Lexical idioms: what is regular and what anomalous in word-formation[1]

Kateřina Vašků (Praha)

ABSTRACT

Traditionally, phraseology deals with multi-word lexical units. However, mentions of idiomatic compounds and derivatives, i.e. lexical idioms, are far from scarce in linguistic literature even though their only systematic treatment appears to have been presented by Čermák (2007a, b). The present paper attempts to analyse complex words in English to explore the relation between the regular and the idiomatic in word-formation. The study is carried out on a sample of 1000 lexemes from BNC. The sample contains 681 complex lexemes of which 381 are regular and 300 contain one or more anomalies. The results point to an unexpectedly high number of formal anomalies and this leads to the conclusion that non-productivity should not be a criterion of idiomaticity in English due to a high number of relatively systematic, but unproductive structures. Idiomaticity is seen as a scalar value influenced by the degree of opaqueness, discrepancy between word-formation and lexical meaning, and presence of formal or collocational anomaly.

KEYWORDS

phraseology, idiom, anomaly, word-formation, semantics

DOI

https://doi.org/10.14712/18059635.2019.1.2

1. INTRODUCTION

Phraseology has traditionally been concerned with multi-word lexemes (cf. Cowie 1998; Burger 1998), focusing especially on fixedness and semantic opacity of idioms, which contrast with regular, non-idiomatic, word-combinations. The present paper examines to what degree it is possible to transfer the concept of idiomaticity below its traditional domain, exploring plausibility of the concept of lexical idiom in English word-formation.

The aim of the present study is to identify what should be seen as regular and what as idiomatic in English word-formation. On a sample of lexemes retrieved from the BNC we will inspect various manifestations of anomaly in complex words in English and attempt to deal with problematic areas peculiar to English vocabulary. In addition, the study describes similarities and differences between the traditional multi-word idioms and lexical idioms.

The focus on the contrast between the regular and the idiomatic in word-formation may be useful for general linguistics, inasmuch as searching for analogies between the lexical and phrasal structure is in accordance with the tendency of the

1 This study was supported by the Charles University project Progres Q10, Language in the shiftings of time, space, and culture.

current cognitive approach to linguistics to blur the lines between levels of language description. In addition, the concept of lexical idioms as anomalous, unexpected combinations of morphemes, may be also beneficial for applied linguistics, especially language acquisition and ELT. So far, there is no practical manual or dictionary which would present idiomatic, and therefore potentially "tricky", words for language users.

2. THEORETICAL BACKGROUND

The study follows the general phraseological literature of the continental and British "phraseological approach", including Burger (1998), Cowie (1998) and Howarth (1998), who all describe phraseology as the study of word-combinations. Idiomaticity is seen by these authors as a scalar phenomenon encompassing mainly the degree of semantic opaqueness, but also the degree of formal fixedness. Table 1 presents Howarth's collocational continuum (1998: 28), which can serve as a representative of these traditional approaches:

	free combinations	**restricted collocations**	**figurative idioms**	**pure idioms**
lexical composites verb + noun	*blow a trumpet*	*blow a fuse*	*blow your own trumpet*	*blow the gaff*
grammatical composites preposition + noun	*under the table*	*under attack*	*under the microscope*	*under the weather*

TABLE 1. Howarth's collocational continuum

As can be seen in Table 1, Howarth describes word-combinations on the continuum from formally free and semantically transparent *free combinations*, through formally less free but semantically still transparent *restricted collocations*, followed by semantically irregular, but still fairly transparent, *figurative idioms*, ending in *pure idioms*, which are both formally fixed and semantically opaque.

The study of irregular combinations on the lexical level reaches back to the 1970's when Dokulil (1978) presented his account of meaning predictability in complex words, which was later developed by Štekauer (2005). Nevertheless, it was Čermák (2007a, b) who first described lexical idioms systematically as a part of phraseology. In contrast with Howarth's classification of collocations, Čermák (2007a) focuses on idiomatic structures only; non-idiomatic structures are not included in the classification. However, the scope of units described within phraseology by Čermák is roughly the same as that of Howarth. This is because Čermák defines idioms based on a combination of anomalous properties, including not only semantics, but also formal anomalies and anomalies in collocability. Čermák (2007a, b[2]) is the main theoretical

2 Čermák (2007b) is also available as a chapter in Čermák (2007a).

source of the present study because he explicitly includes lexical idioms into his classification of phrasemes and idioms,[3] describing the field of lexical idioms as

> "a large area of idiomatic combinations of morphemes [...], neglected by those who are obsessed with (seemingly) free forms only, choosing not to see the very same phenomenon in compounds and elsewhere and ignoring an obvious link between higher levels and the level of morpheme combinations. In general, the identical content may be rendered either as a combination of separate forms (*split hairs, cut corners*) or of morhemes inside a lexical idiom, i.e. a single-word lexeme (*hair-splitting*, *corner-cutting*). This is basically and broadly an Europocentric view, building on familiar isolating, inflectional or agglutinative language types and disregarding polysynthetic languages, where, i.e. inside their large incorporational constructions within the scope of a single textual word, idioms should not, in this view, exist." (Čermák 2007b: 20)

Čermák focuses especially on Czech lexical idioms, presenting his own classification based on their morphological structure, i.e. the word-formation process involved and the resultant word-class. Accordingly, there are derivational and compositional phrasemes which further subdivide into four main classes: nominal, adjectival, verbal and adverbial. This is illustrated in Figure 1.

