, thick as a plank/two planks/*etc.), but also strength (e.g., *strong as iron/ a rock/wood/a stone*), determination (e.g., *tough as steel*), stability (e.g., *consistent/unshakable/steady as a rock*) and authenticity (e.g., *true as steel*). Paradoxical as it may sound, soft materials are not used in the non-literal meaning in this construction (cf. *soft as silk/wool/pillow*).

Overall, the relationship between the semantic categories and the association metrics for each slot in the English A *as* NP construction can be summed up as shown by Figure 20. For each semantic category in each slot, the mean LLR and ΔP_{diff} were computed, and the results were plotted in the already familiar two-dimensional plane LLR ~ ΔP_{diff} where ellipses were drawn around the mean scores for each semantic tag.

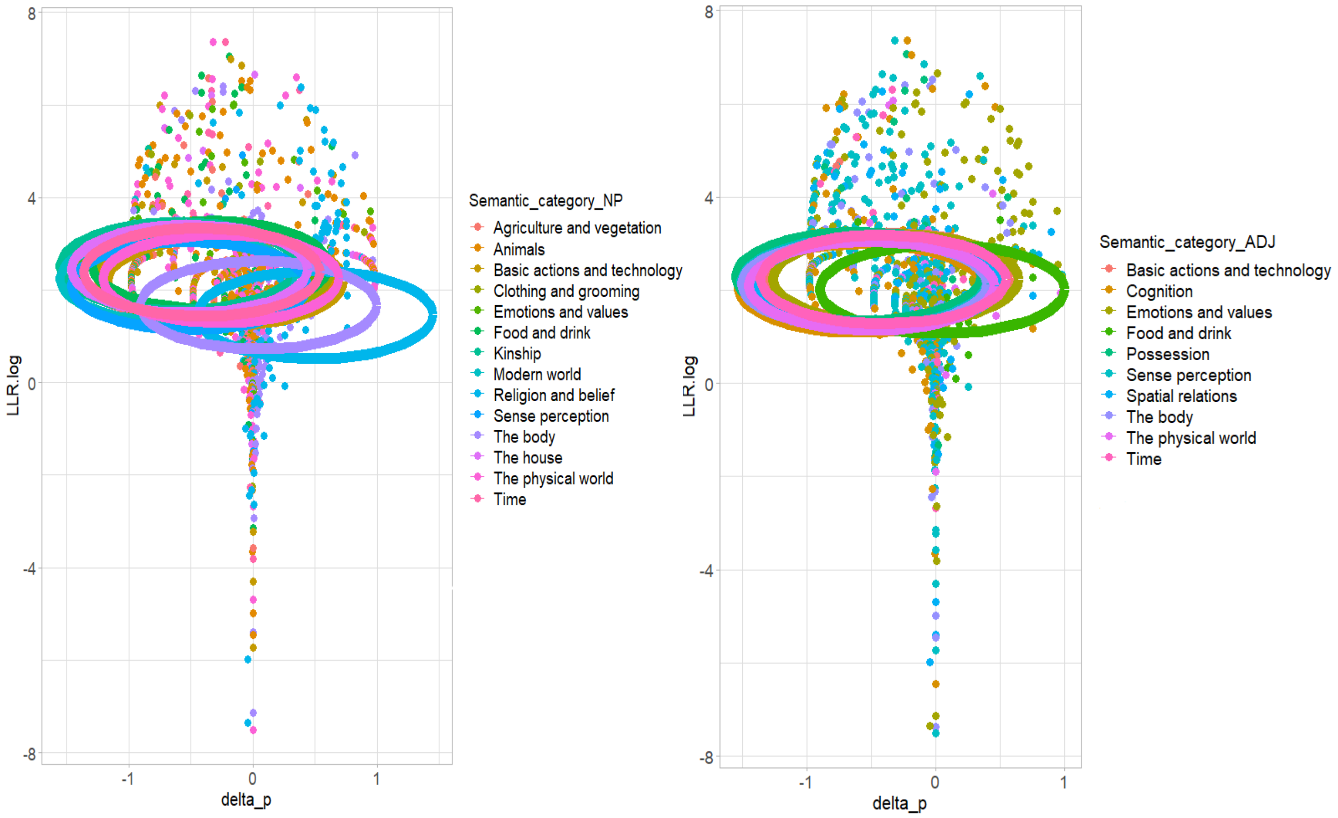


Figure 20. The mean ellipses of the most frequent semantic categories in the NP (left) and ADJ (right) slots, plotted against the log-transformed LLR (y-axis) and the ΔP_{diff} scores (x-axis).

