UDC 811.163.6'373.7
Polona Gantar
Fran Ramovš Institute of the Slovenian Language, Ljubljana

CORPUS APPROACH IN PHRASEOLOGY AND DICTIONARY APPLICATIONS

This paper compares an attempt to identify the phraseological unit on the basis of the degree of semantic motivation of phrasal elements, originating in the Russian phraseological tradition, with various aspects of word combining, as revealed in the corpus environment. The relativisation of relations between single word and multiword lexical units on the one hand and the semantically transparent and opaque phrases on the other broadens the subject-matter of phraseology to different types of language patterning and also offers dictionary solutions based on the contextual treatment of the lexical element.

V članku soočamo poskus identifikacije frazeološke enote na podlagi stopnje pomenske motivacije besednozveznih elementov z izhodišči v ruski frazeolški literaturi in različne vidike besedne povezovalnosti, kot se odkrivajo v korpusnem okolju. Relativizacija razmerij med eno- in večbesednimi leksikalnimi enotami ter pomensko transparentnimi in netransparentnimi besednimi zvezami širi predmet frazeološke problematike na različne tipe jezikovnega vzorčenja in hkrati ponuja slovarske rešitve, ki temeljijo na kontekstualni obravnavi leksikalnega elementa.

Key words: phraseological unit, multiword unit, fixed expression, collocation, phraseme, pure idiom, idiomaticity, phraseologically bound or idiomatic meaning, syntactic patterns, lexical unit; corpus-based approach, lexicographical aspects, dictionary framework
Ključne besede: frazeološka enota, večbesedna enota, stalna besedna zveza, kolokacija, frazem, pravi idiom, idiomatičnost, frazeološki ali idiomatični pomen, stopnja pomenske trdnosti, skladenjski vzorci, pomenska kohezija, leksikalna enota; korpusni pristop, leksikografski vidik, slovarska struktura

1 Phraseology – delimiting the field

The phraseological theory has for some time attempted to delimit in as much detail as possible the field of phraseological research and the basic phraseological unit (PU). For this purpose a set of criteria has been formed according to which the basic and distinctive (in contrast with other lexical units) features of the phraseological unit could be determined. The phraseological theory is most complex where most of the rules recognised and confirmed in similar language samples of the so-called conventional language[1] are blurred; this can be established by the fact that features such as *multiword character, collocability, stability, variation, idiomaticity, connotativity, transformability,* etc. are considered from different angles which leads to opposing ideas about what is essential for the existence of the PU.

[1] In those phraseological papers which are based on a study of language on different levels, conventional language is the one in which the general syntactic and semantic rules operate as opposed to the systemically unexpected realisations (i.e. anomalies) typical of the PU. (c.f. Čermák 1985: 167).

It seems that this state in the field of multiword units is not a coincidence. It is, on the one hand, a consequence of the fact that multiword units present a complex linguistic phenomenon in which the distinctive features are realised to different extents, while on the other hand, the reason for their independence from the syntactic and semantic processes predicted by the system lies in the fact that, due to their idiosyncrasy, their individual constituent parts cannot be considered from separate syntactic and semantic points of view. The traditional treatment of the PU has thus focused on a certain type of multiword units which fitted specific demands, e.g. they are not structurally and semantically fixed, they have a connotation, they are non-terminological, etc., while other units were excluded from the narrower phraseological and consequently dictionary treatment.

1.1 Idiomaticity and the phraseologically bound meaning

The Russian phraseological theory, started by V. V. Vinogradov and N. N. Amosova in the 1950s and 1960s, and the majority of East European phraseological schools built on its foundations tried to form a system of categories which could be used to separate the field of phraseology from the field of general word-combining rules. The fundamental feature of this concept of the PU is based on the ideas of *idiomaticity* and *phraseologically bound meaning*.

Idiomaticity, which applies to the relationship between the entering and the exiting semantics of the constituent parts of the PU as opposed to the meaning of the PU as a whole, can generally be understood as a universal linguistic phenomenon; the distinctive features of morphemes, words and phrases in different languages differ both in form and content. If we leave aside the possibility of idiomatic combinations on the morpheme level and neglect the existence of single word idioms then, as a phraseological issue, idiomaticity is linked to recognising the level of semantic independence of the entire PU in relation to the meaning of the individual parts. This happens in spite of the fact that there have always been differences in understanding the degree of semantic motivation of a concrete PU, while it has been impossible to set sharp boundaries between the various degrees of such a concept of the PU, since determining the type of the PU on such a basis largely depends on the linguistic and cultural experience of the individual speaker (Cowie 1998: 215).