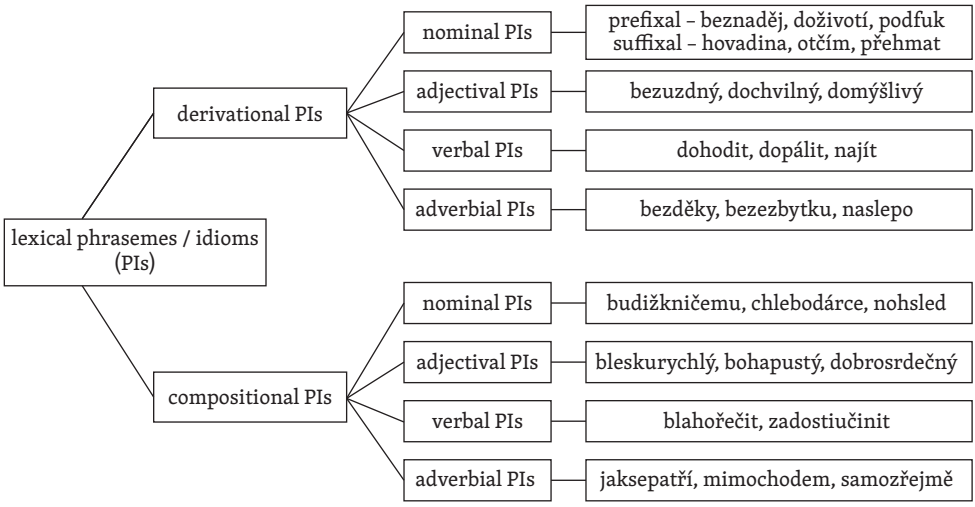

FIGURE 1: Morphological classification of lexical phrasemes and idioms according to Čermák (2007a)

3 The phraseological terminology is unfortunately rather unsettled, and the basic unit has various designations in works of individual authors (*phraseological unit, phraseme, phraseology*, etc.). Čermák uses the term *phraseme* when referring to the formal properties of the unit and *idiom* when referring to semantics of the unit (cf. Čermák 2007a: 33); for simplification, he often uses the designation *phraseme and idiom* when discussing the basic unit of phraseology generally.

Čermák (2007a, b) also mentions some examples of lexical idioms from other languages than Czech, pointing out that „[t]he existence of idioms is most probably universal [...], although its specific manifestation is due to the typological character of the language in question" (2007b: 20). He mentions several problems with identification of lexical idioms. Firstly, it is rather difficult to set a clear line between the assumed usual meaning of a given morpheme and its idiomatic use. He considers the most frequent meaning found in the dictionary to be the prototypical one. However, with idiomaticity being a scalar phenomenon, this approach does not take into account differences between less frequent but still fairly common senses and those which are clearly idiosyncratic. Secondly, Čermák explains that he does not include such words as *jackpot* among lexical idioms referring to such words as depletive, claiming that "[t]he crux of the matter is in difficulty to discern between a very large number of meanings (dictionary senses) and a loss of meaning (depletion). In this light, *jackpot* (with 14 meanings for *jack* and 6 for *pot*, according to *The New English Dictionary of English*) represents a border-line case of sorts, the number of meanings for *jack* being just too large and, primarily, diverse. It seems, namely, that this type of coinage should refer to a basic, prototypical meaning of a constituent if viewed independently, this being, on the other hand, difficult to find for *jackpot*" (Čermák 2007a: 23–24). Thirdly, there is a question of the degree to which one should take into account the diachronic perspective. Čermák claims that synchronic analysis has only limited value as it is problematic to distinguish between the synchronic and diachronic relations without complete data of the whole vocabulary. He therefore concludes that we need to base our analysis on intuition and estimation based on synchronic data (Čermák 2007a: 265).

The practical analysis of Czech lexical idioms is performed by Klötzerová (1997, 1998). It is based on data retrieved from a Czech dictionary and the analysis systematically describes data of the whole dictionary. The study follows the concept of phraseology developed by Čermák, for whom the basic property of (lexical) idioms is their (multiple) anomaly. The presence of anomaly is considered to be the main criterion for idiomaticity. The features taken into account are the range of paradigm, productivity, and the substitution and transformation tests. These tests are applied in accordance with Čermák's definitions. Idiomaticity is treated as a graded phenomenon. Klötzerová works with the terms centre (prototypical instances of the studied phenomenon) and periphery (less prototypical instances of the studied phenomenon, cf. Vachek 1966) and distinguishes between central and peripheral lexical idioms. She identifies approximately 5% of lexical units in the Czech dictionary *Slovník spisovné češtiny* (Filipec et al. 1994) as lexical idioms. Lexical idioms are then categorized according to their nature as compositional idioms, prefixal idioms, valency idioms and reflexive idioms. Klötzerová focuses on morphological classification of lexical idioms and the assignment of a lexeme into the category of idioms seems to be based mostly on intuition. The analysis of lexical idioms based on the complete data of a dictionary is really unique and no other language has been analysed in this range.

A recent contribution to the study of idomaticity on the lexical level is presented by Kos (2018) who follows the model of onomasiological categories developed by Dokulil (1986) and Štekauer (1998) and examines idiomatic nature of naming units in the mutational category.

3. COMPOUNDS AND DERIVATIVES SEEN AS IDIOMS OUTSIDE PHRASEOLOGICAL LITERATURE

Although phraseological literature generally focuses on word combinations only, there are in fact reasons to assume that linguists often intuitively broaden the scope of phraseology to include single-word units too. The fact is that in the literature not dealing with phraseology it is paradoxically not difficult to find mentions of both idiomatic derivatives and compounds. Mentions of idiomatic compounds are far more common, which is probably due to their status closer to phrases in combining (at least) two lexical stems. The following paragraphs present selected mentions of lexical compounds and derivatives by authors with various theoretical and language backgrounds. The presented sample of literature is far from complete and it should serve to illustrate how much the contrast of the regular and the idiomatic on the lexical level is actually spread[4]. The following paragraph presents a study which is concerned with both idiomatic derivatives and compounds and then presents chronologically works dealing with idiomatic compounds and finally works dealing with idiomatic derivatives. If not stated otherwise, the studies focus on idiomaticity in English.