In order to avoid overplotting, the low-frequency values were dropped. These plots provide visual evidence of the strong prevalence of the NP-schemas, related to the conceptual domains of the body and religion and showing low mutual attraction scores. To compare, the semantic affiliation of the adjectives do not seem to differ drastically in terms of symmetric and asymmetric association values; the exception is the A *as* NP instances where the tertium refers to various kinds of victuals and manifests greater predictive power than the NPs, co-occurring with them.

8. Conclusion

The current study has been focused on investigating the lexical association of collocates in the A *as* NP construction in English and Russian and attempted to characterize its semantic nature and potential development patterns. The notion of collocations has been adopted here in the Firthian sense so that all the attested instances of co-occurring lexical items in the construction are viewed as collocations. The analysis, carried out in the present study, has been implemented in two steps: first, we examined the association between the lexical components of the A *as* NP construction in both languages and then investigated the interaction of these measures with the semantics of the slots in the English construction.

The analysis of symmetric association has revealed a continuum of lexicalization between the items in the adjective and noun phrase slots. The instances with the strongest collocational strength can be identified as lexical prefabs, or idioms. They are located on the lowest level of abstraction of the A *as* NP construction taxonomy on account of their representation and activation in the mental lexicon as indivisible chunks. Furthermore, such instances demonstrate a high degree of relevance to the language-specific culture, which is especially conspicuous in Russian, abundant with folklore references. On the other hand, the A *as* NP construction with a looser association link between its lexical components constitutes a source for creative and innovative uses. Its difference from the institutionalized (idiomatic) instances lies in its relative semantic and morphological flexibility of the lexical items. Overall, the symmetric association measure, employed in the present study, reflects the hierarchy, or rather the gradient, of lexical sequences where the highest values reflect the highest level of automatization and conventionalization whereas the lowest values characterize the pairs with certain flexibility in lexical choice.

It has also been demonstrated that the concept of directional collocativity is indispensable to a comprehensive discussion on the topic of collocatons. Not only does the evidence provided here corroborate the psychological nature of lexical associations, but it also meshes well with the assumption of relatively high productivity of the A *as* NP construction. In this sense, an asymmetric association measure is a useful instrument for establishing 1) most productive A *as* NP subschemas and 2) the factors behind the sporadic loci of their extension to novel items. This method has been proved particularly informative for identification of potential grammaticalization patterns. In English, the apparent subjectification and semantic bleaching of the meaning in the____ *as hell/shit/fuck* subschemas are indicative of the ongoing grammaticalization of the intensification function. The Russian A *as* NP construction is noticeably less notorious in this respect, and its subschemas are rather constrained to a specific semantic field or a number of related semantic categories. In fact, the construction instances in Russian are semantically closer to the prototype because, unlike in English, they display a stronger motivational link between the tertium and the vehicle. In other words, the Russian A *as* NP construction is less idiosyncratic than that in English owing to its semantic proximity to the prototypical meaning of similies, in a broader sense.