Based on the concept of phraseological meaning, i.e. the meaning of the PU as a whole, not the sum of its constituent parts, as the key feature of the PU, two basic types of the PU were identified in the phraseological theory: those PUs which can be semantically analysed (their meaning is dispersed to different extents among their constituent parts), and those which show no such relation and are entirely semantically unmotivated (Erbach 1992: 12; Nunberg et al. 1994: 496–497). At the same time the inability to literally translate the phraseological meaning was proposed as one of the basic conditions for recognising a PU, even though recent text-based research has shown that the phenomenon of interlinguistic idiomaticity is relative and dynamic since an expression can be idiomatic in a certain language but its foreign language counterpart may not be idiomatic (Mlacek et. al 1995: 64). The above starting points

were the main reason for treating the PU as a specialised segment of the lexical fund and the content of specialised dictionaries, while they were presented quite ineffectively and unsystematically in general dictionaries.

1.2 The Slovenian linguists' approaches to the PU have, since the first theoretical paper[2] on phraseology (Toporišič 1973/74), followed the attempts to place the PU onto different levels of the language structure. Criteria for determining the PU were formed and they took into consideration the multiword character, permanency and the possibility of automatic reproduction. Including multiword terms among the PU[3] meant setting the groundwork for phraseology in the wider sense, at first on the basis of the Russian theoretical approaches. In the second half of the 1980s, the demand for at least one constituent part to have a »meaning distinct from the dictionary meaning« (Kržišnik-Kolšek 1986: 435) and the elimination of terminological expressions from the narrower phraseological framework established the distinction between fixed expressions and PUs. When the concept of collocability was introduced (Kržišnik-Kolšek 1988: 51–54), the idea of the PU was limited to monocollocable units of the type *priti/spraviti na kant* (Engl. *to go broke/ to make someone bankrupt*), while the so-called limited collocability of the type *kriv + obtožba, ovadba; pričevanje, izpoved, prisega; nauk, vera* (Engl. *false + charge, report; testimony, confession, oath (= perjury); teachings, creed (= heresy)*) (Kržišnik 1994: 33) was not specifically determined in relation to the PU, even though this meant that phrases in which the words in one of their meanings collocate with a relatively limited range of other words, e.g. *star + mama, mati, oče starši; star + celina, kontinent, svet* (Engl. *old + mother (= grandmother), father (= grandfather), parents (= grandparents); old + continent, world = Europe*), were excluded from lexicological and phraseological research. An important criterion which turned the attention of the Slovenian phraseological research to a very restricted segment of multiword referential units (i.e. phraseology in its narrower sense) was focusing on only those units which have important connotative semantic components and an important pragmatic role (Kržišnik 1990: 400); this excluded from phraseological research phrasal verbs such as *držati s kom* (Engl. *side with so.*), *pristati na kaj* (Engl. *agree to sth.*), etc., prepositional collocations of the type (*razlikovati, sortirati*) *po barvi*; (Engl. *(distinguish, sort) by colour; v barvi* (*kože, lesa*) (Engl. *in the colour (of skin, wood)*); (*igrati*) *na mestu* (*branilca*) (Engl. *(play) as a defender*), and units with a so-called grammatical meaning, such as: *ne glede na* (Engl. *regardless of*); *za razliko od* (Engl. *as opposed to*); *v primerjavi*

[2] In Slovenian linguistics, phraseology has been considered as a research topic at least since the late 1950s, when the bases of Pavlič's *Frazeološki slovar v petih jezikih* (Engl. *Phraseological dictionary in five languages*) (1960) were formed and when the grounds were determined for the presentation of phraseology in the *Slovar slovenskega knjižnega jezika* (Engl. *Dictionary of the Standard Slovenian Language*), the Volume One of which was published in 1970.

[3] An important contribution is defining the group of terminological fixed expressions with classifying adjectives of the type *mehki, trdi les; črni bor* (Engl. *soft, hard wood; black pine*), etc. on the basis of formal recognition of the degree of semantic unity as revealed in the phrasal or morphemic composition of technical terms (Vidovič Muha 1988).

z/s (Engl. *in comparison with*), etc., in addition to the above mentioned terms. There have been attempts to determine the highest possible phraseological entity which is not yet a discourse entity and thus resolve the issue of the clausal construction of fixed expressions; the most disputable were expressions consisting of a verb + noun, the so-called false verbal phrasemes (Kržišnik 1994: 83), such as *luna trka koga* (Engl. *the moon knocks so.* (= *be off one's rocker*)), *srce pade v hlače komu* (Engl. *so' heart sinks into his/her pants* (= *so.' heart sinks*), etc. As far as the methodology used in linguistic analyses of PUs is concerned, intuition played an important role (e.g. through recognising the structure and the meaning of fixed expressions and their transformational possibilities in surveys, etc.); due to the lack of substantial corpora (until 1997), the analyses were typically limited to certain types of texts, such as newspapers, works of literature, *the Slovar slovenskega knjižnega jezika* (Engl. *Dictionary of the Standard Slovenian Language*).

