Variation and Variants in the Dictionary of Multiword Expressions (Focussing on Complex Nominals/Noun Compounds)

Martina Ivanová, Prešov University

The paper deals with the lexicographic description of multiword expressions (MWEs) in Slovak. MWEs are defined as lexicalized word combinations that cohere more strongly than ordinary syntactic combinations: that is, they are lexically, semantically, paradigmatically, syntactically or/and statistically idiosyncratic. The paper focuses on MWEs belonging to the group of complex nominals/noun compounds to show the difficulty with processing variability and variants in lexicographic descriptions. Different types of variants are introduced to show the nature of variability occurring in multiword expressions. Different corpus tools are described which help the researcher to stipulate the lexicographic variants on the basis of reliable statistic data. One of the tools, Word Sketch Difference, a tool that is a part of Sketch Engine, is introduced to show how word sketch scores for semantically close lemmas can help the researcher to process the variants in the dictionary of MWEs.

Keywords: *multiword expression, complex nominal, noun compounds, variability, variants, log dice*

1. Introduction

The aim of this paper is:

- to clarify what multiword expression (MWE) is,
- to briefly describe the *Dictionary of Multiword Expressions in Slovak*,
- to illustrate the nature of variability and variants in MWEs,
- to show the difficulty concerning lexicographical description of MWE variants,
- to manifest how corpus tools can be used to pin down different types of variants of MWEs.

The paper looks at recurrent types of variation and considers some lexicographical consequences concerning processing the variants in the dictionary of MWEs. Variability is evident in corpus data, but it is often under-represented in dictionaries. The present paper considers the interaction between data and lexicographical description. It is based on extensive corpora (Slovak National Corpus, corpus version prim-6.0-public.all, Omnia Slovaca).

2. Multiword Expressions

For Fillmore, Kay & Connor (1988: 502) MWEs introduce a distinction between what a speaker can compute automatically from language (on the bases of grammatical rules) and what he must explicitly store. Sinclair (1991: 109) has called this the distinction between the *open-choice principle* and the *idiom-principle* of language.

What goes under the heading of multiword expression is rather heterogeneous. Under the label “multiword expression” one assumes a wide range of linguistic constructions such as

idioms (*storm in a teacup, sweep under the rug*), fixed phrases (*in vitro, by and large*), nominal compounds (*olive oil, laser printer*), compound verbs (*take a nap, bring about*), etc. There are dozens of other terms for various notions of MWEs, including fixed expressions, formulaic sequences, fossilized units, prefabricated patterns, etc. (Moon 1998; Wray 2000).

MWEs are of great interest to linguists and lexicographers, because of their large number in languages, their peculiar syntactic and semantic behaviour, and their unclear lexical status (Jackendoff 1997; Moon 1998; Pauwels 2000; Fellbaum 2006).

Calzolari et al. (2002: 1934) define MWE as a sequence of words that acts as a single unit at some level of linguistic analysis. In addition they exhibit some or all of the following features: 1. reduced syntactic and/or semantic transparency, 2. reduced compositionality, 3. reduced syntactic flexibility, 4. breach of general syntactic rules, 5. high degree of lexicalisation, 6. high degree of conventionality.

A definition referring to the idiosyncratic nature of MWEs can be found in Sag et al. (2002: 2). According to these authors MWEs are idiosyncratic expressions that cross word boundaries (or spaces). Bauer's (1983) basic definition of MWEs as lexicalised or institutionalised phrases can also be mentioned, where lexicalised phrases include any syntactic, semantic or lexical (i.e. word form) element which is idiosyncratic.

Another definition is given in Sprenger (2003: 4):

> *Fixed expressions refer to specific combinations of two or more words that are typically used to express a specific concept. [...] The defining feature of FE is that it is a word combination stored in Mental Lexicon of native speakers that as a whole refers to a (linguistic) concept. This makes FEs "non-compositional" in the sense that the combination and structure of their elements need not be computed afresh, but can be retrieved from Mental Lexicon. However, the degree of lexical and syntactic fixedness can vary.*

It can be concluded that independent of their lexical fixedness or variability, MWEs possess a holistic quality in the sense that they fulfil a specific role in communication as autonomous language units. They can be characterized by idiosyncratic features, be they lexical, syntactic, semantic, pragmatic and/or statistic, and at one or more of these levels (Kim & Baldwin 2010).

To sum it up, MWEs can be defined as lexicalized word combinations that cohere more strongly than ordinary syntactic combinations: that is, they are lexically, semantically, paradigmatically, syntactically or/and statistically idiosyncratic. The nature and the measure of their idiosyncrasy will be described in the following sections.

2.1 *Semantic idiosyncrasy*

Semantic idiosyncrasy refers to the notion of idiomacity concerning the semantic transparency or semantic compositionality of MWEs. The semantic compositionality of a given MWE is defined as the degree to which the meaning of the whole expression results from combining the meanings of its individual words when they occur in isolation. According to Nunberg et al. Wasow (1994) non-compositional (idiomatic) meaning should not count as a defining criterion for MWEs. The idiomacity of MWEs is scalar, reaching from completely transparent word combinations to completely idiomatized ones.