The semantic analysis of the English construction has confirmed the general idea about the use of similies. Our findings are congruent with the evidence, indicative of the importance of embodied and sensory experiences for cognitive processes. Yet another prominent source domain for the metaphorical meaning of the construction is animal imagery and anthropomorphism, which is likewise suggestive of human cognitive tendencies that are quite likely to be observed cross-linguistically. It is also apparent that the semantic affiliation of lexical items, and especially noun phrases, has a considerable impact on the strength of association in the construction. Although the A *as* NP construction could be by and large characterized as left-predictive, there are two semantic domains that clearly dominate the scene in reversing this direction. The presence of the conventionally taboo nouns increases the likelihood of the construction to be right-predictive while, in the adjective slot, the function of reversing the direction of association is fulfilled by the domain of victuals. Similar to English, the Russian A *as* NP construction exhibits overwhelmingly backward association, and an analysis of a larger corpus size could potentially reveal subtle interactions between the semantics of certain lexical items and the association characteristics.

Although the morphological variation was ignored in the present study, we have no doubts that its inclusion in the analysis might yield more comprehensive results, especially in regard to a highly synthetic language, such as Russian. For instance, a preliminary and rather superficial inspection of the Russian data suggests that some noun phrases are more likely to be used in the construction in the diminutive form than others. From the typological perspective, the use of data from more, preferably, genetically unrelated, languages is intuitively likely to corroborate the assumption about cross-linguistic semantic similarities, which may have their roots in general cognitive processes. Even more, a future cross-linguistic analysis of the diachronic development of the A *as* NP construction might have a potential to disclose general grammaticalization patterns, specific to the expressions of comparison.

9. References

Agresti, A. (2002). *Categorical Data Analysis.* Hoboken: John Wiley & Sons.

Allan, L. G. (1980). "A note on measurement of contingency between two binary variables in judgment tasks." *Bulletin of the Psychonomic Society, 15 (3)*: 147-149.

Baayen, R. H. (1993). "On Frequency Transparency and Productivity." In: *Yearbook of Morphology 1992*. Eds. Booij, G. and Marle, J. van. Dordrecht, London: Kluwer. 181-208.

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Tübingen: Narr.

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic.* Amsterdam: John Benjamins.

Benson, M. (1989). "The Structure of the Collocational Dictionary." *International Journal of Lexicography, 2(1)*: 1-14.

Benson, M. and E. Benson and R.F. Ilson (1986). *Lexicographic Description of English.* Amsterdam: John Benjamins.

Biber, D. (2009). "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing." *International Journal of Corpus Linguistics 14(3)*: 275-311.

Biber, D. and S. Conrad (1999). "Lexical bundles in conversation and academic prose." In: *Out of corpora. Studies in honour of Stig Johansson.* Ed. Hasselgård, H. and S. Oksefjell. Amsterdam: Rodopi. 182-190.

Black, M. (1962). *Models and Metaphors.* Ithaca: Cornell University Press.

Breusch, T. S. and A. R. Pagan (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation". *Econometrica*, 47 (5): 1287-1294.

Bušta, J., and O. Herman. JSI Newsfeed Corpus. In The 9th International Corpus Linguistics Conference. Corpus Linguistics 2017 Conference, University of Birmingham.

Bybee, J. (2002). "Sequentiality as the basis of constituent structure." In: *The Evolution of Language out of Pre*-Language. Eds. Givón, T and F. M. Bertram. Amsterdam: John Benjamins. 109–132.

Bybee, J. (2010). *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.

Carter, R. (1998). *Vocabulary: Applied Linguistics Perspectives.* London: Routledge.

Chiappe, D. and J. M. Kennedy (2001). "Literal bases for metaphor and simile." *Metaphor and Symbol, 16(3–4):* 249-276.

Chiappe, D., J. M. Kennedy and T. Smykowski (2003). “Reversibility, aptness, and the conventionality of metaphors and similes.” *Metaphor and Symbol, 18(2*): 85-105.

Choueka, Y. (1988). “Looking for needles in a haystack.” Proceedings, RIAO Conference on User-oriented Context Based Text and Image Handling. Cambridge. 609-623.

Church, K. and R. L. Mercer (1993). “Introduction to the special issue on computational linguistics using large corpora.” *Computational Linguistics, 19(1*): 1-24.

Clausner, T. C., and W. Croft (1997). “Productivity and schematicity in metaphors.” *Cognitive Science 21(3)*: 247-282.