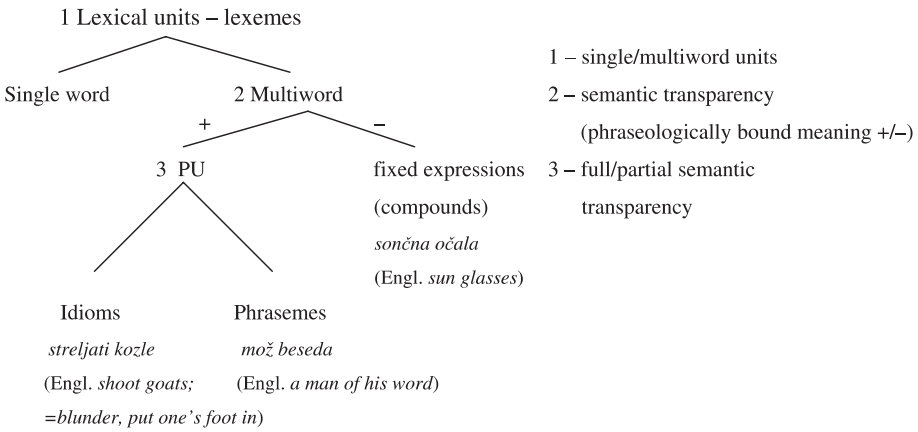


Figure 1: *PU placement into the lexical fund of the language according to their structural and semantic base*

1.3 Collocability and collocations

The Anglo-Saxon approaches to the issue of multiword units, on the other hand, which originate in the traditions of A. S. Hornby and H. E. Palmer, also considered those word combinations which are not strictly semantically unmotivated (i.e. pure idioms). This starting point enabled the recognition of typical word combinations; the level of idiomaticity, demonstrating itself as a relative linguistic phenomenon, seemed less important than the fact that, in the process of language acquisition and learning, certain word combinations cannot be »put together« from their individual constituents but are rather learned as a whole. This was the basis for including multiword units, especially collocations, in learner's dictionaries. At the same time, recognising the trends in word combinations, regardless of their phraseological predispositions in the

sense of how fixed they are semantically and grammatically, how expressive they are, whether they are non-terminological, made it possible to recognise the syntactic patterns in which words and expressions tend to appear and established the starting point for considering the issue of words and phrases within a context.

2 The starting points for a corpus-based approach to phraseology

Such a starting point is a good opportunity to observe, within the corpus environment, the capacity of words to connect in a text with a large or limited number of other words. It has turned out that corpora are especially useful for studying the issue of phrases, since their ability to automatically sort concordance strings and measure word combinations in the form of statistical calculations has shed light on word collocability. In addition to stressing the importance of an empirical analysis of language data, the phraseological literature emphasises the necessity of a suitable quantity of language data, especially in determining the regularity of transformation processes and variation. However, one of the basic problems in studying phraseology and establishing its rules still lies in the fact that conclusions are made on the basis of a small quantity of data (Čermák 2001: 5). As mentioned above, one of the peculiarities of phraseology is that categories familiar from elsewhere are blurred within fixed expressions and it is impossible to determine them from a small sample.

Using a corpus for lexicographical purposes thus offers a chance to identify those word co-occurrences which are typical of a language. Studying the samples obtained also provides, in a real context, an insight into the typical semantic and communicative roles. Both uses of the corpus, as material for analysis, as well as for methodological purposes, give the lexicographer more flexibility in dictionary design especially in relation to the potential user. It is probably no coincidence that dictionary projects contributed, among other things, to the shaping and improving of corpus design in the sense of compiling greater linguistic variation and to corpus tool development. It was above all those dictionary projects which tried to provide as real language data as possible, primarily by choosing to present those headwords, phrases and their forms which are well-represented in real language. It was corpus data that made lexicographers reconsider the issue of including forms which are simply the results of word-formation possibilities of the language and have not been found in real texts. Corpus-based dictionaries can better capture the semantic value of lexical elements and establish their true frequency. Corpora have also provided entirely new possibilities in dictionary use. If we accept the demand for coherence and communicative effect of the text, our starting point is the fact that a text is formed in a number of very sophisticated ways; a made-up example can more or less successfully mimic them, but cannot replace the sensibility and the importance of the context. This also makes it possible to determine on the basis of a corpus measurable and thus fairly objective criteria for collecting the essential features of multiword units.

2.1 The features of a PU – the starting points and determining the criteria

In a corpus environment, the features of word combinations – especially to identify various types of multiword units as potential dictionary entities – can be established from three different starting points: frequency, functional or semantic. These starting points are further determined with the basic procedures of corpus analysis (Teubert 1999): the identification of language data, where the word or its form is the basis; correlating language data with the use of statistical methods, where aspects of word combination and language sampling are observed, and the interpretation of the results. *FIDA, A Reference Corpus of the Slovenian Language* and the concordancer available to its users were used for this purpose.

The frequency aspect, which is the centre of our analysis, refers to the recognition of obvious word co-occurrences and determining the typical collocators of the word studied within the concordance string. Figure 2 shows that *FIDA* provides at least three possibilities, with MI^3 yielding the best results, especially when considered together with the data on absolute frequency of a corpus element studied in a concordance string (Gorjanc and Krek 2001).