The definition of the notion of semantic transparency was elaborated for semantic characteristics of complex words (Libben et al 2003; Marelli & Luzzatti 2012). Dressler

(2005: 272) proposes an asymmetric model (i.e. a model that assigns unequal value to the semantic transparency of head and modifier). The concept of asymmetric models can be adopted to MWEs when analysing their semantic structure. Four degrees of semantic transparency in MWEs can be distinguished:

1. transparency of both members of MWE: *pena na holenie* 'shaving foam' "*foam* applied to the face, or wherever else hair grows, to facilitate *shaving*", *zbraň hromadného ničenia* '*weapon of mass destruction*' "a nuclear, radiological, chemical, biological or other *weapon* that can kill and bring significant harm to *a large number* of humans or cause great damage to man-made structures (e.g. buildings), natural structures (e.g. mountains), or the biosphere";

2. transparency of the head component, opacity of the modifying component of MWE: *vysoká pec* 'blast furnace' "type of metallurgical *furnace* used for smelting to produce industrial metals, generally iron, but also others such as lead or copper", *biela káva* 'white coffee' "regular black *coffee* that has had milk or cream added to it";

3. transparency of the modifying component, opacity of the head component of MWE: *československá jar* 'Czechoslovak Spring' "period of revival processes in the former *Czechoslovakia* from the end of 1967 until the military Warsaw Pact invasion of Czechoslovakia", *tepelný most* 'thermal bridge' "part of the construction in which extensive *heat* penetration occurs";

4. opacity of both members of MWE: *pastierska kapsička* 'shepherd's pouch' "white-flowered annual European herb bearing triangular notched pods", *biely trpaslík* 'white dwarf' "stellar remnant composed mostly of electron-degenerate matter".

To summarize, we have identified four classes of MWEs on the continuum of semantic idiosyncrasy. Within this theoretical approach transparency of the head component is assigned a higher value as it provides a greater number of important characteristics for semantic description of the whole MWE.

2.2 *Syntactic Idiosyncrasy*

MWEs are traditionally viewed as syntactically fixed expressions. However, it is widely accepted that the criterion of syntactic fixedness is not a defining criterion of MWEs. In theoretical literature it is recognized that MWEs are restricted with regards to some syntactic operations. For example, the Dutch *rode kool* 'red cabbage' allows neither the modification of the adjective by a measure adverb nor the insertion of another prenominal adjective, cf. **erg rode kool* 'very red cabbage', **rode dure kool* 'red expensive cabbage', cf. Booij (2009). However, the first criterion applies not only to MWEs but also to free word combinations in which the modifying adjective has the status of relational adjective, e.g. *masová účasť* 'mass attendance' – **veľmi masová účasť* 'very mass attendance' (syntactic phrase); *masové médium* 'mass medium' – **veľmi masové medium* 'very mass medium' (MWE). The second criterion can be addressed too as there are specific rules concerning the word order of prenominal adjectives also in free word combinations in Slovak, e.g. *drahé červené pero* 'expensive red pen' – **červené drahé pero* 'red expensive pen' (syntactic phrase); *drahá*

červená kapusta ‘expensive red cabbage’ – **červená drahá kapusta* ‘red expensive cabbage’ (MWE).

There are also some limits concerning the change of word order. Within certain groups of MWEs certain word order structures are preferred, either A + N *zubná kefka ‘toothbrush’*, or N + A *vrabec domový* ‘sparrow’. Nevertheless, there are also MWEs which have word order variants, e.g. *anjel strážny – strážny anjel* ‘guardian angel’, *gama lúče – lúče gama* ‘gamma rays’.

2.3 *Paradigmatic Idiosyncrasy*

Paradigmatic idiosyncrasy involves restrictions with regard to some paradigmatic operations. It refers to the fact that the parts of a given lexicalized (multiword) expression cannot be substituted by another word of similar meaning without losing its semantic integrity. This phenomenon is also referred to as non-substitutability (Manning & Schütze 1999).

The components of an MWE cannot be freely substituted by their synonymous or antonymous counterparts, e.g. *skúška správnosti – *test správnosti* ‘true-false test’, *rozličný tovar – *rozmanitý tovar* ‘various goods’, *prvá pomoc – *posledná pomoc* ‘first aid’ – *‘last aid’, *nový roman – starý roman* (‘nouveau roman’ – * ‘old roman’), whereas free syntagmas do not show such kind of restrictions, e.g. *rozličné predmety – rozmanité predmety* ‘various items*’* – ‘diverse items’*, rozličné povahy – rozmanité povahy* ‘different characters*’* – *‘*manifold characters’)*, rozličné rady – rozmanité rady* (‘different advice’ – ‘variable advice’, *nový hotel – starý hotel* ‘new hotel*’* – ‘old hotel’), *prvý návštevník – posledný návštevník* ‘first visitor’ – ‘last visitor’), (cf. Ološtiak 2011).