Cribari-Neto, F. and A. Zeileis (2010). “Beta Regression in R.” *Journal of Statistical Software*, 34(2): 1–24.

Daudaravičius, V., and R. Marcinkevičienė (2004). “Gravity counts for the boundaries of collocations.” *International Journal of Corpus Linguistics, 9(2):* 321-348.

Davies, M. (2013). Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). Available online at: <https://www.english-corpora.org/glowbe/> 10 May 2020.

De Smedt, T. and W. Daelemans (2012). “Pattern for Python.” *Journal of Machine Learning Research*, 13, 2031–2035.

Desagulier G. (2016). “A lesson from associative learning: Asymmetry and productivity in multiple-slot constructions.” *Corpus Linguistics and Linguistic Theory 12 (2*): 173-219.

Diessel, H. (1999). *Demonstratives: Form, Function, and Grammaticalization.* [Typological Studies in Language. 42]. Amsterdam: John Benjamins.

Diessel, H. (2019). *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge: Cambridge University Press.

Dilts, P. (2010). “Good nouns, bad nouns: what the corpus says and what native speakers think”. In *Corpus-linguistic applications*. Eds. S. Th. Gries, S. Wulff and M. Davies. Leiden: Brill | Rodopi. 103-117.

Dunning, Ted (1993). “Accurate methods for the statistics of surprise and coincidence.” *Computational Linguistics 19(1):* 61-74.

Ellis, N. (2006). “Cognitive perspectives on SLA: The associative-cognitive CREED.” *AILA Review, 19:* 100-121.

Ellis, N., and F. G. Ferreira-Junior (2009). “Constructions and their acquisition: Islands and the distinctiveness of their occupancy.” *Annual Review of Cognitive Linguistics, 7:* 188-221.

Erman, B and B. Warren (2000). “The Idiom Principle and the Open Choice Principle.” *Text, 20*: 29-62.

Evans, V. and M. Green. (2006). *Cognitive Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

Evert, S. (2005). The Statistics of Word Co-Occurrences: Word Pairs and Collocations. Ph.D. thesis, Stuttgart: University of Stuttgart.

Evert, S. (2009). “Corpora and collocations.” In: *Corpus Linguistics. An International Handbook.* Eds. A. Lüdeling and M. Kytö. Berlin: Mouton de Gruyter. 1212-1248.

Evert, S. and B. Krenn (2001). “Methods for the qualitative evaluation of lexical association measures.” Proceedings of 39th Annual Meeting of the Association for Computational Linguistics, 188–195.

Evert, S., P. Uhrig, S. Bartsch and T. Proisl (2017). “E-VIEW-Alation – a Large-Scale Evaluation Study of Association Measures for Collocation Identification.” In: *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2017 Conference*. Eds. Iztok, K., T. Carole, J. Miloš, K. Jelena, K. Simon B. Vít . Leiden, Brno: Lexical Computing. 531-549.

Espinheira, P. L., S. L. Ferrari and F. Cribari-Neto (2008). “Influence diagnostics in beta regression.” *Computational Statistics & Data Analysis*, 52(9): 4417-4431.

Fernando, C. (1996). *Idioms and Idiomaticity.* Oxford: Oxford University Press.

Ferrari, S. L. P. and F. Cribari-Neto (2004). “Beta Regression for Modelling Rates and Proportions.” *Journal of Applied Statistics*, 31(7): 799-815.

Ferrari, S. L., P. L. Espinheira and F. Cribari-Neto (2011). “Diagnostic tools in beta regression with varying dispersion.” *Statistica Neerlandica*, 65(3): 337-351.

Fillmore, C. (1988). “The mechanisms of „Construction Grammar‟.” *Berkeley Linguistic Society, 14*: 35–55.

Firth, J. R. (1957). *Modes of Meaning*. Papers in Linguistics, 1934-1951. Oxford: Oxford University Press.

Fox, J. and S. Weisberg (2019). An R Companion to Applied Regression, 3rd edition. Sage, Thousand Oaks CA. Available online at: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/> 14 September 2020.

