Collocability and contextualization in the process of acquiring lexis

Jiří Dvořák a Daniela Dvořáková

Abstract: Significant changes have been taking place in the way lexis is acquired by language learners as a result of new technology for data-processing and text-based research. Dictionaries are founded on authentic usages of words, their collocations, and semantic domains they are associated with. However, large, complex units of meaning are often absent from dictionaries and terminological glossaries. The importance of collocability and contextualization is highlighted in the contribution. Although study of lexis in collocations and particular contexts is not an overarching concept in language training, the authors believe it is beneficial not only for the needs of students' linguistic training in general, but also for the needs of experts working with the terminology of their specializations.

Key words: collocability, collocation, context, corpus, frequency, management

Introduction

One approach to the acquisition of lexis puts emphasis on structure, while the other is aimed at language use, that is, when the occurrences of similar structures in different contexts are investigated. The question to be answered is whether there are different preferences for some structures over others. The analysis of frequent collocations may not only provide answers to the question above, but also serve as empirical evidence of possible misconceptions about language use. Contextualization is often called on for help when decoding the meanings of words.

Collocations

The concept of collocation and its use is the central concept of linguistic training. Attention paid to lexical combinations in particular contexts is enabled by the development of corpora, which are becoming more extensive and more available. The term collocation is often defined differently. As it is stated by Křen, "je to často pojem velice široký a může zahrnovat různé jazykové jevy od odborných termínů přes víceslovné předložky, frazémy a idiomy, až po nejrůznější statisticky významné souvýskyty slov, lingvisticky nepříliš snadno popsatelné." [it is often a broad term and may include various linguistic phenomena from specialized terms, multi-word prepositions, phrasemes and idioms, to various statistically significant co-occurrences of words, which are not quite easy to be described linguistically.] (Křen 2006: 223; translated by authors) Melčuk distinguishes *idiomacity* from *stability of collocation* (i.e. the probability of a constituent appearing with other constituents) and thus classifies "four basic types of collocations: 1) stable

and idiomatic, 2) stable and non-idiomatic, 3) non-stable and idiomatic, 4) non-stable and non-idiomatic." (quoted in Lipka 1972: 78)

Collocation in corpus linguistics is defined broadly as a combination of words that co-occur, the meaning of which is based on the meaning of its components. Whether or not a phrase has become a fixed expression, a collocation, a unit of meaning, is revealed by paraphrases. Teubert and Čermáková state that the paraphrases of *friendly fire* do not tell us what *friendly* means, they explain what *friendly fire* is. (quoted in Halliday, Teubert, Yallop & Čermáková 2004: 149) While collocations or fixed expressions are established on the basis of usage, paraphrases tell us whether indeed they are understood as units of meaning. The same authors propose the concept of the single word to be replaced by the new concept of a collocation or a unit of meaning in order to overcome the polysemy of lexical items in dictionaries. Thus the ambiguity in traditional linguistics will disappear. (ibid., 151) Lipka (1992: 9) believes that "it is impossible to capture the exact differences of meaning, unless we consider the combining potential of these words, the so-called collocations."

The meaning of a term results from its usage in collocations. Such usage may sometimes be overgeneralized, however. This overgeneralization is due to the fact that the normally accepted meaning of a word is based only on a few of its occurrences. Furthermore, the assumption that terms should have identical meanings, unaffected by context, is often contradicted with the fact that the same terms have different meanings supported by different definitions in various branches of science and their sub-fields. There is also an increasing transition of terminology into general language and vice versa.

Čermák sees that "úhrn kolokací daného lexému, od nejčastějších a opakovaných (jádro) až po řídké a příležitostné (periférie, nové metafory), představuje kombinatorický profil takového lexému, …" [The summary of collocations of a given lexeme from the most frequent and repeated collocations (core) to the rare and occasional ones (periphery, new metaphors) represents a combinatorial profile of such a lexeme.] (Čermák 2006: 57; translated by authors) Collocability is the formal and semantic ability of a lexical unit to be connected with other lexical units. The collocation relation is described by Jackson and Amvela (2004: 114) in statistical terms as "greater than chance likelihood of occurrence." From this they conclude that "the mutual expectancy of two words could be stronger or weaker, depending on both the direction of expectancy and the number of alternative predictable words." (ibid., 114) The higher the co-occurrence of words, the closer the connection among them is. The collocability of most words is quite broad, but never unlimited. The term collocate means a word which joins the initial word as a component of a given collocation.