2.4 *Lexical Idiosyncrasy*

Lexical idiosyncrasy can be attested for some MWEs. First, it concerns foreign phrases adopted as a whole into Slovak in which no component exists independently in Slovak, e.g. *avant la lettre, art brut, art déco, paso doble, fin de siécle, laterna magica.* Another type is represented by MWEs which include at least one foreign component non-existing in Slovak outside a MWE, e.g. *steel* in *steel gitara* ‘steel guitar’, *head-up* in *head-up displej* ‘head-up display’.

Lexical idiosyncrasy can be also identified for those MWEs which contain at least one monocollocable component, e.g. *mimoúrovňový* in *mimoúrovňová križovatka* ‘interchanges’, *gregoriánsky* in *gregoriánsky chorál* ‘Gregorian chant’, *gregoriánsky kalendár* ‘Gregorian calendar’. Monocollocable words can be defined as words whose usage is severely restricted to one or a few combinations only.

2.5 *Morphological Idiosyncrasy*

When an MWE includes a noun as its component, this noun typically occurs both in singular or plural form, e.g. *vysoká škola – vysoké školy* ‘university/universities’, *kyslý dážď – kyslé dažde* ‘acid rain(s)’. However, some MWEs often limit the possibilities to only one of those in spite of the fact that the given noun behaves as countable outside an MWE, e.g. *akcie na doručiteľa* ‘bearer shares’*, zimné pneumatiky* ‘winter tyres’*, letné pnemumatiky* ‘summer tyres’*, pivné kvasnice* ‘brewer's yeast’*, zemiakové lupienky* ‘potato chips’*, čínske paličky* ‘chopsticks’ occur only in plural in Slovak.

2.6 *Statistical Idiosyncrasy*

Corpus studies – especially computer-aided corpus studies – reveal much more reliably than native speaker intuition that many combinations of units in language tend to recur. The frequency of occurrence of particular word combination within the same immediate context is an empirically verifiable feature of collocations. Many definitions of collocations incorporate some notion of frequency or recurrence as a defining feature of collocations.

As opposed to collocation, MWEs' identification in corpus needs specialized tools. Data obtained from a corpus enable one to set the association scores as a measure of attraction between words. The most common score defines co-occurrence by surface proximity measured by number of particular word combination tokens. Nevertheless, although MWEs are often employed in general and in technical language, their automatic identification based on association measures is often limited by their low token frequency in standard corpora.

For testing tan association measure, different tools have been developed. MI-score is a measure of how strongly two words seem to associate in a corpus, based on the independent relative frequency of two words. T-score is a measure of how certain we can be that the collocation is the result of more than the vagaries of a particular corpus. According to Křen (2006) MI-score tends to identify non-conventionalized or even random collocations whereas, on the basis of t-score more systematic, conventionalized collocations can be captured.

To illustrate practical problems of using the above-mentioned corpus tools to detect MWEs, i.e. conventionalized, systematic, lexicalized collocations, the unit *futbal* 'football' with its adjectival collocates has been investigated. The results are presented in Table 1.

Table 1: The corpus frequency of the lemma *futbal* collocates (SNK, prim-6.0-public-all)

lemma	**MI-score**	**t-score**	**absolute frequency**
slovenský 'Slovak'	6.388	77,17	6100
malý 'small'	7.098	66,76	4522
dobrý 'good'	5.748	56,69	3372
americký 'American'	6.868	46,93	2241
útočný 'offensive'	10.59	41,18	1698
svetový 'world'	6.429	38,04	1481
pekný 'nice'	7.12	37,83	1452
sálový 'indoor'	13.8	35,04	1228

The statistics presented in Table 1 shows that it is probably impossible to choose a single most appropriate association measure for detecting MWEs in the corpus. The non-lexicalized word combination *slovenský futbal* 'Slovak football' has a similar frequency as the MWE *americký futbal* 'American football'. These word combinations also have a similar MI-score. As to the t-score, similar statistical results can be seen for the non-lexicalized word combination *slovenský futbal* 'Slovak football' and the MWE *malý futbal* 'minifootball'; the same holds for *útočný futbal* 'offensive football' as a non-lexicalized word combination and *americký futbal* 'American football' as an MWE, *pekný futbal* 'nice football' as a non-lexicalized word combination and *sálový futbal* 'indoor football' as an MWE.