Rodriguez and Rio-Torto's study (2013) compares several types of both derivatives and compounds in Portuguese with the aim to explore the way meaning construction occurs in derivation and compounding. It addresses three questions: how do derivatives and compounds get their meaning, which factors are involved and, most importantly for my study, does the semantics of derivative and compounds constituents follow the same rules? It also acknowledges that both derived words and compounds may have compositional and idiomatic meanings, i.e. meanings either computable or not computable from the meaning of their constituents.

Idiomatic compounds are mentioned for example in Levi (1978), who examines complex nominal structures with noun–noun nominal compounds (*apple cake, doghouse*) as one subgroup. She introduces a "continuum of derivational transparency" (p. 63) for compounds, with completely transparent compounds at one end (*mountain village*), followed by less transparent ones (*briefcase*, *polar bear*), and the third group consisting of exocentric compounds which include the most opaque cases and compounds that are partially or wholly idiomatic (e.g. *flea market* and *honeymoon*).

Lipka (2002: 95) says: "A complex lexeme may be synchronically analysable but no longer motivated, like *blackboard* or *watchmaker*. If its complete meaning is not derivable from its morphological structure and the pattern exhibited in parallel formations, as in *callgirl, highwayman, streetwalker, pushchair, wheelchair,* we say that such lexical items are idiomatic."

Katamba and Stonham (2006: 74) in their monograph *Morphology* explain the rise of idiomatic compounds, identifying two causes of idiomaticity, non-adherence to standard rules of word-formation, noticing that "[n]o synchronic rules can be devised to account for the meaning of a semantically unpredictable compound like

4 It should be added that mentions of irregularity on the lexical level are not limited to English. Such mentions can be found across languages. The present paper focuses on English mentions to simplify the matter as they serve quite well to illustrate the point.

stool pigeon" and metaphorical extension, accounting for compounds which originally had a literal meaning that was replaced by later metaphorical extensions (e.g. *redlegs, deadline*).

Benczes (2005, 2006 and 2015) deals with noun + noun compounds and bases her analysis on the theory of conceptual metaphor and metonymy (Lakoff and Johnson 1980). Benczes (2006) subsumes both endocentric and exocentric compounds under the concept "creative compound" which refers to metaphorical and metonymical compounds alike since as she points out even endocentric compounds (*armchair, handwriting*) very much like the exocentric *hammerhead* are creative compounds that involve metaphor and metonymy and require the use of creative imaginative, associative processes to be understood. However, these three compounds differ in the degree of creativity they involve and in addition to being "lexicalised to various degrees, a noun–noun combination such as *hammerhead* can be considered to be more creative than *armchair* or *handwriting* in the sense that a greater effort is required from the listener to understand its meaning" (Benczes 2006: 187–188). In short, what others call idiomatic compounds Benczes (2006: 184) regards as a part of the spectrum of creative compounds which "are not unanalysable, nor semantically opaque: in fact, they can be analysed within a cognitive linguistic framework, by the combined application of metaphor, metonymy, blending, profile determinacy and schema theory".

Kavka (2009) devotes one whole chapter of *The Oxford Handbook of Compounding* (Lieber and Štekauer 2009) to compounds from a phraseological point of view, in which he assesses "the relationship between compounds and idioms, arguing that both exhibit a gradience from mildly to wildly idiosyncratic interpretation that begs us to consider them together" (Kavka 2009: 18).

As far as mentions of idiomatic derivatives are concerned, they range from an occasional remark to a more systematic type of treatment. Kastovsky (1982) distinguishes between *systematic lexicalization*, such as in the regular addition of very general features such as [+PROFESSIONAL] in derivations by means of *-er* (*lecturer*, *reporter*, *writer*), and non-systematic, i.e. truly idiomatic semantic lexicalization.

Beard (1987) introduces the term *semantic drift* that affects items stored in the lexicon in both systematic and random ways. The drift resulting in semantic irregularity may start from primary transparent meanings (such as seen in *construction*, *painting*) or an idiomatic meaning may be subsequently attached to the output of a regular process (as in *transmission* "gearbox"). An affected item "disengages from the productive L-derivation rule which generates it" (Beard 1987: 26) and becomes listed in the lexicon.

Haspelmath (2002: 74–75) explicitly discusses idiomaticity of derivatives: "[w]e can distinguish two kinds of idiomaticity. In weak idiomaticity, the semantic contribution of the derivation is present, but the meaning of the derived lexeme is not exhaustively described by the base meaning and the derivational meaning. [...] In strong idiomaticity, the regular derivational meaning is not present at all, and the meaning of the derived lexeme cannot even be guessed from the meanings of the components."

Lieber (2009: 63) writes: "Hand in hand with the notion of transparency comes the related notion of lexicalization. When derived words take on meanings that are not transparent — that cannot be made up of the sum of their parts — we say that

the meaning of the word has become lexicalized. Meanings of complex words that are predictable as the sum of their parts are said to be compositional. Lexicalized words have meanings that are non-compositional. So the words *oddity* and *locality* that we looked at above have developed lexicalized or non-compositional meanings. Sometimes the meanings of derived words have drifted so far from their compositional meanings that it's quite difficult to imagine the compositional meaning for them. Consider, for example, the word *transmission*, which denotes a part of a car."

An occasional reference to an idiomatic derivative appears in Bauer et al. (2013: 30–31): "We should also note that although idiomatization typically occurs with the passage of time, it is nevertheless possible for words to be coined with meanings that are idiomatic from their inception; for example, according to the *OED,* the verb *cannibalize* was attested from the very beginning with the meaning 'to take parts from one machine to use in another'. It has never had the compositional meaning 'to act like a cannibal'."