According to Cruse (1986: 40),

> *the term collocation refers to sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent. …The semantic integrity or cohesion of a collocation is more marked if the meaning carried out by one or more of its constituent elements is highly restricted contextually, and different from its meaning in more neutral contexts.*

Expressions such as *safety catch and heavy casualties* are examples of collocations. However, the sense of *safety* in *safety catch* and *heavy* in *heavy casualties* requires specific contextual conditions as opposed to *safety* in *safety equipment* and *heavy* in *heavy backpack*. Collocations whose constituents do not like to be separated are termed by Cruse as "bound collocations" (ibid., 41) and by Hnátková as "monocollocability". (Hnátková 2006: 143) An example of such a bound collocation might be *non-commissioned officer* (NCO).

Collocations vs. free combination of words

Bearing in mind the fact how broadly the term *collocation* is defined it is sometimes difficult to distinguish collocation from a free combination of words, which might be equally important in the process of language acquisition. However, drawing a subjective dividing line between collocations and free combinations of words would be of no benefit to anyone concerned. Thus, we use the term *collocation* in a broad sense as sequences of lexical items which habitually co-occur and we also recommend to pay attention to any frequent free combinations of words during language training, because, as Partington, Duguid and Taylor (2013: 8) claim, "if something is seen to happen frequently in a language, then it is significant." Moreover, what is perceived as being a free combination by one linguist may be perceived as being a weak collocation by another. As it is written in the *Oxford Collocations Dictionary for Students of English* (2002: vii),

> *Combinations of words in a language can be ranged on a cline from the totally free—see a man—to the totally fixed and idiomatic—not see the wood for the trees. …Between these two extremes, there is a whole range of nouns that take the verb see in a way that is neither totally predictable nor totally opaque as to meaning. These run from the fairly weak collocation see a film through the medium strength see a doctor to the stronger collocations of see danger. All these combinations, apart from those at the very extremes of the cline, can be called collocation. And it is combinations such as these—particularly in the medium strength area—that are vital to communicative competence in English.*

Noam Chomsky, quoted in *The BBI Combinatory Dictionary of English* (1986: IX), points out that

decide on a boat, meaning choose to buy a boat contains the collocation decide on (in his terminology: close construction), whereas decide on a boat, meaning make a decision while on a boat is a free combination (in his terminology: loose association). It is believed that "free combinations should generally not be included in dictionaries", but "the inclusion of free combinations is sometimes essential to illustrate a sense of a polysemous entry in a general-purpose dictionary." (ibid., IX)

Contextualization

Biber et al. (2000: 1) state that "studies of language can be divided into two main areas: studies of structure and studies of use." The same authors mention that linguistic analyses have traditionally put more emphasis on structure and described the combinations of smaller units forming larger grammatical units. The second perspective is to emphasize language use and investigate the occurrence of similar structures in different contexts. Biber et al. (2000) suggest that the questions to be answered are whether these structures are used preferentially for different specialized meanings and whether there are different preferences for one of the forms over others.

Contextualization may be called on for help when decoding the meanings of terms. Context is, according to Cann (1993: 22), "needed to restore ellipses, resolve ambiguity, provide referents for deictic elements and resolve anaphoric dependencies." At the same time it should be borne in mind that contextualization is just one approach to meaning and Frawley (1992) considers five approaches to meaning, i.e. meaning as reference, as logical form, as context and use, as culture, and as conceptual structure.

Hauser mentions that "závažnost kontextu pro zpřesnění slova vystupuje do popředí u slov mnohovýznamových a u homonym" a "frekvence nejčastějších kontextů, v nichž se uživatel se slovem setkává, ovlivňuje i chápání významu slova." ["he importance of context for specifying a word comes to the fore in the cases of polysemous words and homonyms," and "frequency of the most common contexts in which the user encounters a word, affects also the understanding of the word's meaning."] (Hauser 1980: 66–67; translated by authors)

Filipec claims that

Vztah významu a oblasti užití lexikální jednotky je sice velmi důležitý a vzájemný – vždyť význam je v kontextu konkretizován a dotvářen, ba i přetvářen, a na druhé straně se v kontextu přímo konstituuje (srov. např. otázku přenášení významu), principiálně jde však o dva různé aspekty lexikální jednotky [the relationship between meaning and the area of use of lexical unit is very important and mutual, because the meaning is being specified and completed, even

> *changed and directly formed in context (cf. for example the issue of transfer of meaning), but in principle there are two different aspects of a lexical unit.] (Filipec 1973: 78; translated by authors)*

The question is what the range of context should be in order to have the lexical meaning of a word specified satisfactorily. According to Těšitelová (1992), the collocational span of one and two words left and right of the collocate are considered to be a minimum for such a lexical meaning specification. Verbs require wider context, though. Cvrček categorizes the following three types of context: "bezprostřední, blízký a vzdálený" [immediate, close, and distant contexts.] (Cvrček 2014: 25; translated by authors) The distant context in such a classification has four and more words left and right of the key word.