3. Variation in the Dictionary of Multiword Expressions

On the basis of semantic, syntactic, paradigmatic, lexical and morphological idiosyncrasies the most frequent collocates are sorted out and those which can be subsumed under the label of MWE are included in the prepared *Dictionary of Multiword Expressions in Slovak*. In the dictionary, MWEs labelled as (i) noun compounds, complex nominals, e.g. *olivový olej* 'olive oil', or multiword terminology, e.g. *umelá inteligencia* 'artificial intelligence', and (ii) verb compounds termed also as verbo-nominal expressions or light verb constructions, e.g. *dať príkaz* 'to give an order' are processed.

Linguists have proposed various definitions of multiword expressions based on their fixed characteristics. It is one of the most widely-held assumptions in linguistics that fixed expressions cannot be modified. This feature is mentioned as non-modifiability in Manning & Schütze (1999).

In fact, we can see a lot of multiword expressions violating the principles of their fixed characteristics. Linguistic variability can be counted among the major properties of MWEs and it can appear on different levels.

3.1 *Variation and synonymy*

Numerous extensive studies have been carried out on variation and its differentiation from synonymy. In order to make a clear distinction between variants and synonymous MWEs, two criteria are focused on: meanings and structural characteristics.

MWEs are said to be synonymous when they have the same content but different components highlighting different aspects of their semantic structure. Synonymous MWEs arise either as a result of borrowing units from foreign languages or due to selection of two different constituents as onomasiological marks to form two different MWEs.

The former procedure can be illustrated with a number of adopted MWEs having Slovak counterparts, e.g. *mail art – poštové umenie*, *obligačné právo – záväzkové právo* 'bond law', *masovokomunikačné prostriedky – hromadné oznamovacie prostriedky* 'mass media', *termonukleárna reakcia – termojadrová reakcia* 'thermonuclear reaction', *hard rock – tvrdý rock*.

The latter process is responsible for forming synonymous MWEs having the same content based on different images. It is the result of selection of some semantic components from the semantic structure of a unit to function as an onomasiological mark in the structure of an MWE, e.g. *plynový mechúr* 'gas bladder' – *vzduchový mechúr* 'air bladder' "internal *air-filled* organ that contributes to the ability of a fish to control its buoyancy, enables to equalize *gas pressure* in the body of the fish to the aqueous medium external pressure and thus to stay at its current water depth without having to waste energy in swimming" (two semantic components from the semantic structure of the unit have been chosen to act as onomasiological marks in an MWE: *air-filled, gas pressure*), *vševediaci rozprávač* 'omniscient narrator' – *autorský rozprávač* 'author's narrator' "the voice in which a story is written that is, *similarly to author*, outside the story and that *knows everything* about the characters and events in the story" (two semantic components have been chosen to function as onomasiological marks in an MWE: *similarly to author, knows everything*).

Some synonymous MWEs result from both the adoption process and the variable onomasiological selection, e.g. *zložené oko* 'compound eye' – *facetové oko* 'facet eye' "eye *composed of* many simple *facets* which, depending on the details of anatomy, may give either

a single pixelated image or multiple images, per eye" (the semantic component *composed of* is expressed using the Slovak item, *facets* is expressed using the loaned item).

3.2 *Types of MWE Variants*

Variants in the *Dictionary of Multiword Expressions in Slovak* come in quite a range of types. The overview below presents a possible classification of variation types occurring among Slovak MWEs:

I syntagmatic
1 quantitative
a additive: *sťahovanie národov – veľké sťahovanie národov* 'migration period'
b reductional: *perodická sústava prvkov – periodická sústava* 'periodic table'
2 qualitative
a with construction change: *alergia na peľ – peľová alergia* 'pollen allergy'
b without construction change: *anjel strážny – strážny anjel* 'guardian angel'
II paradigmatic
a orthographic: *Versaillský systém – versailleský systém* 'Versailles system'
b phonematic: *segedínsky guláš – segedínsky guľáš* 'Szegedin goulash'
c morphological: *borovica limba – borovica limbová* 'stone pine'
d word-formation: *žalúdkový/žalúdočný vred* 'peptic ulcer'

I Syntagmatic variants
Many MWEs allow variations concerning their syntactic structures with regard to quantitative changes (processes of extension/reduction), or qualitative changes (construction change and change of word order).

I1) Extension/reduction variants
A special type of syntactic variability is represented by extension/reduction variants of MWEs. It concerns MWEs with facultative component the elimination or addition of which does not violate the integrity of an MWE, e.g. *periodická sústava prvkov/periodická sústava* 'periodic table of elements/periodic table', *bodové odporové zváranie/bodové zváranie* \resistance spot welding/spot welding', *subtropický dažďový les/subtropický les* 'subtropical rain forest/subtropical forest', *štátna záverečná skúška/štátna skúška* 'final state exam/state exam'.

I2a) Construction variants
Construction variants arise as the result of change in a syntactic structure, they usually concern different syntactic codings of the modifying element in the form of either prenominal adjectival attribute or post-nominal noun attribute, e.g. *peľová alergia/alergia na peľ* 'pollen allergy', *korelačný koeficient/koeficient korelácie* 'correlation coefficient', *bielizňový kôš/kôš na bielizeň* 'laundry basket'. Structures with prenominal adjectival attribute dispose of a high degree of condensability, those with post-nominal noun attribute dispose of a higher degree of explicitness.