Figure 2: *The frequency starting point; the keyword of the concordance string: šala (Engl. joke)*

(a) 10 most frequent collocators left of the keyword

1	za	544
2	v	348
3	biti	188
4	in	59
5	dober	56
6	kota	55
7	Kot	55
8	zbijati	55
9	se	49
10	pripovedovati	46

(b) 10 mutually connected collocators left of the keyword – MI value

1	=privihnil	14.987226
2	=jelušičeve	14.987226
3	=yorkshirskim	14.987226
4	=gattinonijevi	14.987226
5	=vsplahutale	14.987226
6	=yorkshirskimi	14.987226
7	=carmichaelovega	14.987226
8	=vzorčan	14.987226
9	=vidrnim	14.987226
10	=pinčevi	14.987226

(c) 10 mutually inter-connected collocators left of the keyword – MI^3 value

1	zbijati	24.799923
2	=prvoaprilska	24.164425
3	neslan	23.591403
4	za	21.953701
5	=prvoaprilsko	20.973203
6	zbijanje	19.883292
7	neslano	19.783085
8	=prvoaprilske	19.708851
9	=severin	19.211598
10	v	18.915057

The functional aspect is based on recognising the typical syntactic patterns in which the keyword of the concordance string occurs and establishing semantic links between them. These patterns are typically the result of the grammatical and semantic features of a language and are therefore some sort of grammatical and lexical conglomerates. As such they become, in a corpus environment, the starting point for various grammatical, lexical and syntactic analyses and turn the attention from the level of the language system to studying examples of textual realisations; the typological rules created on such grounds also take into consideration all the »violations« which represent the basis of a topical linguistic description. Thus for instance the verb *veljati* (Engl. *to be in force, to be worth*) – in addition to typical collocations where the indi-

vidual collocators are semantically distinctive – with the preposition *za* (Engl. *for, as*) forms syntactic patterns which are of lexicographic interest in various stages of the process of lexicalisation (cf. a1 and a2, which are phrasal verbs, as opposed to b).

Figure 3: *The functional starting point; the keyword of the concordance string:* ***veljati za*** *(Engl.* ***to be considered as, to apply to****)*

syntactic pattern	**textual pattern**	**meaning**	**textual example**
(a1) veljati za *kakšnega* (Engl. *be considered)*	veljati za *pomembnega/ nedolžnega/spoštovanega/ uglednega …* (Engl. *be considered important/ innocent/respectable/ distinguished)*	*'be such'*	*V nasprotju z Nizozemci, ki* ***veljajo za*** *skromnejše turiste, Belgijci na počitnicah ne stiskajo pri izdatkih.* (Engl. *As opposed to the Dutch, who* ***are considered*** *more moderate tourists, the Belgians are not unwilling to spend money when on holiday.*)
(a2) veljati za *koga/kaj* (Engl. *be considered someone or something)*	veljati za *favorita/ utemeljitelja/ začetnika/ predhodnika …* (Engl. *be considered a favourite/the founder /the beginner/the predecessor*)	*'be someone or something'*	*Te freske, ki žal niso v celoti ohranjene,* ***veljajo za*** *glavno delo slikarske šole v Ferari.* (Engl. *These frescoes which have unfortunately not been entirely preserved* ***are considered*** *the main work of the Ferara school of painting.*)
(b) veljati za *koga/kaj* (Engl. *apply to someone or something*)	veljati za *vse/večino/oba…* (Eng. *apply to all/ most/both*)	*'be relevant for someone or something'*	*Prišli boste do spoznanja, da je življenje borba, in če to* ***velja za*** *vas, zakaj ne bi* ***veljalo za*** *druge.* (Engl. *You will see that life is a struggle and that if this* ***applies to*** *others, why should it not apply to you.)*

The semantic aspect is based on recognising cognate semantic realisations on the basis of lengthy concordance strings and enables the typical semantic features to be transferred into dictionary definitions. The semantic aspect is a key feature in the identification of fixed expressions and cannot be treated separately from the frequency and syntactic aspects; the recognition of the lexical role of an expression notable for its frequency is a highly complex phenomenon and very much linked to the typical elements of the context, text type and other extra linguistic phenomena. As native speakers, we never entirely abandon the intuition in the interpretation of the semantic content; this is also true in the case of multiword units. However, it is possible, on the basis of numerous textual realisations revealed by concordance strings and various possibilities of sorting textual materials in a corpus environment, to recognise with a great degree of certainty cognate semantic realisations and abstract them in the sense of dictionary definitions.