Garcia, M., M. García-Salido and M. Ramos (2019). “A comparison of statistical association measures for identifying dependency-based collocations in various languages.” *MWE-WN@ACL 2019*: 49-59

Giacinti, F (2019). “Liscio come l'olio. Un'analisi corpus-based del pattern Adj-come-NP in italiano.” *CLUB Working Papers in Linguistics, 3 (1)*: 69-92.

Glucksberg, S. (2001). *Understanding Figurative Language: From Metaphors to Idioms.* Oxford University Press, Oxford.

Goldberg, A. E. (2006). *Constructions at Work.* Oxford: Oxford University Press.

Goldberg, A. E. (2013). "Constructionist approaches." In: *Handbook of Construction Grammar*. Eds. Hoffmann T. and G. Trousdale. Oxford: Oxford University Press. 15-31.

Goldberg, A. E. (2016). "Partial productivity of linguistic constructions: Dynamic categorization and statistical preemption." *Language and Cognition, 8(3):* 369-390.

Greaves, C. and M. Warren (2010). "What can a corpus tell us about multi-word units." In: *The Routledge Handbook of Corpus Linguistics.* Ed. McCarthy, M. and A. O'Keeffe. Abingdon: Routledge. 212-226.

Gregory, M. L., W. D. Raymond, A. Bell, E. Fosler-Lussier and D. Jurafsky (1999). "The effects of collocational strength and contextual predictability in lexical production." Proceedings of the 35th annual Chicago Linguistic Society. Chicago.

Gries, S. Th. (2007). Coll.analysis 3.2a. A program for R for Windows 2.x. Available online at: <http://www.stgries.info/teaching/groningen/index.html> 17 September 2020.

Gries, S. T. (2013). "50-something years of work on collocations: What is or should be next…" *International Journal of Corpus Linguistics, 18(1):* 137-165.

Gries, S. Th. And A. Stefanowitsch (2004a). "Extending collostructional analysis: A corpus-based perspectives on 'alternations'." *International Journal of Corpus Linguistics 9 (1)*: 97-129.

Gries, S. Th. And A. Stefanowitsch (2004b). "Co-varying collexemes in the into-causative." In: *Language, Culture, and Mind*. Eds. Achard, M. And S. Kemmer. Stanford: CSLI. 225-236.

Hanks, P. (2005). "Similes and sets: the English preposition "like"." In: *Languages and Linguistics: Festschrift for Professor F. Čermák*. Eds. R. Blatná and V. Petkevič. Prague: Philosophy Faculty, Charles University.

Haspelmath, M., and O. Buchholz (1998). "Equative and similative constructions in the languages of Europe." In: *Adverbial constructions in the languages of Europe*. Ed. Auwera J. v. D. Berlin: de Gruyter. 277-334.

Hausmann, F. J. (1989). *Le dictionnaire de collocations. In In Wrterbcher, Dictionaries, Dictionnaires. Ein internationales Handbuch*. Berlin: De Gruyter.

Hausmann, F. J. (2003). "Was sind eigentlich Kollokationen?" In: *Wortverbindungen – mehr oder weniger fest.* Ed. K. Steyer. Berlin: Walter de Gruyter. 309-334.

Hiramatsu, Ch. and K. Fujita. (2015). "Visual categorization of surface qualities of materials by capuchin monkeys and humans." *Vision research*, 115: 71-82.

Hoang, H.H., S. Kim and M. Kan (2009). “A re-examination of lexical association measures.” Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE '09). Association for Computational Linguistics, 31-39.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hoffmann, S., and S. Evert (2006). “BNCweb (CQP edition) - the marriage of two corpus tools.” In: *Corpus technology and language pedagogy: new resources, new tools, new methods*. Eds. Braun, S., K. Kohn and J. Mukherjee. Frankfurt am Main: Peter Lang. 177-195.

Hudson, R. (2016). *The Christian Writer's Manual of Style: 4th Edition*. Zondervan.

Jenkins, H. M. and W. C. Ward (1965). “Judgement of contingency between responses and outcomes.” *Psychological Monographs, 79(1):* 1-17.