Finally, the concept of semantic transparency in complex words is covered in great detail by Körtvélyessy, Štekauer and Zimmermann (2015). The authors, approaching language from the onomasiological perspective, demonstrate "that individual onomasiological types result from the omnipresent conflict between semantic transparency and economy of expression. […] The more complete the representation of the onomasiological structure, the higher the semantic transparency and, at the same time, the lower the economy of expression (ibid.: 92).

To summarize the theoretical section, it may be said that while authors have apparently no problem assigning idiomaticity to complex monolexical units, phraseologists consider it canonical to define phraseology as dealing (only and exclusively) with multi-word units. It seems therefore plausible to remedy this discrepancy and view lexical idioms as a part of the phraseological study.

4. DEFINITION OF LEXICAL IDIOMS AND AIMS OF THE STUDY

Based on the literature presented above, especially on Čermák (2007a, b), lexical idioms are defined for the purposes of the present study as single-word lexemes formed as a combination of components which are anomalous semantically and/or collocationally and/or grammatically.

Semantic anomaly is understood as non-compositionality and it is seen as a scalar quality including both more or less transparent units with slight discrepancy between the lexical and the word-formation meaning[5] to highly opaque idioms.

Collocational anomaly occurs when one of the components in a lexeme has very low collocability, or when the components in a lexeme are semantically incompatible (which is naturally accompanied by semantic non-compositionality in addition).

Formal anomaly is slightly more problematic. The most typical examples of grammatical anomalies of multi-word phraseological units (based on fixedness of the grammatical structure) cannot be simply transferred to the level of single-

5 The terms are used by Dokulil (1978) to discuss meaning predictability of complex words.

word lexemes because grammatical fixedness is typical of all lexical units that have undergone the process of lexicalisation and are thus stable in form. Therefore, grammatical anomaly in single-word lexemes may only be represented by the use of non-productive affixes (cf. Čermák 2007a: 74) or by the presence of some idiosyncratic formal irregularity (e.g. *spokesperson, showbiz*).

The study attempts to find out if all three types of anomaly, i.e. semantic, formal and collocational, are of equal importance to the definition of lexical idioms, or whether any of them is more important for the identification of lexical idioms than the others.

5. METHOD AND DATA

The data were retrieved from the British National Corpus (BNC), Version 3 (BNC XML Edition) and they consist of a random sample of 1000 lemmas of all relevant (i.e. open) word-classes — nouns, adjectives, verbs and adverbs, with no other restrictions but the lowest frequency of 10 occurrences.[6] The sample collected according to these principles thus contains both simple and complex words and both words formed within English and loanwords. The lexemes in the sample are then classified according to their morphological structure and complex lexemes are examined for the presence or absence of all three types of anomaly presented above, also focusing on specific features of the anomaly and distinguishing between anomalies occurring (to some degree) systematically and true idiosyncrasies.

6. RESULTS

The first step of the analysis was to identify complex lexemes which can be further analysed. However, although the differentiation of simple and complex lexemes is only the first step towards the analysis, it is evident already at this stage that the borderline is not very clear. Besides obvious instances of simple words, such as *white, rare, catch* and *money*, and obvious instances of complex words, such as *unrecognizable, nippy, government-owned* and *milkman*, there are also words such as *conflict, prepare, impact, odour* which are placed somewhere in between because they are partly analysable, especially for a native speaker of English with some knowledge of classical languages. Moreover, there are lexemes such as *superficial*, *amenity*, but also native *Tuesday* and *beware*, which are more likely to be classified as complex, although it is also problematic to fully describe their structure from the synchronic point of view.

Given the purpose of the study, I have drawn a somewhat artificial line between simple and complex lexemes, classifying the following problematic cases as instances of simple lexemes:

6 Some irrelevant items were excluded from the sample, especially proper names and initialisms.

1. borrowed lexemes with some traces of the original complex structure with no obvious link between the "components" and their meaning: *vagabond, anthem, insect, perplex;*
2. simple lexemes whose one part accidentally resembles an existing affix: *privy, beaver, gutter, mayor;*
3. borrowed lexemes with complex structure in the original language, with a component both formally and semantically recognizable in English and another one which does not occur systematically in English: *defend, repeal, collide, figment.*

The last group (3), however, is on the borderline to complex lexemes and if the instances were seen as polymorphemic units, they would be analysed as anomalous (collocational anomaly of the non-systematic component) and thus idiomatic.

Based on these criteria, the sample contains 319 simple lexemes and 681 complex lexemes. The distribution of word-classes in the sample is presented in Table 2:

WORD-CLASS	SIMPLE	COMPLEX	TOTAL
N	206	322	528
ADJ	13	264	277
ADV	2	47	49
V	98	48	146
TOTAL	319	681	1000

TABLE 2. Distribution of word-classes and simple/complex lexemes in the sample

The complex lexemes were examined further with a focus on their structure and the presence or absence of anomaly. According to the latter criterion, the lexemes were assigned to categories 0–4. A lexeme can be assigned to more categories if more types of anomaly occur at the same time (e.g. semantic and formal anomaly). Therefore, the number of items assigned to the categories is higher than the total number of complex lexemes in the sample (681). In total, there were 407 anomalies identified in the 300 lexemes assigned to categories 1–4 (cf. Table 3).

Category 0 signifies absence of anomaly, and it is thus assigned to non-idiomatic complex lexemes. These lexemes are semantically regular (i.e. transparent), formally regular (i.e. they contain a free base and are formed by productive processes), and also collocationally regular (there is no anomaly in the combination of constituents, i.e. there is no semantic incompatibility). The category was assigned to 381 complex lexemes (48.3% of all complex lexemes); the distribution of word-formation types does not differ significantly from the distribution in the rest of the sample (81.6% derivatives, 10.5% compounds, 7.9% combined formations[7]).

7 Combined formations are complex lexemes consisting of more than two elements which involve both composition and derivation (*bad-tempered, time-dependent, lexicographer*, etc.)