We therefore anticipate that each frequently occurring mutually bound multiword unit is a potential lexical unit or a typical syntactic pattern of the language studied, and this broadens the narrow phraseological field to the entire concept of the fixed expression, where its actual lexical role still needs to be established. Slovenian materials too

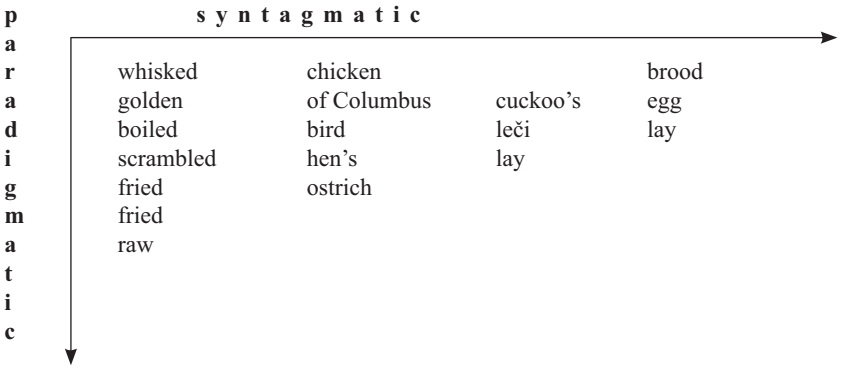
have shown that typical co-occurrences need not have notable lexical values as well. Co-occurrences with grammatical and semantic elements (prepositions and conjunctions) are particularly common; they become interesting from a lexicographic point of view as part of broader syntactic patterns, e.g. *kar tako **za šalo**; bolj **za šalo** kot zares; gre **za šalo;** kot **za šalo** (premagati, opraviti z/s kom/čim, pomesti s kom); malo/malce **za šalo** (in) malo/malce zares; napol **za šalo** napol zares; **za šalo** (povprašati, reči ...); vzeti, jemati **za šalo;** imeti smisel **za šalo*** (Engl. *just like that, **as a joke**; more **as a joke** than for real; **it's a joke**; easily (beat so., deal with so. or sth., sweep aside); half-**joking,** half-serious; (ask, say,…); **as a joke**; take as **a joke**; to know how to take **a joke***), etc.

2.2 A typology of fixed expressions based on corpus data

Research anticipates three relatively independent types of fixed expressions which can be determined by recognising syntagmatic and paradigmatic relations revealed by corpus data, obtained according to the starting points outlined above. The syntagmatic features are reflected in the ability of a word to form collocations on the horizontal axis, while the paradigmatic powers are revealed in the possibility to accumulate words on the vertical axis within an individual semantic field. The anticipated types of fixed expressions also predict potential dictionary headwords and a hierarchical relation within the dictionary entry, as we will see below.

The relativity of relations between single and multiword lexical units and between the semantically transparent and semantically opaque fixed expressions is blurred in a corpus. This of course does not mean that it is impossible to determine the basic subject-matter of lexical study on the basis of corpus data; it means surpassing the discrete separation into two categories: words and phrases on the one hand and fixed and free expressions on the other. By enabling the recognition of word co-occurrences, the corpus has also given new value to concepts such as a lexical or grammatical unit. When transferred into lexicographic practice along with the fact that the focus of

Figure 4: *The sorting of collocators left of the keyword according to syntagmatic and paradigmatic relations; the keyword of the concordance string: jajce (Engl. egg).*



lexicography is ascribing meaning to the linguistic sign which manifests itself always and exclusively within a text, this has resulted in equivalently treating the topics of the word and the phrase, while fixed expressions and obvious syntactic patterns were no longer limited to specialised dictionaries, but were included in general dictionaries as well.

The concept of collocations as a phenomenon of formal probability and at the same time a semantic phenomenon which is revealed through the mutual interconnection of lexical elements presents two types of typical word co-occurrences: those that indicate individual meanings of a polysemous lexeme, e.g. (*rdeča, modra, svetla, temna* ipd.) *barva*; (*tiskarska, oljna*) *barva* (Engl. (*red, blue, light, dark,* etc.) *colour;* (*printing (=printing ink), oil) colour*) – 'material used for colouring'; (*politična, klubska*) *barva* – (Engl. *(political, club) colour*) – 'a reflection of belonging to sth'; and those that occur only in the chosen sense of the word and thus form more or less limited collocational paradigms[4] (Čermák 1985: 171, Kržišnik 1988: 51), e.g. *barva kože* (Engl. *colour of skin*) – 'race'; *osnovne barve* (Engl. *primary colours*) 'basic colours of the colour spectrum,' etc. The first type of collocations, which appears under the heading I in the table below, creates a direct link to single word elements and presents a typical contextual placement of the word in question. The second type (appearing under the heading II) comprises those fixed expressions and syntactic patterns which are between the semantically transparent, of the type *osnovne barve* (Engl. *primary colours*), (*cvetje* etc.) *vseh barv* (Engl. (*flowers* etc.) *of every colour*), (*razlikovati, razvrstiti, ločiti*) *po barvi* (Engl. *(distinguish, sort, separate) by the colour*); *obrniti (kaj) na/v šalo* (Engl. *turn sth. into a joke*), etc. and semantically opaque, where the semantic link between the collocating elements is mutual, e.g. *spreminjati barve* (Engl. *change colour*) 'express anger, distress' etc.; *priti s pravo barvo na dan* (Engl. *show one's true colour*) 'express one's true, secret intentions or character'. In the phraseological literature, this type, along with semantically opaque expressions, presents the central part of the phraseological field and is generally referred to as a *phraseme*. On the dictionary level, the term refers to phrases which need an explanation, while they tend to be more or less linked to one of the meanings of the word forming such a phrase; this presents possible starting points for the dictionary hierarchy. Idioms are an extreme in the lack of expression of the meaning or anticipating the meaning from the constituents of the phrase; this is why they are generally treated as semantically relatively independent units within the dictionary.