Jones, S. and J. Sinclair (1974). “English lexical collocations: A study in computational linguistics.” *Cahiers de Lexicologie 24(1)*: 15-61.

Kapetanios, E., S. Alshahrani, A. Angelopoulou and M. Baldwin (2018). “What Do We Learn from Word Associations? Evaluating Machine Learning Algorithms for the Extraction of Contextual Word Meaning in Natural Language Processing.” Preprint.

Kenny, D. (2000). “Lexical hide-and-seek: Looking for creativity in a parallel corpus.” In: *Intercultural faultlines. Research models in translation studies I. Textual and cognitive* aspects. Ed. Olohan, M. Manchester: St. Jerome Publishing. 93-104.

Kjellmer, G. (1991). “A mint of phrases.” In: *English Corpus Linguistics.* Eds. K. Aijmer and B. Altenberg,. New York: Longman. 111-127.

Kjellmer, G. (1994). *A Dictionary of English Collocations.* Oxford: Clarendon Press.

Koenker, R. (1981). "A Note on Studentizing a Test for Heteroscedasticity". *Journal of Econometrics,* 17: 107-112.

Koplenig, A. (2018). “Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis.” *Corpus Linguistics and Linguistic Theory*, 14(1): 1-34.

Kövecses, Z. (2000). *Metaphor and Emotion.* Cambridge: Cambridge University Press.

Kövecses, Z. (2002). *Metaphor. A Practical Introduction.* Oxford: Oxford University Press.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago: Chicago University Press.

Langacker, R. W. (2006). “Subjectification, Grammaticization, and Conceptual Archetypes.” In: *Subjectification: Various Paths to Subjectivity*. Eds. A. Athanasiadou, C. Canakis, and B. Cornillie. Berlin and New York: Mouton de Gruyter. 17-40.

Langacker, R. W. (2008). *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.

Lea, D. (2002). *Oxford Collocations Dictionary for Students of English.* Oxford: Oxford University Press.

Leech, G. (1974). *Semantics: The Study of Meaning*. Harmondsworth: Penguin Books.

Lebedeva, L. A. (2017). *Ustojchivye sravnenija russkogo yazyka: Kratkij tematicheskij slovar'* [Russian fixed similes: A short thematic dictionary]. Moscow: Flinta.

Levshina, N. (2014). Rling: A companion package for how to do linguistics with R. R package version 1.0. Available online at: <https://benjamins.com/sites/z.195/> 19 October 2020.

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Lin, D. (1999). “Automatic identification of noncompositional phrases.” Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 317-324.

List, J.-M. "Exporting Sublists from a Wordlist with LingPy and Concepticon," in Computer-Assisted Language Comparison in Practice. Available online at: <16/07/2018, https://calc.hypotheses.org/58> 27 September 2020.

List, J.-M., Ch. Rzymski, S. Greenhill, N. Schweikhard, K. Pianykh, A. Tjuka, M.-S. Wu, R. Forkel (2020). Concepticon 2.3.0. Jena: Max Planck Institute for the Science of Human History. Available online at <http://concepticon.clld.org> 19 September 2020.

MacWhinney B., J. Leinbach, R. Taraban and J. McDonald (1989). “Language learning: Cues or rules?” *Journal of Memory and Language, 28*: 255-277.

Makkai, A. (1972). *Idiom Structure in English.* The Hague: Mouton.

Manning, C. and H. Schuütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Mauranen, A. (2000). “Strange strings in translated language: A study on corpora.” In: *Intercultural Faultlines Research Models in Translation Studies.* Ed. Olohan, M. Manchester: St. Jerome Publishing. 119-141.

McCarthy, M. (1998). *Spoken Language and Applied Linguistics.* Cambridge: Cambridge University Press.

Mel’cuk, I. (1998). “Collocations and lexical functions.” In: *Phraseology. Theory, analysis and applications.* Ed. Cowie, A. P. Oxford: Clarendon Press. 23-53.