I2b) Word order variants
For Slovak MWEs, the change of word order of adjectival components is attested especially in the names of Earth periods or historical periods, e.g. *doba ľadová/ľadová doba* 'Ice Age', *doba bronzová/bronzová doba* 'Bronze Age'and religious terms, e.g. *anjel strážny/strážny anjel* 'guardian angel', *litánie loretánske/loretánske litánie* 'Litany of Loreto/Loreto litanies'. The post-nominal position of the congruent adjective can be explained on the basis of the Latin influence on Slovak in its historical development.

II) Paradigmatic variants
Many MWEs allow variations concerning their paradigmatic dimension. Paradigmatic variation relates to the possibility of replacing a MWE component with a paradigmatically related component so that the semantic integrity of the MWE is preserved.

IIa) Ortographic variants
Orthographic variability in Slovak MWEs usually has two sources: 1. Slovak imports many words from other languages using transliteration resulting in the situation that two possible forms (with an original and an adopted spellings) coexist, 2. The source of orthographic variability is rooted in existence of different forms with regards to Slovak orthographic rules.
(1) Orthographic variants arise in the processes of adaptation of loan words in Slovak resulting in coexistence of an original and an adopted forms, e.g. *jazzový/džezový vek* 'jazz age'.
(2) Orthographic variability based on parallel existence of two forms with regard to the existing orthographic standards usually concerns orthographic rules referring to the way of writing capital letters. In Slovak, orthographic variants are attested in MWEs containing adjectives from the religious sphere *Boží/boží*, e.g. *Božia muka/božia muka* 'wayside cross', *služby Božie/božie* 'worship services'.

IIb) Phonematic variants
Phonematic variants differ in one or more phonemes preserving the formal and semantic identity of a lexeme, e.g. *segedínsky guláš/guľáš* 'Szegedin goulash' (cf. Jarošová 2009). Phonematic variants may arise in the processes of adaptation of loan words into Slovak, e.g. *projektový manažér/menežér* 'project manager'. With regard to lexicographic practice there are two possible approaches: to incorporate into a dictionary only correctly spelled words and, on the other hand, to inform the user about the actual usage of the units (as attested in the corpus). To overcome these contradictory tendencies the notion of graded variation can be introduced. When marking a word as a variant of another, it is either classified as a fully equivalent variant or as a non-preferred variant, or even a no longer existing variant. This classification can be used when analysing forms that occur in a corpus, e.g. the form *menežér* is not found in a dictionary; it is treated as a non-existing variant.

Another source of phonetic variability is caused by truncation. In Slovak, truncation variants are typical for adjectives formed from geographical names. It is usually the phoneme *-g-* that can be deleted in the adjectival component of MWEs, e.g. *pekingský/pekinský palácový psík* 'Pekingese/Pekinese', *hongkongská/hongkonská chrípka* 'Hong Kong flu'. Truncation often occurs in cases where two vocalic phonemes combine on morphemic boundaries; in such cases one of these vocalic phonemes is usually deleted, e.g. *letná paraolympiáda/ paralympiáda* 'summer Paralympics'.

IIc) Morphological variants

Morphological variation can be delimited on the basis of different parts of speech used in the same constructional type of MWEs. It usually applies to nomenclature names in which the modifying attribute in post-nominal position can be expressed either using an adjective or a noun in Slovak, e.g. *repka olejná*$_{ADJ}$*/olejka*$_{NOUN}$ 'rape, Brassica napus', *borovica limba*$_{NOUN}$*/limbová*$_{ADJ}$ 'stone pine', *borovica sosna*$_{NOUN}$*/sosnová*$_{ADJ}$ 'Scots pine'.

Another source of morphological variation is connected to different morphological forms of adjectives. In the MWE *Morseho/Morseova abeceda* 'Morse code' two different types of possessive adjectives (the form *Morseho* belongs to the declention type *pekný*, the form *Morseova* to the declention type *otcov*) are used.

IId) Word-formation variants

In Slovak, there exist frequent rival pairs or even groups of adjectival derivatives entering MWEs based on the systematic competition of domestic word-formation types (Nábělková 1996: 258). Competition exists especially between productive formants *-ový/-ný* and other specialized formants, e.g. *piesková/piesočná/piesočnatá pláž* 'sandy beach', *meračský/merací prístroj* 'measuring instrument, meter', *herný/hrací plan* 'game plan', *sprchová/sprchovacia hlavica* 'shower head' and between productive formants with each other *vojenský/vojnový konflikt* 'military conflict', *rozvodná/rozvodový skriňa* 'switchboard', *pevninový/pevninský ľadovec* 'continental glacier'. There is also a special type of variant caused by competition of word-formation types with domestic formants as opposed to foreign formants, e.g. *telefónna/telefonická linka* 'telephone line'.