COMPLEX LEXEMES	TYPE	EXAMPLES OF THE CATEGORY	TOTAL OF ITEMS ASSIGNED
CAT 0	regular	N: *dryness, meeting, welder, ill-health, sociolinguist, scriptwriter, showbizz* V: *localize, rethink, untangle, outperform* ADJ: *horrified, removable, greyish, sleepless, audio-visual* ADV: *swiftly, smoothly, dryly*	381
CAT 1	form. anomalous	N: *atonement, scripture, breadth, proficiency* V: *migrate, maximise, pressure* ADJ: *astral, obligatory, venerable, olden*	158
CAT 2	coll. anomalous	N: *consumerism, deadlock, father-in-law, mulberry, mohair* V: *commentate, interface, overhaul* ADJ: *garrulous, defunct, newfound, supercilious* ADV: *best, henceforth*	81
CAT 3	sem. anomalous I	N: *oddity, offshoot, tipper, scholarship* V: *unearth, reconstitute* ADJ: *mouth-watering, skinny, thorny, proven*	40
CAT 4	sem. anomalous II	N: *dreadnought, pullover, mistletoe, whirlpool, livestock, casualty* V: *dismember, pin-point, depress, mortify* ADJ: *extra-mural, foolproof, sensible, catching, institutionalized* ADV: *gingerly, abroad*	128
TOTAL			789

TABLE 3. Categories of complex lexemes in the sample

When we take a closer look at the word class of regular complex lexemes, we will notice that not all word-classes behave in the same way.

As can be seen in Table 4, the proportion of regular complex lexemes is highest for adverbs, and also fairly high for adjectives. In contrast, most nouns and even more prominently verbs contain some anomaly. The regularity of the group of adverbs is not very problematic to explain. Derived adverbs are formed in English by quite a limited set of suffixes (*-ly, -wise, -ward(s), -ways, -s*). Of these suffixes, only *-ly* is highly productive. In fact, its productivity is so high that it is sometimes considered to be an inflectional morpheme in the literature (cf. Giegerich 2012). All 40 instances are adverbs ending in *-ly*. The group of adjectives is more heterogeneous. However, there is a large group of participial adjectives (*burned, confusing, understanding*) which tend to be regular (although idiomatic instances do occur occasionally, cf. *institutionalized, catching*).

WORD-CLASS	COMPLEX LEXEMES	CATEGORY 0 COMPLEX LEXEMES	CL / CATEGORY 0 CL PROPORTION (%)
N	322	158	48.8
ADJ	264	171	64.8
ADV	47	40	85.1
V	48	12	25.0
TOTAL	681	381	

TABLE 4. The word-class ratio between Category 0 and all complex lexemes

Category 1 contains words with a formal anomaly. The anomalies were divided into three groups: bound lexical bases, unproductive affixes and other anomalies. Bound lexical bases are not used in productive word-formation processes, unless they are combining forms. Therefore, from the synchronic point of view they represent an anomaly, although they are certainly quite common in English. Bound lexical bases came into English with extensive borrowing from Romance languages and Greek and they may be used systematically with Latinate affixes (cf. *nation, native, natality, neonate*). The category comprises 87 lexemes with bound bases, many of which combine this type of anomaly with one or more others. Here are some examples of lexemes with bound bases: *granary, serial, renunciation, resumption, publicise, paralytic.*

Unproductive affixes, like bound bases, do not participate in regular word-formation processes although they are very common in derivatives formed at some stage in history when they were productive (including both non-native and native affixes). In addition, a great many derivatives with unproductive affixes were borrowed into English as complete units, with word-formation taking place in the source language. Examples of both types include *atonement, dilatory, reversal, scripture, literate, scientific.* There are 64 lexemes with an unproductive affix.

Other anomalies include affixes specific for a different word-class (mostly due to conversion): *forward* (v.), *engineer* (v.), *pressure* (v.). However, conversion is very productive in English and it seems that this type of anomaly is not perceived to be very conspicuous. In addition, there are phrasal compounds (*father-in-law, check-in*), words involving clipping (*showbiz, compstation*), and lexemes with a truly idiosyncratic anomaly (*olden* — archaic inflectional affix lexicalized, *topmost* — anomalous superlative, *casualty* — anomalous form of the affix *-ity*, *gunwale* — anomalous pronunciation). It is this last group of formally idiomatic lexemes in Category 1 that seems to be most relevant for the definition of idioms. It includes lexemes that one would probably call lexical idioms based on intuition only, in which they differ from words such as *nation, publicise* or *sweeten* from the former two categories.

Category 2 was assigned to lexemes with a collocational anomaly. The sample includes 81 instances assigned to this category. Collocational anomaly is generally of three types: low collocability of a component, formal incompatibility between components and semantic incompatibility between components.

Especially the category of lexemes with low collocability is very hard to define. To achieve some objective classification, each component would have to be analysed in terms of its overall frequency of use and the frequency of occurrence within various

lexemes, assuming that the threshold frequency of occurrences will be different for an affix (which is in the case of regular use expected to occur in a higher number of cases) and for a lexical base, especially of a specialized meaning (which could probably occur in a limited set of words even if regular in use). In addition, it is sometimes difficult to decide whether an anomaly is collocational or formal. The reason for this is that often the anomalous combination leads to an anomalous form (a base which occurs in one structure only must be consequently also analysed as a bound base). The sample includes instances of Romance or Greek origin borrowed as ready-made complex units, e.g. *garrulous, duplicate, cadence, tirade, relegate*. There are 38 examples in the category.

A subgroup of this type are lexemes whose bases occur in a set of relatively frequent words, but the meaning cannot be retrieved by reference to the word family because their meaning is too divergent. This concerns for example *revolve* from the sample, which can be grouped together with e.g. *devolve, involve, evolve* and *convolve*. It is not easy to see any system in the use of *-volve* and its meaning in the complex words. Moreover, the same can be said for each of the words about the relation of the affix to the base *-volve*. These lexemes are certainly peripheral in the category of lexical idioms. They are close to the group mentioned above as a subgroup of simple lexemes and there are certainly reasons for seeing these lexemes as simple.