Nevertheless, it is impossible to draw a sharp boundary between semantically transparent and opaque fixed expressions, or phrasemes motivated by association, such as e.g. *ubiti dve muhi na en mah; obrniti komu hrbet, imeti zvezane roke* (Engl.

[4] The concept of a collocational paradigm is derived from various aspects of word combining where the (linguistic) meaning can be taken into consideration. The lack of limitation on the one hand and the limited number of elements (which morphologically and semantically function as a logical whole within the collocational paradigm) on the other offer two basic sets of phrases among which the basic unit of phraseology is determined: the broader collocational paradigm, which is an open set and has an unlimited number of elements and the limited collocational paradigm which is generally a closed set and has a limited number of elements.

kill to birds with one stone; turn one's back on so., tie one's hands), etc., and the so called »pure« idioms, which include phrases without an obvious semantic link between any of their constituent parts, e.g. *iti se gnilo jajce, železna zavesa, na vrat na nos* (Engl. *play rotten egg, iron curtain, out of the blue*), etc. However, since the dictionary, with its user friendly nature of a reference book, demands a consistent structure, it is reasonable to think through the relationship between single word and multiword lexical units as potential keywords in the stage of dictionary design. The following table presents a possible general solution.

DEGREE OF SEMANTIC TRANSPARENCY/OPACITY

LEXICAL UNIT	**collocation** I.	**collocation** II.	**fixed expression** phraseme	**fixed expression** idiom
single word	***barva*** *(rdeča, modra; svetla, temna ...)* Engl.: ***colour*** *(red, blue, dark, light...)* ***barva*** *(tiskarska, oljna)* Engl.: ***colour*** *(printing (=printing ink), oil)* ***barva*** *(politična; klubska)* Engl.: ***colour*** *(political, club)* ***šala*** *(posrečena, robata, opolzka ...)* Engl.: ***joke*** *(good, rude, obscene)*			
multiword		***osnovne barve*** Engl.: ***primary colours*** ***barva kože*** Engl.: ***colour of skin*** *(cvetje)* ***vseh barv*** Engl.: *flowers* ***of every colour*** ***v barvi*** *(kože, lesa)* Engl.: *in the* ***colour of skin/wood*** ***prvoaprilska šala*** Engl.: ***an April Fool*** ***neslana/neokusna šala*** Engl.: ***improper joke*** ***v šali*** *(dejati, praviti)* Engl.: *(say sth.)* ***in joke*** *(kot)* ***za šalo*** *(premagovati ...)* Engl.: *easily (overcome,...)* ***ni šale*** *(s kom/čim)* Engl.: *(sth./so.) is* ***no joke*** *(razumeti, vzeti, jemati)* ***kot šalo*** Engl.: *(understand, take)* ***as a joke*** ***obrniti*** *(kaj)* ***na/v šalo*** Engl.: ***make a joke of*** *sth.* ***za šalo*** *(vprašati, poskusiti ...)* Engl.: ***for fun*** *(ask, try,...)* *(ne)* ***biti za šalo*** Engl.: *(not)* ***be able to take a joke*** ***iti za šalo*** Engl.: ***to be a joke***	***spreminjati barve*** Engl.: ***change colour*** ***priti s (pravo) barvo na dan*** Engl.: ***to show one's (true) colours*** ***zbijati/stresati šale*** Engl.: ***tell jokes*** ***šalo na stran*** Engl.: ***stop joking*** ***malo za šalo malo zares*** Engl.: ***half-joking, half-serious***	***hudič/vrag je odnesel šalo*** *Engl.:* ***the devil took the joke*** *(=things are more serious than they seemed at first)*

The types of fixed expressions mentioned above present potential dictionary units. The degree of semantic transparency/opacity of a potential dictionary unit seems irrelevant for the usefulness of information provided by the dictionary, above all if that means excluding phrases which are not lexicalised enough. Nevertheless, it seems important for the organisation of the dictionary itself to present, on various levels of the dictionary entry, the different possibilities of word combinations which are obvious in a language and have, at the same time, a visible lexical role.