This special type of variants is represented by competition of derivational and compound adjectives used as components of MWEs. Adjective compounds explicitly express semantic information which is only implicitly expressed in adjective derivatives, e.g. *denný/celodenný lístok* 'day ticket/whole day ticket' "transport ticket valid *the whole day*" (the semantic component 'the whole day' is expressed in the form of a compound adjective), *farebná/viacfarebná mapa* 'colour map/multi-colour map' "a picture or chart that shows the rivers, mountains, streets, etc., in a particular area using *more colours*" (the semantic component 'more colour' is expressed by the compound adjective).

In Slovak, there also exist frequent rival pairs or even groups of nominal derivates entering MWEs which belong to the same word-formation type and share the same onomasiological meaning. Competition exists especially between productive formants and other specialized formants, e.g. *platca/platiteľ DPH* 'VAT payer'. Another source of variability is given by competition of different word-formation procedures, e.g. transflection (a word-formation procedure common in Slavic languages in which a new word is coined by a change of grammatical morpheme) vs. suffixation, e.g. *priama úmera/úmernosť* 'direct correlation', *nepriama úmera/úmernosť* 'inverse correlation'. Frequently, word-formation variants arise as the result of both perfective and imperfective forms of the same motivating verb, e.g. *výjazdové zasadanie/zasadnutie* 'external meeting'. A special type of variants is represented by competition of derivational and compound nouns. Nominal compounds explicitly express semantic information which is only implicitly expressed in noun derivatives, e.g. *bezpodielové vlastníctvo/spoluvlastníctvo* 'joint ownership, joint tenancy' "a type of ownership of real or personal property *by two or more persons* in which each owns an undivided interest in the whole" (the semantic component 'by two or more persons' is expressed in the compound noun).

4. Identifying MWE Variants in the Dictionary

Identifying variants of MWEs is not a simple task for lexicography. It is not easy to stipulate which MWEs should be processed with variants. One of the tools that can be used to solve this problem is the *Word Sketch Difference*, a tool that is a part of the *Sketch Engine*, an example of a syntax-based concordance program (Kilgarriff et al. 2004).

Sketch differences in the *Sketch Engine* specify for two semantically related words what behaviour they share and how they differ. Semantically close units tend to share some, but not all collocates. The sketch differences show the patterns which are shared by both semantically close words; they also provide information in a colour scheme for the user to grasp immediately whether and where the lemmas are semantically similar with respect to the collocates they choose. Sketch Difference is a neat way of comparing two very similar words: it shows those patterns and combinations that two items have in common, and also those patterns and combinations that are more typical of, or unique to, one word rather than the other.

To measure the similarity of two lemmas with regards to the collocates they choose, the *logDice* function is used. It is based not only on the frequency of a particular relation, but also on the frequency of the headword in the same syntactic position (with any collocate) and the frequencies of collocates (in any syntactic position), cf. Rychlý (2008). Values of the *logDice* have the following features: (1) Theoretical maximum is 14 if all occurrences of X co-occur with Y and all occurrences of Y co-occur with X. Usually the value is less than 10. (2) Value 0 means there is less than 1 co-occurrence of XY per 16,000 X or 16,000 Y. It can be suggested that negative values indicate no statistical significance of XY collocation. (3) Comparing two scores, plus 1 point means a double frequency of a collocation, plus 7 points means roughly 100 times higher frequency of a collocation. (4) The score does not depend on the total size of a corpus. The score combines relative frequencies of XY in relation to X and Y.; (5) By comparing the value of *logDice*, the collocation preferences of two lemmas can be stipulated. If the difference between the *logDice* of two lemmas is between 6.0 and 2.0, no variant is proposed for a particular lemma if the difference is between 2.0 and 0; variants are present in a dictionary.

To illustrate this method the word sketch scores for the semantically close lemmas *vojenský* and *vojnový* 'military' vs. 'related to war' are presented in Table 2 and Table 3, the third column indicates the value of *logDice*.

Table 2: The frequency of nominal collocates of adjectives *vojenský* in Sketch Engine.

služba	15099	4.78
operácia	5797	6.07
základňa	5784	6.83
sila	5672	4.26
jednotka	5130	5.12
technika	4403	5.04
nemocnica	4295	4.79
súd	3645	3.92
polícia	3566	4.61
akcia	3239	3.33

spravodajstvo	3152	6.85
lietadlo	3101	5.12
konflikt	3055	5.48
akadémia	3039	5.75
správa	2797	2.81
história	2738	4.21
škola	2688	1.99
prokuratúra	2609	6.2
tabor	2587	5.06
cintorín	2539	5.75
zásah	2492	5.01
obvod	2425	5.61
cvičenie	2423	5.0
útvar	2414	5.91
priestor	2347	2.39
material	2261	3.21
letectvo	2217	6.88
veliteľ	2157	5.76
posádka	1668	5.34
intervencia	1612	6.38
ústav	1522	4.41