Other, more idiosyncratic instances of low collocability include *mohair* (a loanword in which the second part was identified with the form *hair* by folk etymology and *mo-* thus became a monocollocable component), *Tuesday* (containing an opaque and monocollocable component *tues-*), *mistletoe* (*mistle-* occurs also alone, but normally only in the collocation *mistle thrush* and *carpenter* (which has the form of an agentive noun, but the verb *carpent* is rare).

The second type, formal incompatibility between the components, is represented by lexemes resulting from a formally anomalous combination of components. Some instances are an idiosyncratic anomaly, some are members of a less prototypical word-formation pattern: *dogged, walling* (both lexemes with an affix which is prototypically deverbal, although marginally it can also be denominal), *urinal* (*-al* is added to a noun to form another noun), *polyunsaturated* (anomalous combination of prefixes), *consumerism* (anomalous combination of suffixes).

The third type, semantic incompatibility, was recognized mostly in compounds (20 compounds, 4 derivatives and 1 combined formation). This is not surprising given the fact that incompatibility of two lexical meanings is more probable due to a large number of lexical fields and specificity of lexical meanings, than semantic incompatibility between a lexical component and a grammatical component (which is more general in meaning and thus more compatible). Instances of semantic incompatibility are illustrated by *deadlock, spendthrift, rainbow, goalmouth, starfish, wholesale*. This subcategory is on the borderline between collocational and semantic anomaly. Some of these lexemes are instances of what we would probably intuitively call lexical idioms (*deadlock, spendthrift, rainbow*), others are based on a quite transparent semantic shift (*goalmouth, starfish, wholesale*) and these would be therefore more peripheral in the category of lexical idioms. There is a higher number of compounds among the Category 2 lexemes than in the whole sample of complex words. This reflects the fact

that collocational anomalies include semantic aspects, which appear to be more common in nominal compounds.

Category 3 is the first type of semantic anomaly. Lexemes included in this group are less idiomatic than lexemes included in Category 4 below. They have both literal and transferred meaning and the transferred meaning is usually quite transparent as it is based on some common semantic shift. The peripheral position in the category of lexical idioms is also supported by the finding that in the sample, lexemes of this category hardly ever display another type of anomaly. It is probably the case, however, that when the shifted meaning is used conventionally (i.e. lexicalized), it becomes partly independent of the original non-idiomatic meaning. The speaker knows that this specific non-compositional meaning is linked to this lexeme, and therefore, it is idiomatic in this non-compositional sense. Examples of category 3 include metaphorical senses of *mouth-watering* "looking delicious", *tubby* "plump" and *unearth* "to reveal or discover", senses based on metonymy of *oddity* "an odd person or thing", *skinny* "thin".

Category 4 was also assigned to lexemes with semantic anomaly. Lexemes in this category do not have literal meaning, only the idiomatic one. From this point of view, they are therefore more idiomatic than category 3 and the two categories can also be seen as two stages of the same aspect. There are 128 instances in the sample, including instances of metaphor, e.g. *pin-point* "to locate or identify exactly", *catching* (adj.) "infectious" or "captivating", *foreman* "a person, often experienced, who supervises other workmen", metonymy, e.g. *mindful* "keeping aware", *flipper* "the flat broad limb of seals, whales, penguins, etc.", *on-board* "on or in a ship, boat, aeroplane, or other vehicle", meaning specialization, e.g. *institutionalized* "placed in an institution, esp. a psychiatric hospital or penal institution or a children's home or home for elderly people", *washing-machine* "a machine for washing clothes". There are also occasional instances of meaning deterioration (*opportunism*) and amelioration (*collegiality*).

Other, non-systematic kinds of semantic anomaly include *double-sided* "usable on both sides", *nursery* "a room in a house set apart for use by children", *shorthand* "a system of speed writing". These last examples are probably closest to what we would intuitively call lexical idioms by displaying idiosyncratic discrepancies between lexical meaning and word-formation meaning. Another group with a high degree of idiomaticity includes exocentric formations: *dreadnought, pullover, spendthrift*.

The last group which is highly idiomatic from the semantic point of view is the group of particle compounds, e.g. *logon, roll-out, break-out, overhaul.* From one point of view, particle compounds, i.e. compound nouns formed from phrasal verbs, are indisputably lexical idioms in one aspect: they are analogous to phrasal verbs which are generally counted among idioms because of their polylexicality, opaque meaning and fixed structure. However, in most cases they are also formed regularly from the corresponding phrasal verbs and their meaning corresponds to the meaning of the phrasal verbs. This problem is also associated with a more general question of how to decompose complex units when their transparency or opaqueness is assessed.

The distribution of word-classes and word-formation types in category 4 is quite specific: there are considerably more nouns than in the whole class of complex lexemes (67.2% in category 4 vs. 47.3% in the whole sample of complex lexemes), and the

proportion of compounds is also higher (47.6% in category 4 vs. 17.5% in the whole sample of complex lexemes). These results indicate that compounds (and nouns) incline more often to anomalous meaning (which is also in accordance with the expectations based on a high number of mentions of idiomatic compounds in literature).

The relation between word-class / word-formation type and the category based on anomaly present (or absent) is provided below in Tables 5 and 6.