Michelbacher, L., S. Evert and H. Schütze (2007). "Asymmetric Association Measures." Proceedings of the International Conference on Recent Advances in Natural Language Processing. Borovets.

Michelbacher, L., S. Evert and H. Schütze (2011). “Asymmetry in corpus-derived and human word associations.” *Corpus Linguistics and Linguistic Theory, 7*: 245-276.

Mokienko, V. M. (2016). *Ustojchivye sravnenija v sisteme frazeologii* [Fixed similes in the phraseological system]. Saint Petersburg/Greifswald: LEMA.

Monti, J., V. Seretan, G. C. Pastor and R. Mitkov (2018). “Multiword units in machine translation and translation technology.” In: *Multiword units in machine translation and translation technology (Current issues in linguistic theory, 341*). Eds. R. Mitkov, J. Monti, G. C. Pastor, and V. Seretan. Amsterdam and Philadelphia: John Benjamins.

Moore, R. (2004). “On Log-Likelihood-Ratios and the Significance of Rare Events.” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 333-340.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-based Approach.* Oxford: Oxford University Press.

Moon, R. (2008). “Conventionalized as-similes in English A problem case.” *International Journal of Corpus Linguistics, 13:* 3-37.

Morgenstern, Y. and F. Schmidt and R. Fleming. (2019). “One-shot categorization of novel object classes in humans.” *Vision research*, 94: 62-75.

Murmann, M. (2019). *Inchoative Emotion Verbs in Finnish: Argument Structures and Collexemes*. Tübingen: Gunter Narr Verlag.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nelson, D.L., C. L. McEvoy and T. A. Schreiber (2004). “The University of South Florida free association, rhyme, and word fragment norms.” *Behavior Research Methods, Instruments, and Computers 36:* 402-407.

Norrick, N. (1986). “Stock similes.” *Journal of Literary Semantics, XV (1):* 39-52.

Omazić, M. (2002). “O poredbenom frazemu u engleskom i hrvatskom jeziku.” *Jezikoslovlje, 3(1–2)*: 99-129.

Onnis, L., and E. Thiessen (2013). “Language experience changes subsequent learning.” *Cognition, 126(2):* 168-284.

Ortony, A. (1993). “The role of similarity in similes and metaphors.” In: *Metaphor and Thought.* Ed. Ortony, A. Cambridge: Cambridge University Press. 342-356.

Parizoska, J. and I. Filipović Petrović. (2017). “Variation of Adjectival Slots in kao (‘as’) Similes in Croatian: A Cognitive Linguistic Account.” International Conference on Computational and Corpus-Based Phraseology: 348-362.

Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Pecina, P. (2010). “Lexical association measures and collocation extraction.” *Lang Resources & Evaluation 44:* 137-158.

Pecina, P. and P. Schlesinger (2006). “Combining association measures for collocation extraction.” Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 651-658.

Pianykh, K. (2019). Productivity of the A as NP construction in Russian and English: clear as mud or clear as day? Unpublished term paper.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton and Ch. D. Manning (2020). “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.” Association for Computational Linguistics (ACL) System Demonstrations.

Saffran, J., E. Newport and R. N. Aslin (1996). “Word segmentation: The role of distributional cures.” *Journal of Memory and Language, 35:* 606-621.

Sag, I. A., T. Baldwin, F. Bond, A. A. Copestake and D. Flickinger (2002). “Multiword expressions: A pain in the neck for NLP.” Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing, 2276/2010, 1-15.

Schneider, U. (2018). “ΔP as a measure of collocation strength. Considerations based on analyses of hesitation placement in spontaneous speech.” *Corpus Linguistics and Linguistic Theory, 16(2):* 249-274.

Seretan, V. (2003). Syntactic and Semantic Oriented Corpus Investigation for Collocation Extraction, Translation and Generation. Ph.D. thesis. Language Technology Laboratory, Department of Linguistics, Faculty of Arts, University of Geneva.