According to Cruse (1986), every aspect of the meaning of a word is reflected in a characteristic pattern of semantic normality (and abnormality) in grammatically appropriate contexts. Every difference in the semantic normality profile between two items betokens a difference of meaning. Cruse (1986: 16) notes that

> *the meaning of a word can be pictured as a pattern of affinities and disaffinities with all the other words in the language with which it is capable of contrasting semantic relations in grammatical contexts. Affinities are of two kinds, syntagmatic and paradigmatic. A syntagmatic affinity is established by a capacity for normal association in an utterance (e.g. between dog and barked). A syntagmatic disaffinity is revealed by a syntagmatic abnormality …(e.g. the lions are chirruping).*

He also claims that, "paradigmatically, a semantic affinity between two grammatically identical words is the greater the more congruent their patterns of syntagmatic normality. So, for instance, dog and cat share far more normal and abnormal contexts than, say, dog and lamp-post." (ibid., 16)

Misunderstandings in the area of terminology as well as in general English are caused by the fact that we do not know the collocations which are sometimes fixed and cannot be replaced by alternatives. We quite often do not know that a particular collocation we use is a term with its definition and a specific area of use either. *Contingency plan* is a fixed expression and there are no words which may be used instead of *contingency*. The NATO Terminology Management System (NTMS) includes *disaster recovery plan* as an admitted synonym to *contingency plan*. The term *contingency plan* may also be replaced by the collocation *back-up plan* under less formal circumstances, but the term *alternative plan* is not an appropriate "alternative". To *secure the safety* is an example of "Czenglish" from a sentence in which the expression *to provide security* (or *safety*, depending on the context) should have been written by a student. Students should look at context for help in their effort to comprehend a text or a part of speech without consulting

their dictionaries. When being asked to explain the meaning of a word e.g. *re-ject*, teachers should use the word either in a clear collocation or in a particular context in their effort to introduce new vocabulary without dictionaries. The statement "*the proposal will be either adopted or rejected*" is focused on antonymy and also on the fact that the word *adopt* is considered to be a true friend, not a false friend, having similar meaning both in the English and Czech languages. We believe that focus on collocability and contextualization may reduce such mistakes to a minimum.

The term *management* in the British National Corpus

The term *management* has been analysed in collocations on the position of determinatum (Dm = a component being determined) and determinant (Dt = a component determining another component). The term is displayed from the British National Corpus (BNC) with the help of the *Word Sketch* programme. Only the first 10 occurrences of the term are recorded in Table 1.

Tab. 1: *Management (noun) in BNC*
      freq = 21,886 (195.09 per million)

modifier			modifies		
	10,908	0.50		8,888	0.41
senior	262	8.73	system	727	7.36
system	213	5.56	team	448	8.12
network	188	7.68	committee	278	7.37
quality	179	7.12	group	207	5.76
project	162	6.85	information	199	6.49
top	158	7.92	structure	182	6.98
financial	157	7.40	skill	161	7.28
local	157	6.43	development	149	6.01
database	154	8.08	training	147	6.90
resource	153	7.16	service	147	5.35

The first column shows the word frequency and the second column represents logDice score, which indicates the amount of statistical association between two words. Rychlý (2008: 6) states that "the word pairs with the highest score are presented as collocation candidates." He adds that

> values of the logDice have the following features: a) theoretical maximum is 14, in case where all occurrences of X co-occur with Y and all occurrences of Y co-occur with X. Usually the value is less than 10; b) value 0 means there is less than 1 co-occurrence of XY per 16,000 X or 16,000 Y. We can say that negative values means there is no statistical significance of XY collocation. (ibid., 9)

There are only 10 out of 727 occurrences of the collocation *management system* in a distant context selected as an example and recorded in Table 2. It is beyond the scope of this paper to list all 21,886 occurrences of the term *management* in all its collocations, but it is obvious that the same procedure may be applied for all the selected words and their collocations recorded in the corpus. As can be seen in Table 3, the Sketch engine can even provide us with the whole paragraph in which the analysed collocation occurs.

