Lexical Semantics: Collocation and Idioms

Collocational Relations

"**A collocational relation** holds between two lexemes L1 and L2 if the choice of L1 for the expression of a given meaning is contingent on L2, to which this meaning is applied. Thus, between the following pairs of lexical units collocational relations hold: *to do*: *a favour*, *to make*: *a mistake*, *close*: *shave*, *narrow*: *escape*, *at*: *a university*, *in*: *a hospital*."

A collocational relation holds between components of non-free word combinations, i.e. such combinations whose semantics is not fully compositional and has to be partially or entirely derived from the phrase as a whole. Non-free combinations are opposed to free word combinations where syntagmatic relations hold between words in a phrase with purely compositional semantics. Examples of free word combinations are: *a black cable*, *a different number*, *to put a chair* [*in the corner*], *to write a story*, *to run quickly*, *to decide to do something*. Examples of non-free word combinations are: *a black box*, *as different again*, *to put into practice*, *to write home about*, *to run the risk of* [*being fired*], *to decide on* [*a hat*]. The distinction between free and non-free combinations is a general distinction usually made in linguistic research with respect to syntagmatic relations.

Collocational relations can be classified according to lexical, structural and semantic criteria. The most fine-grained taxonomy of collocations based on semantic and structural principle was given by Mel'čuk (1996). This taxonomy uses the concept of lexical functions.

Collocation: definitions

• Collocations of a given word are statements of the habitual or customary places of that word (**Firth, 1957**). As to the **lexical criterion,** a word is used in a fixed position with respect to another element of collocation and the **statistical criterion** has to do with the frequency of word co-occurrence. Firth was the first to introduce the term 'collocation' from Latin *collocatio* which means 'bringing together, grouping'. He believes that speakers make 'typical' common lexical choices in collocational combinations. Collocation is a concept in Firth's theory of meaning: "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*."

• Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a*, *b*, *c* ... (**Halliday, 1961**). **Lexical criterion** (a word is used in a fixed position with respect to another element of collocation) and **statistical criterion** (high cooccurrence frequency). If a lexical item is used in the text, then its collocate has the highest probability of occurrence at some distance from the lexical item. Collocations cut across

grammar boundaries: e.g., the phrases *he argued strongly* and *the strength of his argument* are grammatical transformations of the initial collocation *strong argument*.

• Collocations are binary word-combinations; they consist of words with limited combinatorial capacity, they are semi-finished products of language, fine combinations of striking habitualness. In a collocation one partner determines, another is determined. In other words, collocations have **a basis** and a co-occurring **collocate** (**Hausmann, 1984**). **Lexical criterion: (**the lexical choice of the collocate depends on the basis). All word combinations are classified into two basic groups, i.e., fixed and nonfixed combinations, with further subdivisions, and in this classification, collocations belong to the category of non-fixed combinations. Internal structure of collocation is emphasized, i.e., collocation components have functions of a basis and a collocate, and the basis (not the speaker) 'decides' what the collocate will be.

• Collocation is a group of words that occurs repeatedly, i.e., recurs, in a language. Recurrent phrases can be divided into grammatical collocations and lexical collocations. **Grammatical collocations** consist of a dominant element and a preposition or a grammatical construction: *fond of*, (*we reached*) *an agreement that...* **Lexical collocations** do not have a dominant word; their components are "equal": *to come to an agreement*, *affect deeply*, *weak tea* (**Benson *et al*., 1986**). **Functional criterion** dictates that collocations are classified according to functions of collocational elements and **statistical criterion** includes a high co-occurrence frequency). This understanding of collocation is broad, and collocations are classified according to their compositional structure.

• Collocations should be defined not just as 'recurrent word combinations', [but as] 'ARBITRARY recurrent word combinations' (**Benson, 1990**). **Lexical and statistical criteria** include arbitrariness and recurrency respectively. 'Arbitrary' as opposed to 'regular' means that collocations are not predictable and cannot be translated word by word.

• Collocation is "that linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its 'synonyms' because of constraints which are not on the level of syntax or conceptual meaning but on that of usage" (**Van Roey, 1990**). **Statistical criterion** proves high co-occurrence frequency in corpora. Van Roey summarizes the statistical view stated by Halliday in terms of expression or 'usage'. A collocate can thus simply be seen as any word which co-occurs within an arbitrary determined distance or *span* of a central word or *node* at the frequency level at which the researcher can say that the co-occurrence is not accidental. This approach is also textual in that it relies solely on the ability of the computer program to analyze large amounts of computer readable texts.

• Collocations are associations of two or more lexemes (or roots) recognized in and defined by their occurrence in a specific range of grammatical constructions

(**Cowie, 1994**). **Structural criterion** shows that collocations are distinguished by patterns. Collocations are classified into types according to their grammatical patterns.