2.3 The relationship between single word and multiword lexical units – a dictionary example

The fact that multiword lexical units are equal to their single word counterparts in their lexical role does not justify their subordination within the dictionary entry, at least not in the sense of omitting expected dictionary information (meaning, pronunciation, part of speech classification, examples of usage, etc.). Multiword placement within a dictionary system is, however, more complicated than it seems at first glance, at least for two reasons. The first is the already mentioned fact that multiword units are composed of elements of »conventional« language, which means that they generally also exist outside the concrete idiomatic combination, and the second is that they tend to keep, to various extents, their extra-idiomatic grammatical and semantic features within the idiomatic combination. Establishing semantic associations between the constituent of a fixed expression and the word which also exists independently of the fixed expression presents a possibility to sort fixed expressions within a dictionary according to previous semantic and grammatical data of the constituent parts of the superordinate (single word) headword. Since the context of the word studied as the keyword in a concordance string is the focus of our attention, it is possible to select, among the typical co-occurrences, those possibilities of word combinations which are created for instance by the metaphorical potential of a polysemic word, which becomes a constituent of a fixed expression. The degree of semantic transparency/opacity can then present a solid basis for placing a fixed expression under a certain meaning of a word in the role of the superordinate single word headword[5]. Let us consider some of the possibilities.

kisel (Engl. ***sour***) adj.
1. 'taste'
collocations kislo (grozdje, sadež, solata)
Engl: *sour (grapes, fruit, salad)*
kislega (okusa)
Engl: *sour (tasting)*
fixed expressions kisla smetana
Engl: *sour cream*
kislo zelje
Engl: *sauerkraut*
kisla repa
Engl: *sour turnip*
kislo mleko
Engl: *sour milk*
kisla juha
Engl: *sour soup*
kisla kumarica
Engl: *sour gherkin (=pickled gherkin)*
2. 'unhappy, unpleasant'
collocations kisel (nasmeh, obraz, vreme)
Engl: *sour (smile, face, weather)*
fixed expressions ugrizniti v kislo jabolko
Engl: *bite a sour apple (=start sth. unpleasant)*
3. 'having low ph value'
collocations kisla (tla, prst, zemlja; okolje)
Engl: *acid (soil, earth; environment)*
fixed expressions kisli dež/padavina
Engl: *acid rain/percipitation*
kisla voda
Engl: *mineral water*

idiom čas, sezona kislih kumaric
Engl: *period of pickled gherkins (=period, during the holiday, when there is no news)*

lisica (Engl. ***fox***) noun
1. 'animal'
collocations (stekla, povožena) lisica
Engl: *(rabid, run-over) fox*
(pokončati, upleniti) lisico
Engl: *(to kill, to hunt) fox*
(ceplenje, vakcinacija) lisic
Engl: *fox (vaccination)*
lov na lisico
Engl: *fox hunting*
fixed expressions šakalska lisica
Engl: *crab-eating fox*
leteča lisica
Engl: *flying fox*
morska lisica
Engl: *thresher shark*
polarna/arktična/bela/srebrna lisica
Engl: *arctic fox*
puščavska lisica
Engl: *desert fox*
2. 'fox fur'
collocations krznena lisica
Engl: *fox fur*
fixed expressions (mariborska, pohorska) zlata lisica
Engl: *(Maribor, Pohorje) golden fox*
3. 'clever person'
collocations (prebrisana) lisica
Engl: *(clever) fox*
(zvit) kot lisica
Engl: *(clever) as a fox*

lisice (Engl. literally: ***foxes; =handcuffs; clamps***) noun
1. 'a device for securing a prisoner's wrists'
collocations (natakniti, nadeti, sneti) lisice
Engl: *(put on, snap on, remove) handcuffs*
fixed expressions policijske lisice
Engl: *police handcuffs*
vkleniti v lisice
Engl: *shackle in handcuffs*
2. 'a device for immobilizing an illegally parked car'
collocations (odklepanje, priklenitev) lisic
Engl: *(unlocking, locking of) clamps*
odpraviti lisice
Engl: *do away with clamps*
uvedba lisic
Engl: *the introduction of clamps*

3 Conclusion

Attempts to delimit the field of phraseology based on determining the degree of idiomaticity have not resulted in a single concept of the PU, especially when compared to free and semantically transparent fixed expressions. The concept of an idiomatic meaning of a phrase, i.e. a meaning which does not depend on its constituent parts, narrows the field of phraseology to semantically opaque fixed expressions

which used to be the subject-matter of specialised phraseological dictionaries, while all other forms of language patterning were excluded from them and presented ineffectively and unsystematically in general dictionaries. The presentation of aspects of word combining regardless of the degree of their grammatical and semantic fusion, and the possibility to automatically sort concordance strings and measure word collocability in the form of statistical calculations in a corpus environment make it possible to form objective bases for recognising typical word combinations which may be of dictionary interest in various stages of the process of lexicalisation. Even though the relations between single word and multiword lexical units and the relations between semantically transparent and semantically opaque fixed expressions are blurred in a corpus, it is possible to quite objectively show both the structural and the semantic extension of the word to the level of the phrase or a longer syntactic pattern by considering the lexical element in its context.

V angleščino prevedla
Agnes Pisanski Peterlin.

References

Cowie, Antony, 1998: Phraseological Dictionaries: Some East-West Comparisons. *Phraseology.* A. P. Cowie (eds.). *Theory, Analysis, and Applications*. Oxford University Press. 210–228.