Tab. 2: *The collocation management system in a distant context*

Environment Ministers, who are establishing a	**management**	*system* to control pollution. Ministers
discharged is the introduction of" energy	**management**	*systems* " microprocessor-based products
by the British Government stand" energy	**management**	*systems* ", microprocessor-based products
Government's current approval of" energy	**management**	*systems* " and Jeremy Bentham's proposals
perceived need to save energy per se. Energy	**management**	*systems* are collections of equipment constructed
arrangement of the Panopticon. *</p>* Energy	**management**	*systems* : beyond Panopticon? *<p>* The basis
has endorsed the introduction of energy	**management**	*systems* because, inter alia, they enable
concerned with rights and liberties, energy	**management**	*systems* are designed to protect interests
And, in order to accomplish this, energy	**management**	*systems* , like the internal arrangement
exercise of power. For the design of energy	**management**	*systems* offers a new diagram of a mechanism

Tab. 3: *The collocation management system in the whole paragraph*

< previous from a ship travelling the North Sea. The NERC survey is the Department of the Environment's contribution to the North Sea Task Force, a body established by nine North Sea Environment Ministers, who are establishing a *management* system to control pollution. Ministers have agreed to reduce emissions of 37 pollutants by 50 per cent by 1995, cutting a handful of more dangerous ones, such as cadmium and lead, by 70 per cent. *</p>* next >

Conclusion

Frequency analysis of terms and their collocations in various contexts as part of linguistic analysis may help experts and students to use the terms consistently and, in the case of no terms, be aware of prevailing tendencies of word combinations from their area of interest. The analysis of frequent collocations being recorded in a corpus may serve as empirical evidence of possible misconceptions about language use. The lessons learned in this area may also lead to an improved teaching process.

References

The BBI Combinatory Dictionary of English: A Guide to Word Combinations. Amsterdam, Philadelphia: John Benjamins Publishing Company, 1986. ISBN 9789027232618

BIBER, D., CONRAD, S., & REPPEN, R. (2000). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: CUP. ISBN 0521499577

CANN, R. (1993). *Formal Semantics.* Cambridge: CUP. ISBN 0521376106

CRUSE, D. A. (1986). *Lexical Semantics.* Cambridge: CUP. ISBN 0521276438.

CVRČEK, V. (2014). *Kvantitativní analýza kontextu*. Praha: UK. ISBN 9788074222641.

ČERMÁK, F. (2006). Polysémie a kolokace: případ adjektiva *měkký*. In ČERMÁK, F., & ŠULC, M. (Eds.) *Kolokace* (s. 56–93). Praha: NLN. ISBN 8071068632.

FILIPEC, J. (1973). K otázce sémantického popisu lexikálních jednotek. *Slovo a slovesnost: časopis pro otázky teorie a kultury jazyka, 34*, 78–84. ISSN 00377031.

FRAWLEY, W. (1992). *Linguistic Semantics*. New Jersey: Lawrence Erlbaum Associates. ISBN 9780805810752.

HALLIDAY, M. A. K., TEUBERT, W., YALLOP, C., & ČERMÁKOVÁ, A. (2004). *Lexicology and Corpus Linguistics.* London, New York: Continuum. ISBN 0826448623.

HAUSER, P. (1980). *Nauka o slovní zásobě.* Praha: SPN.

HNÁTKOVÁ, M. (2006). Typy a povaha komponentů neslovných frazémů z hlediska lexikálního obsazení. In ČERMÁK, F., & ŠULC, M. (Eds.) *Kolokace* (142–167). Praha: NLN. ISBN 8071068632.

JACKSON, H., & AMVELA, E. Z. (2004). *Words, Meaning and Vocabulary.* London, New York: Continuum. ISBN 0826460968

KŘEN, M. (2006). *Kolokační míry a čeština: srovnání na datech ČNK.* In ČERMÁK, F., & ŠULC, M. (Eds.), *Kolokace* (s. 223–248) Praha: NLM. ISBN 8071068632.

LIPKA, L. (1972). *Semantic Structure and Word Formation*. Munchen: Wilhelm Fink Verlag.

LIPKA, L. (1992). *An Outline of English Lexicology.* Tubengen: Max Niemeyer Verlag. ISBN 9783484410039.

Oxford Collocations Dictionary for Students of English (2002). Oxford: OUP. ISBN 0194312437.

PARTINGTON, A., DUGUID, A., & TAYLOR, CH. (2013). *Patterns and Meanings in Discourse*. Amsterdam, Philadelphia: John Benjamins Publishing Company. ISBN 9789027203601.

RYCHLÝ, P. (2008). *A lexicographer-Friendly Association Score* [online]. [Retrieved 10 July 2016] http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf

TĚŠITELOVÁ, M. (1992). *Quantitative Linguistics*. Praha: ČAV. ISBN 802000131X.

Authors

M.A. Jiří Dvořák, Ph.D., e-mail: jiri.dvorak@unob.cz, University of Defence Language Centre
Author is an ESP lecturer, who participated in projects focused e.g. on military activities environmental impact assessment and risk assessment. He is also interested in crisis management terminology.

M.A. Daniela Dvořáková, e-mail: 103675@mail.muni.cz, Masaryk University Language Centre
Author is an ESP lecturer involved in developing materials for the courses in the M.A. programmes. She is also interested in the assessment of students' performance and testing.