• Collocations are composite units which are placed in a Howarth's lexical continuum model on a sliding scale of meaning and form from relatively unrestricted (collocations) to highly fixed (idioms). Restricted colloca-tions are fully institutionalised phrases, memorized as wholes and used as conventional form-meaning pairings (**Howarth, 1996**). **Syntactic criterion** has to do with commutability: the extent to which the elements in the expression can be replaced or moved (*to make*/*reach*/*take decision* vs. *to shrug one's shoulders*). **Semantic criterion** (motivation: the extent to which the semantic origin of the expression is identifiable, e.g., *to move the goalposts* = to change conditions for success **vs**. *to shoot the breeze* = to chatter, which is an opaque idiom). Classification includes four types of expressions with no reference to frequency of occurrence:

- free collocations (*to blow a trumpet* = to play a trumpet),
- restrictive collocations (*to blow a fuse* = to destroy a fuse/to get angry),
- figurative idioms (*to blow your own trumpet* = to sell oneself excessively),
- pure idioms (*to blow the gaff* = to reveal a concealed truth).

The problem with this classification is that it is difficult to determine what is meant by 'syntactically fixed', 'unmotivated' or 'opaque'. This is seen in the previous ambiguous example of *to blow a fuse*.

• Collocation is the co-occurrence of two items in a text within a specified environment. Significant collocation is a regular collocation between two items, such that they co-occur more often than their respective frequencies. Casual collocations are "non-significant" collocations. (**Sinclair *et al.*, 2004**). **Lexical criterion** dictates recurrency of co-occurrence and **statistical criterion** (high co-occurrence frequency). The degree of significance for an association between items is determined by such statistic tests as Fischer's Exact Test or Poisson Test.

• Collocation is a combination of two lexical items in which the semantics of one of the lexical items (the base) is autonomous from the combination it appears in, and where the other lexical item (the collocate) adds semantic features to the semantics of the base (**Mel'čuk, 1998**). Gledhill (2000) explains that for Mel'čuk, a collocation is a **semantic function** operating between two or more words in which one of the words keeps its 'normal' meaning. As to the **Semantic criterion**. the meaning of a collocation is not inferred from the meaning of the base combined with the meaning of the collocate. Semantics of a collocation is not just the sum of the meaning of the base and the meaning of the collocate, but rather the meaning of the base plus some additional meaning that is included in the meaning of the base. According to Fontenelle (1994) 'the concept of collocation is independent of grammatical categories: the relationship

which holds between the verb *argue* and the adverb *strongly* is the same as that holding between the noun *argument* and the adjective *strong*'.

Idioms

An **idiom** is a complex, multiword expression whose meaning is **non-COMPOSITIONAL**, that is, not predictable from the meanings of the constituent parts. For example, one cannot work out that *spill the beans* means 'reveal the information' or *cut the mustard* means 'meet an expected standard' just on the basis of knowing the meanings of each of the individual words in the expressions and the rules of English grammar. Instead, one has to learn the expressions as whole units and store them in the lexicon as **LEXEMES**.

Because idioms are fixed expressions, the idiomatic meaning is typically not preserved if any of the component words are replaced with a (near) **SYNONYM**, as in *spill the **pulses***. The grammatical form of an idiom is also usually restricted. For example, *Peter kicked the bucket* cannot be put into passive **VOICE** while still retaining the idiomatic meaning: *The bucket was kicked by Peter* does not mean 'Peter died'. Some idioms are **METAPHORICALLY** motivated – for example, *let off steam* 'release pent-up emotions' can be seen as involving a metaphorical conceptualization of a person as a pressurized steam cooker.

Idioms are exceptions. An expression is an **idiom** if its meaning is not compositional, that is to say it cannot be worked out from knowledge of the meanings of its parts and the way they have been put together. *Come a cropper* means 'fall heavily' but we cannot derive this meaning from the meanings of *come*, *a*, *crop* and *-er*. *Browned off* (meaning 'disgruntled'), and *see eye to eye* (meaning 'agree') are other examples. Idioms simply have to be learned as wholes.

Ordinary one-morpheme words are also, in a sense, idioms. The best we can hope to do for the word *pouch* is to pair it with its meaning, 'small bag'. The meaning of *pouch* cannot be worked out compositionally from the meaning of *ouch* and a supposed meaning of *p*.

A good starting place is Makkai's *Idiom Structure in English* (1972), a book that both in thoroughness and explicit statement makes clear what is often only implied in semantic analyses, regarding idioms and otherwise. Consider first two relatively minor, but revealing, points. Makkai cites an English sentence *Kim drives at sixty miles an hour* and comments that in French, German and Russian the preposition would correspond to English *with* rather than *at;* he then reaches the puzzling conclusion that by using *at* here English speakers "conform to an **a-logical** construction whose existence is justified by a majority of speakers" (p. 57). Such a direct comparison of isolated sentences from several languages, while a common practice, mocks the idea that languages are systematic. Makkai neither gives nor appeals to a complete treatment of *at* and *with*. The claim: a language is a-logical if it is out of step, in some immediately evident way, with other languages.