WC	CL	%	CAT 0	%	CAT 1	%	CAT 2	%	CAT 3	%	CAT 4	%
N	322	47.3	158	41.3	74	46.8	49	60.5	21	51.2	86	67.2
ADJ	264	38.8	171	45.0	56	35.4	15	18.5	17	41.5	26	20.3
ADV	47	6.9	40	10.5	3	1.9	3	3.7	0	0	4	3.1
V	48	7.0	12	3.2	25	15.9	14	17.3	3	7.3	12	9.4
T	681	100	381	100	158	100	81	100	41	100	128	100

TABLE 5. Word-class distribution in all categories of complex lexemes in the sample

WF	CL	%	CAT 0	%	CAT 1	%	CAT 2	%	CAT 3	%	CAT 4	%
D	523	76.8	311	81.6	146	92.4	51	63	30	73.1	60	46.9
C	119	17.5	40	10.5	12	7.6	29	35.8	9	22.0	61	47.6
D+C	39	5.7	30	7.9	0	0	1	1.2	2	4.9	7	5.5
T	681	100	381	100	158	100	81	100	41	100	128	100

TABLE 6. Distribution of word-formation processes in all categories of complex lexemes in the sample

7. DISCUSSION

The above analysis of a sample of 1 000 random lexemes from the BNC has shown that when analysing English vocabulary in terms of phraseology, one has to deal with various degrees and subtypes of each of the three main types of anomaly. Moreover, sometimes it is even problematic to assign a lexeme unequivocally to the category of simple or complex lexemes.

One of the main questions arising in connection with the defining criteria of lexical idioms is whether the criterion of (non-)productivity should be included in the list of relevant formal anomalies of lexical idioms. The main reason given as to why it should is that speakers must store and retrieve a formation based upon unproductive processes as one unit, without segmenting it into its components. However, there is no real evidence that this condition really distinguishes unproductively formed words from those formed productively. The assumption that regular combinations are stored and retrieved by segments while irregular combinations are stored and retrieved as whole units is much less problematic on the syntactic level, where the productivity rules generally have much greater validity (and it is extensively questioned even on that level by distributional phraseologists). Instead,

unproductivity should not be considered a defining criterion of lexical idioms in English because it does not represent an idiosyncratic phenomenon for two main reasons. First, the amount of unproductively formed English words is quite high. In fact, 18.6% of complex words in the sample (126 out of 681) are formed in this manner, which indicates that they are not exceptions in the proper sense of the word. Indeed, an overwhelming majority of these words are of common origin (Latin or Greek loanwords and words formed within English by analogy with these Latin and Greek words) and therefore, they are not idiosyncratic instances of irregularity — they exist rather within a system with certain rules, although the rules are not used (normally) any more to form new lexemes. Second, we cannot approach the production of words in the same way as we approach the production of word combinations. The form of an actual word is always fixed, and the free choice of a suffix is only theoretical. In view of the generally high occurrence of formal anomalies at the level of a word, I therefore assume that the overall relevance of formal anomalies in both production and (especially) perception is smaller than that of other types of anomaly and that unproductivity should better be excluded from the defining criteria of English phraseological units.

Moreover, the results indicate that semantic anomaly should be seen as the primary, most important type of anomaly. Although morpheme combinations are significantly different from word-combinations in the aspects of fixedness and combinability, semantic non-compositionality remains the most conspicuous sign of idiomaticity even in the case of morpheme combinations, both the anomaly of the components (here referred to as *semantic anomaly*) and the anomaly of their combinations (here referred to as *semantic incompatibility*).

The practical analysis also brought some additional questions about the scope of phraseology on the lexical level. Should we, for example, include only complex lexemes formed within English, or should we also include borrowings with a transparent structure (especially Latin and Greek formations)? If we study language from the synchronic point of view, this should not pose a problem — all lexemes with a clear morphological structure should be included in the study. Another practical problem is the problem of decomposition for semantic analysis. Should we see, for example, *dismemberment* as a non-idiomatic formation, analysing its immediate constituents *dismember* and *-ment*, where the independent meaning of *dismember* semantically corresponds to its meaning within the derivative? Or should we decompose the word completely into *dis-*, *member* and *-ment*, in which case it would be seen as idiomatic, given that the prototypical meaning of *member* is not the one it has in *dismemberment*? From the structurally-theoretical point of view, the former analysis would probably be the correct one. However, for some purposes (e.g. language acquisition, ELT) it would make more sense to treat the lexeme as idiomatic, because although the (three) basic components are all common in English, their sum does not correspond to the overall meaning (we could add that *dismember* is less likely to be known to the speaker than its individual parts).

If we exclude non-productivity from formal anomalies and consider the semantic anomaly as the primary and defining type, sometimes accompanied also by other types of anomaly, we may retrieve a list of central lexical idioms. They are defined as

lexemes exhibiting semantic anomaly in combination with some other type of anomaly. There are 39 instances in the sample and most of them are really what we might see as good examples of lexical idioms intuitively.[8] They are presented in Table 7.

NOUNS	VERBS	ADJECTIVES	ADVERBS
biplane, bloodstock, casualty, compstation, consumerism, data-base, deadlock, dismemberment, father-in-law, goalkeeping, goalmouth, check-in, livestock, mistletoe, mulberry, onset, payroll, rainbow, seascape, showbiz, spendthrift, starfish, stompie, understudy, upkeep, urinal, workload	*engineer, forward, interface, overhaul*	*defunct, dogged, godly, underhand, wholesale*	*hereby, underway*

TABLE 7. Central lexical idioms

A more comprehensive list of such central lexical idioms could be a useful resource for practical application (in textbooks, manuals, etc.).

The results have also shown that in accordance with the traditional literature on phraseology, idiomaticity on the lexical level is scalar in nature and the degree of idiomaticity is given by the combination of several factors: the prominence of semantic anomaly is not given only by the degree of discrepancy between the lexical meaning and the word-formation meaning, but also by the degree to which this discrepancy is transparent or opaque, and, in addition to semantic aspects, formal and collocational anomalies increase the level of idiomaticity, especially if they are truly idiosyncratic.