Cowie, Anthony, 1999: Phraseology and corpora: some implications for dictionary making. *International Journal of Lexicography* 12 (4). Oxford: Oxford University Press. 307–323.

Čermák, František, 1985: Frazeologie a idiomatika. František Čermák, Josef Filipec: *Česká lexikologie.* Praha: Academia. 166–248.

– – František, 2001: Substance of idioms: perennial problems, lack of data or theory? *International Journal of Lexicography* 14 (1). 1–20

Erbach, Gregor, 1992: Head-Driven Lexical Representation of Idioms in HPSG. M. Everaert, idr. (eds.). *International Conference on Idioms Tilburg, NL: Proceedings of Idioms* I–II.

Gantar, Polona, 2004: Frazem in njegovo besedilno okolje. Doktorska disertacija. Univerza v Ljubljani, Filozofska fakulteta.

Gorjanc, Vojko, Krek, Simon, 2001: A Corpus-Based Dictionary Database as the Source for Compiling Slovene-X Dictionary. *COMPLEX 2001. 6th Conference on Computational Lexicography and Corpus Research: »Computational Lexicography and New EU Languages«*. Birmingham: The University of Birmingham. 41–47.

Gorjanc, Vojko, Krek, Simon, Gantar, Polona, 2005: Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50/2. 3–19.

Hunston, Susan in Francis, Gill, 2000: Pattern Grammar. A corpus-driven approach to the lexical grammar of English. John Benjamins Publishing: Amsterdam, Philadelphia.

Korpus slovenskega jezika FIDA: http://www.fida.net

Krek, Simon, 1999: Računalniški korpusi v slovaropisju. *Razgledi* 13. 14–16.

Kržišnik-Kolšek, Erika, 1986: Poskus razvrstitve stalnih besednih zvez v Trubarjevi Cerkovni ordningi. Obdobja 6, 16. stoletje v slovenskem jeziku, književnosti in kulturi. Ljubljana: Filozofska fakulteta. 435–445.

– – 1988: *Frazeologija v moderni: Magistrsko delo*. Mentorica B. Pogorelec. Ljubljana: Filozofska fakulteta.

Kržišnik, Erika, 1990: Tipologija frazeoloških prenovitev v Cankarjevih proznih besedilih. Slavistična revija 4/38. 399–420.

– – 1994: Slovenski glagolski frazemi (ob primeru frazemov govorjenja). Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Mlacek, Jozef, eds., 1995: *Frazeologická terminológia*. Bratislava: STIMUL – Centrum informatiky a vzdelávania FF UK.

Moon, Rosamund, 1998: *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Oxford: Clarendon Press (Oxford Studies in Lexicography and Lexicology).

Nunberg, Geoffrey, eds., 1994: Idioms. *Language* 3. 491–538.

Teubert, Wolfgang, 1999: Korpuslinguistik und Lexikographie. Deutsche Sprache 4/99. John Benjamins. V: Študije o korpusnem jezikoslovju. V. Gorjanc in S. Krek (eds.). Ljubljana: Krtina. 2005. 103–136.

Toporišič, Jože, 1973/74: K izrazju in tipologiji slovenske frazeologije. *Jezik in slovstvo* 8. 273–279.

Vidovič Muha, Ada, 1988: Nekatere jezikovnosistemske lastnosti strokovnih besednih zvez. XXIV. seminar slovenskega jezika, literature in kulture. Ljubljana: Filozofska fakulteta. 83–91.

Povzetek

Poskusi zamejitve področja frazeologije na podlagi ugotavljanja stopenj idiomatičnosti niso zagotovili enotnega pojmovanja FE, zlasti ne v razmerju do prostih in pomensko transparentnih SBZ. Pojmovanje idiomatičnega tj. od sestavin neodvisnega celostnega pomena zveze, oži področje frazeologije na pomensko netransparetne SBZ, ki so bile navadno predmet specializiranih frazeoloških slovarjev, medtem ko so bile vse druge oblike jezikovnega vzorčenja iz njih izključene, v splošnih slovarjih pa predstavljene neučinkovito in nesistematično. Z izpostavitvijo vidikov besedne povezovalnosti ne glede na stopnjo medsebojne gramatične in pomenske zlitosti ter z možnostjo avtomatičnega urejanja konkordančnih nizov in merjenja besedne povezovalnosti v obliki statističnih izračunov v korpusnem okolju je mogoče oblikovati objektivna izhodišča za prepoznavanje tipičnih besednih kombinacij, ki so v različnih stopnjah leksikalizacijskega procesa tudi slovarsko zanimive. Čeprav se razmerja med eno- in večbesednimi leksikalnimi enotami ter med pomensko transparentnimi in pomensko netransparentnimi SBZ v korpusu zabrisujejo, je s kontekstualno obravnavo leksikalnega elementa v leksikografski praksi mogoče povsem suvereno prikazati tako strukturno kot pomensko širitev besede na raven besedne zveze ali obsežnejšega skladenjskega vzorca.