Seretan, V. (2018). “Bridging Collocational and Syntactic Analysis.” In: *Lexical Collocation Analysis. Quantitative Methods in the Humanities and Social Sciences*. Eds. Cantos-Gómez, P., M. Almela-Sánchez M. Cham: Springer.

Sinclair, J. (1987). “Collocation: A progress report.” In: *Language Topics: Essays in Honour of Michael Halliday.* Ed. Steele, R. and T. Threadgold. Amsterdam: John Benjamins. 319-331.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1995). *Collins COBUILD English Dictionary*. London: Harper Collins.

Simas, A. B. and A. V. Rocha (2006). betareg: Beta Regression. R package version 1.2. Available online at: <http://CRAN.R-project.org/src/contrib/Archive/betareg/> 5 October 2020.

Smithson, M and J. Verkuilen (2006). "A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables." *Psychological Methods,* 11(1): 54-71.

Stubbs, M. (1996). "Collocations and semantic profiles: On the cause of the trouble with quantitative studies." *Functions of Language, 1:* 23-55.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stefanowitsch, A. and S. Th. Gries (2003). "Collostructions: Investigating the interaction between words and constructions." *International Journal of Corpus Linguistics 8(2*): 209-243.

Stefanowitsch, A. and S.T. Gries (2009) "Corpora and grammar." In: *Corpus Linguistics. An International Handbook*. Eds. A. Lüdeling and M. Kytö. Berlin and New York: Mouton de Gruyter. 933-951.

Rescorla, R. A. (1968). 'Probability of shock in the presence and absence of CS in fear conditioning,' *Journal of Comparative and Physiological Psychology, 66*: 1-5.

Rocha, A. V. and A. B. Simas (2011). "Influence diagnostics in a general class of beta regression models." *Test*, 20(1): 95-119.

van der Ploeg, T., P. C. Austin and E. W. Steyerberg (2014). "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints." *BMC Med Res Methodol*, 14 (137).

Vejdemo, S. and S. Vandewinkel (2016). Extended uses of body-related temperature expressions. In: The lexical typology of semantic shifts. Eds. Juvonen, P. and M. Koptjevskaja-Tamm. Berlin: de Gruyter Mouton.

Wahl, A. (2015). "Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures." *Review of Cognitive Linguistics 13(1):* 191-219.

Waskom, M., O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, B. Fonnesbeck, C. Fonnesbeck, A.

Lee and A. Qalieh (2017). mwaskom/seaborn: v0.8.1 (September 2017), Zenodo. Available at: <https://doi.org/10.5281/zenodo.883859> 15 October 2020.

Ward, W. C. and H. M. Jenkins (1965). “The display of information and the judgement of contingency.” *Canadian Journal of Experimental Psychology, 19(3):* 231-241.

Watson Todd, R. (2019). “Exploring the direction of collocations in eight languages.” *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 64(1): 146-154.

Wehrli, E, and L. Nerima (2015). “The Fips multilingual parser. Festschrift in honour of Michael Zock.” In: Language production cognition, and the lexicon. Eds. N. Gala, R. Rapp and G. Bel-Enquix. Springer. 473-489.

Wehrli, E. and L. Nerima (2018). “Anaphora resolution, collocations and translation.” *Current Issues in Linguistic Theory, 341*: 244-256.

Wiechmann, D. (2008). “On the Computation of Collostruction Strength: Testing Measures of Association as Expressions of Lexical Bias.” *Corpus Linguistics and Linguistic Theory 42*: 253-290.

Zeileis, A. and T. Hothorn (2002). “Diagnostic Checking in Regression Relationships.” *R News*, 2(3): 7-10.

Zeldes, A. (2012). *Productivity in Argument Selection: From Morphology to Syntax*. Berlin: Walter de Gruyter.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge: Addison-Wesley.

Statement of authorship

I hereby confirm that I wrote this master’s thesis on my own and I did not use any other aids or sources, except those indicated. I furthermore confirm that I did not submit the thesis as assessed course work elsewhere nor it was published in German or any other language.

The author has objections to making the present master’s thesis available to the public.

11.12.2020, Jena