8. CONCLUSION

The study presented an analysis of English vocabulary with the aim to explore the regular and the idiomatic in complex words in English. The study has shown that it is indeed possible to distinguish the regular and the anomalous on the level of morpheme combinations and that it therefore makes sense to approach single-word lexemes as phraseological units.

Nevertheless, there are differences in the combinability of morphemes and words. In particular, the regular combination of components on the lexical level is never as free as is the regular combination of components on the syntactic level. This makes the difference between the non-idiomatic and the idiomatic on the lexical level less conspicuous.

It has been also revealed that the specific nature of English influenced by rich borrowing especially from Romance languages demands that the definition of idiomaticity be amended when transferred from Czech and that it should be based on analogy in the language system rather than on current productivity.

8 The most questionable is the idiomatic status of verbs in this group, as they are in this category due to the combination of opaque semantics and conversion.

Idiomaticity is seen as a scalar value influenced by three factors: the degree of opaqueness, the degree of discrepancy between word-formation and lexical meaning, and the degree of formal or collocational anomaly.

REFERENCES

Bauer, L., R. Lieber and I. Plag (2013) *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Beard, R. (1987) Lexical stock expansion. In: Gussmann, E. (ed.) *Rules and the Lexicon: Studies in Word Formation*. Lublin: Redakcja Wydawnictw KUL, 23–41.

Benczes, R. (2005) Metaphor-and metonymy-based compounds in English: a cognitive linguistic approach. *Acta Linguistica Hungarica* 52(2–3), 173–198.

Benczes, R. (2006) *Creative compounding in English: The semantics of metaphorical and metonymical noun-noun combinations*. Amsterdam-Philadelphia: John Benjamins Publishing.

Benczes, R. (2015) Are exocentric compounds really exocentric? *SKASE Journal of Theoretical Linguistics* 12(3), 54–73.

Burger, H. (1998) *Phraseologie: Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.

Cowie, A. P. (1998) Introduction. In: Cowie, A. P. (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Čermák, F. (2007a) *Frazeologie a idiomatika česká a obecná*. [Czech and general phraseology]. Praha: Karolinum.

Čermák, F. (2007b) Idioms and Morphology. In: Burger, H., D. Dobrovol'skij, P. Kühn and N. R. Norrick (eds) *Phraseology: An International Handbook of Contemporary Research*, 20–26. Berlin-New York: Mouton de Gruyter.

Dokulil, M. (1978) K otázce prediktability lexikálního významu slovotvorně motivovaného slova. *Slovo a Slovesnost* 39, 244–251.

Dokulil, M. (1986) III Tvoření slov. In: Petr, J. et al. (eds) *Mluvnice češtiny* (Vol. 1) Praha: Academia.

Filipec, J. et al. (1994) *Slovník spisovné češtiny pro školu a veřejnost*. Praha: Academia.

Giegerich, H. J. (2012) The morphology of-ly and the categorial status of 'adverbs' in English. *English Language and Linguistics* 16(3), 341–359.

Haspelmath, M. (2002) *Understanding Morphology*. London: Hodder Arnold.

Howarth, P. (1998) Phraseology and second language proficiency. *Applied Linguistics* 19, 24–44.

Kastovsky, D. (1982) *Wortbildung und Semantik*. Düsseldorf-Bern-München: Bagel/Francke.

Katamba, F. and J. Stonham (2006) *Morphology* (2nd ed.) Houndsmills-New York: Palgrave Macmillan.

Kavka, S. (2009) Compounding and idiomatology. In: Lieber, R. and P. Štekauer (eds) *The Oxford Handbook of Compounding*. Oxford: Oxford University Press.

Klötzerová, P. (1997) *Lexikální frazémy a idiomy v češtině*. Unpublished MA thesis. Praha: FF UK.

Klötzerová, P. (1998) Hranice frazeologie se posouvají. Lexikální frazémy v češtině. *Slovo a slovesnost* 59, 277–280.

Kos, P. (2018) O idiomatickém charakteru pojmenování v mutační kategorii. *Časopis pro moderní filologii* 100(1), 9–23.

Körtvélyessy, L., P. Štekauer and J. Zimmermann (2015) Word-formation strategies: Semantic transparency vs. formal economy. In: Bauer, L., L. Körtvélyessy and P. Štekauer (eds) *Semantics of Complex Words*, 85–113. Cham-Heidelberg-New York-Dordrecht-London: Springer International Publishing.

Lakoff, G. and M. Johnson (1980) *Metaphors We Live By*. Chicago: University of Chicago Press.

Levi, J. N. (1978) *The Syntax And Semantics Of Complex Nominals*. New York, NY: Academic Press.

Lieber, R. (2009) *Introducing Morphology*. Cambridge: Cambridge University Press.

Lieber, R. and P. Štekauer (2009) *The Oxford Handbook of Compounding*. Oxford: Oxford University Press.

Lipka, L. (2002) *English lexicology. Lexical structure, word semantics and word-formation*. Tübingen: Gunter Narr Verlag.

Rodrigues, A. and G. Rio-Torto (2013) Semantic coindexation: evidence from Portuguese derivation and compounding. In: Pius ten Hacken, C. T. (ed.) *The Semantics of Word Formation and Lexicalization*, 161–179. Edinburgh: Edinburgh University Press.

Štekauer, P. (1998) *An Onomasiological Theory Of English Word-Formation*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Štekauer, P. (2005) *Meaning Predictability in Word Formation: Novel, context-free naming units*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Vachek, J. (1966) *Travaux linguistiques de Prague: Les problèmes du centre et de la périphérie du système de la langue*. Prague: Academia.

Kateřina Vašků
Department of English Language and ELT Methodology
Faculty of Arts, Charles University
nám. Jana Palacha 2, 116 38 Praha 1
ORCID ID: 0000-0002-0169-082X
e-mail: katerina.vasku@ff.cuni.cz