Table 3: The frequency of nominal collocates of adjectives *vojnový* in Sketch Engine.

konflikt	5552	6.52
zločin	5453	7.26
zločinec	3385	8.03
cintorín	3024	6.25
stav	2769	2.93
veterán	2679	7.57
štát	2478	2.4
loď	2357	4.7
zajatec	2341	8.24
udalosť	1957	3.89
rok	1885	-0.98
hrob	1850	5.54
film	1254	1.87
dráma	971	5.29
čas	913	0.44
obdobie	881	1.11
republika	825	1.67
sekera	766	6.02
korisť	666	6.08
námorníctvo	566	6.63
hrdina	559	3.42
výprava	548	4.17

zóna	524	3.09
ťaženie	486	5.89
štváč	426	7.17
hra	416	-0.02
operácia	400	2.35
obeť	378	2.23
hrôza	357	4.25
kríž	356	2.69
škoda	341	1.55

The value of *logDice* for the collocation *vojenská operácia* is 6.07, the *logDice* for *vojnová operácia* is 2.35, the difference is 3.72 in favour of *vojenská operácia* which means that in this case, no variant of the MWE *vojenská operácia* 'military operation' is proposed. The value of *logDice* for the collocation *vojenský konflikt* is 5.48, the *logDice* for *vojnový konflikt* is 6.52, the difference is 1.04 in favour of *vojnový konflikt* which means that, in this case, the two variants *vojnový konflikt/vojenský konflikt* 'military conflict' are proposed for a dictionary.

This is visualised in Table 4. The green colour means that MWEs with the adjective *vojenský* are proposed, e.g. *vojenská intervencia* 'military intervention', *vojenská prokuratúra* 'military prosecution', *vojenský útvar* 'military formation', *vojenská hodnosť* 'military rank', *vojenská rozviedka* 'military intelligence service', *vojenská základňa* 'military base', *vojenská junta* 'military junta'; red colour means that MWEs with the adjective *vojnový* are proposed, e.g. *vojnový zločin* 'war crime', *vojnový zločinec* 'war criminal'; white colour indicates the cases in which variants of MWEs are proposed, e.g. *vojenské/vojnové ťaženie* 'military campaign', *vojenské/vojnové námorníctvo* 'marine', *vojnový/vojenský cintorín* 'military cemetery', *vojnový/vojenský konflikt* 'military conflict'.

Table 4: The frequency of nominal collocates of *vojenský* and *vojnový* according to Sketch Difference.

XNn	**336,592**	**87,362**	**1.4**	**1.4**
junta	983	0	6,5	--
intervencia	1,612	0	6,4	--
prokuratúra	2,609	0	6,2	--
kontrarozviedka	841	0	6,2	--
diktatúra	976	0	6,0	--
útvar	2,414	0	5,9	--
hodnosť	1,028	0	5,8	--
rozviedka	662	0	5,8	--
základňa	5,784	23	6,8	-0,9
prevrat	1,249	11	6,1	0,1
uniforma	1,149	12	5,8	-0,2
veliteľ	2,157	52	5,8	0,7
spravodajstvo	3,152	70	6,8	1,8
letectvo	2,217	65	6,9	2,5
operácia	5,797	400	6,1	2,3
ťaženie	900	486	5,8	5,9

námorníctvo	1,036	566	6,3	6,6
cintorín	2,539	3,024	5,7	6,3
konflikt	3,055	5,552	5,5	6,5
mašinéria	289	264	4,5	5,7
loďstvo	144	299	3,6	6,2
veterán	513	2,679	4,6	7,6
invalid	49	306	1,9	5,7
zajatec	250	2,341	4,0	8,2
korisť	49	666	1,5	6,1
zločin	97	5,453	1,2	7,3
sekera	18	766	-0,1	6,0
zločinec	21	3,385	0,0	8,0
reparácia|reparácie	0	227	--	6,2
štváč	0	426	--	7,2

5. Conclusions

Lexical variation within MWEs raises many theoretical and practical problems. This paper brings a theoretical insight into the conditions and factors surrounding lexical variability and variants of MWEs in Slovak. It describes diverse types of variants of MWEs which can be identified when processing a dictionary of MWEs. It has been demonstrated that corpus tools can be useful to identify the status of variants on the basis of frequency information that enables users to develop a framework allowing for a 'neutral', i.e. non-stigmatizing, description of linguistic variants in dictionaries of MWEs.

Acknowledgements
This work was supported (50 %) by the Slovak Research and Development Agency on the basis of the contract n. APVV-0342-11 *Dictionary of Multiword Expressions (Lexicographic, Lexicological and Comparative Research)* and (50 %) by the European Science Foundation (ESF) COST Action n. IS1305 *European Network of e-Lexicography*.

References

Bauer, Laurie. 1983. *English Word-formation.* Cambridge: Cambridge University Press.

Booij, Geert. 2009. Phrasal names: a constructionist analysis. *Word Structures* 2. 219–240.

Buzássyová, Klára & Jarošová, Alexandra (eds.). 2006. *Slovník súčasného slovenského jazyka. A – G. 1st vol.* Bratislava: Veda.

Calzolari, Nicoletta & Fillmore, Charles & Grishman, Ralph & Ide, Nancy & Lenci, Alessandro. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of HLT 2002 (Human Language Technology Conference), San Diego, California, March 2002,* 1934–1940.

Dressler, Wolfgang U. 2005. Compound types. In Libben, Gary & Jarema, Gonia (eds.), *The Representation and Processing of Compound Words,* 23–44. Oxford: Oxford University Press.

Fazly, Afsaneh & Stevenson, Suzanne. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *MWE '07 Proceedings of the Workshop on a Broader Perspective on Multiword Expressions,* 9–16. Stroudsburg: Association for Computational Linguistics.

Fellbaum, Christiane. 2006. Corpus-based studies of German idioms and light verbs. *International Journal of Lexicography* 19. 349–360.

Fillmore, Charles J. & Kay, Paul & O'Connor, Mary Catherine. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language* 64. 501–538.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty.* Cambridge, MA: The MIT Press.

Jarošová, Alexandra. 2009. Problematika lexikálnej variantnosti a spôsoby jej lexikálneho zachytenia. In Považaj, Matej (ed.). *Dynamické tendencie v slovenskom pravopise,* 98–134. Bratislava: Veda.

Jarošová, Alexandra & Buzássyová, Klára (eds.). 2011. *Slovník súčasného slovenského jazyka. H – L. 2nd vol.* Bratislava: Veda.

Kilgarriff, Adam & Rychlý, Pavel & Smrz, Pavel & Tugwell, David. 2004. The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress,* 105–116. Lorient: Universite de Bretagne-Sud.

Křen, Michal. 2006. Kolokační míry a čeština: srovnání na datech ČNK. In Čermák, František & Šulc, Martin (eds.). *Kolokace,* 223–248. Praha: Nakladatelství Lidové noviny, Ústav Českého národního korpusu.

Libben, Gary & Gibson, Martha & Yoon, Yea Bom & Sandra, Dominiek. 2003. Compound fracture: the role of semantic transparency and morphological headedness. *Brain and Language* 84. 50–64.

Manning, Christopher & Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

Marelli Marco & Luzzatti, Claudio. 2012. Frequency effects in the processing of Italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language* 66. 644–664.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Oxford: Oxford University Press.

Nábělková, Mira. 1996. Adjectival Variants in Monolingual Dictionaries. In Gellerstan, Martin & Järborg, Jerker & Malmgren, Sven-Göran & Norén, Kerstin & Rogström, Lena & Röjder Papmehl, Catarina (eds.). *Proceedings 1 – 2. Paper submitted to the Seventh Euralex International Congress on Lexicography in Göteborg. Sweden. Part 1,* 257–263. Göteborg: University Department of Swedish.

Nunberg, Geoffrey & Sag, Ivan A. & Wasow, Thomas. 1994. Idioms. *Language* 70. 491–538.

Ološtiak, Martin. 2011. *Aspekty teórie lexikálnej motivácie.* Prešov: Filozofická fakulta PU v Prešove.

Pauwels, Paul. 2000. *Put, Set, Lay and Place: A Cognitive Linguistic Approach to Verbal Meaning*. Munich: Lincom Europa.

Rychlý, Pavel. 2008. A lexicographer-friendly association score. In Sojka, Petr & Horák, Aleš (eds.). *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008*, 6–9. Brno: Masaryk University.

Sag, Ivan A. & Baldwin, Timothy & Bond, Francis & Copestake, Ann A. & Flickinger, Dan. 2002. Multiword expressions: A pain in the neck for NLP. In *CICLing '02 Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, 1–15. London: Springer-Verlag.

Sinclair, John. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Slovenský národný korpus – verzia prim-6.0-public-all. 2014. Bratislava: Jazykovedný ústav Ľ. Štúra SAV. (http://korpus.juls.savba.sk). (Accessed 2016-04-22).

Sprenger, Simone A. 2003. *Fixed expressions and the production of idioms*. Nijmegen: Radboud University Nijmegen. (Doctoral dissertation)

Wray, Alison. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics* 21. 463–489.

Martina Ivanová
Institute of Slovak and Media Studies
Faculty of Arts
Prešov University
Prešov
Slovakia
martina.ivanova@unipo.sk

In SKASE Journal of Theoretical Linguistics [online]. 2016, vol. 13, no.3 [cit. 2016-12-19]. Available on web page http://www.skase.sk/Volumes/JTL33/pdf_doc/02.pdf. ISSN 1336-782X.