Elsewhere, Makkai discusses certain expressions with apparently empty *take*, such as *take a train, take a bath*, and *take a hint*. He observes: "Even *cursory investigation* reveals that they fall into neatly classifiable categories. . . ." (p. 56, emphasis mine)· Although he later considers his classification tentative, it is surprising for someone to announce conclusions that were the result of cursory investigation. He could do this only if he believed that semantic facts can be directly taken from selected data, and expected his readers to believe this too. Yet it is quite possible that additional data would show that his categories overparticularize a larger category.

Makkai's objective is to provide Stratificational Grammar [SG] with a large body of interpreted data, since it is partly the lack of such that put SG at a disadvantage with TG. Idioms are a strategic choice of subject, because they are the best examples of discrepancy of form and meaning, and thus can exploit fully the SG distinction between 'morpheme' and 'lexeme,' the former of which is only a form and not directly related to meaning. Makkai is willing to grant *kick* in *kick the bucket* a morphemic status, since it has the usual morphological variations of *kick (kicks, kicked, kicking),* but he denies it lexemic status: it seems to contribute no meaning to the phrase.

Makkai distinguishes between ENCODING (the *drive at* example earlier) and DECODING idioms, the latter both LEXEMIC *(kick the bucket, hot dog, red herring)* and SEMEMIC (proverbs such as *Don't count your chickens before they're hatched)*. Decoding idioms create DISINFORMATION: interpreters are misled if they try to compute the parts. This criterion is determined by his definition of IDIOM, which (SG technical language aside 1) is simply: *an expression whose full structured meaning is not equal to the sum of its parts*. Ironically, this implies that "free syntax", supposedly the overstated mistake of TG, is the sole linguistic norm; it also assumes that compositional computation never, or only very superficially, involves pragmatic factors.

Makkai correctly notes the inadequacy of some attributes that are often considered criterial for idioms. Regarding frozen or formulaic forms, he cites a number that are not idiomatic: *assets and liabilities, man and wife*, *each and every, facts and figures*, and others (p. 316). Also, while idioms are often figurative, a figurative expression is not necessarily an idiom. Makkai notes: "*Go down* in the sense 'sink, perish' as said of ships is . . . not an idiom, because a simple metaphorical extension of each constituent lexon will easily suggest the meaning" (p. 142). (I don't consider either word figurative, but I accept the general point.) By stipulating that each word of an idiom must occur elsewhere with a meaning, he also eliminates constructions involving uniquely occurring words, as *kith* in *kith and kin*, because they do not create disinformation; he calls them PSEUDO-IDIOMS (p. 123).

Makkai's treatment is extensive and carefully reasoned; but his overcommitment to compositionality creates the idioms he describes. The problem (as I see it) is not the SG model he assumes; though he considers his book a confirmation of

SG and refutation of TG, I find the claim irrelevant. Neither SG nor TG rises or falls because of his arguments; his intuitive judgments of idiomaticity could also be those of someone in CTG. Nor is the problem primarily due to Makkai's strange assumption, in his standard of Disinformation, that pedagogical and theoretical grammars can be the same. Obviously, a foreigner has trouble to the extent that language-in-the-whole (with pragmatic modulations) is not a self-contained entity with immediately evident systematicity. The foreigner's major handicap is, in fact, no different from the native's, or the intuitional linguist's: the stereotypic conscious mind. Proceeding from oversimplified, overconscious expectations, Makkai implies in his definition of compositionality that non-idiomatic expressions *should* exhibit unfettered generativity and that contributing contextual effects *should* be irrelevant.
Nor does the problem concern the word *idiom* as such. Obviously, it can be (and is) used for various different kinds of individuation: a speaker, a dialect, a style, a language, etc. Admittedly, all of Makkai's phrases present some degree of idiosyncrasy. My question is simply: at what level of abstraction does a phrase individuate? By giving constituent words morphemic, but not lexemic, status, Makkai judges the individuation to be relatively abstract, intralinguistic. For some phrases, this may be correct. My claim is, however, that for (most if not all) "phrasal verbs", the individuation is pragmatic, extralinguistic; thus, in the SG framework, constituent words should have both lexemic and sememic status.

Assuming then that "idiom" should apply (for present purposes) only to phrases that are intralinguistically idiosyncratic, we need this definition: *an idiom is an expression whose words occur elsewhere but never with the same (inherent) meaning as in this expression*. This definition allows what Makkai assumes must be denied: the possibility that constituent words may contribute semantically to an expression, yet not account fully for its perceived meaning (a circumstance we noted often with *hit*). Not only words, but also combinations of words, are open to modulation of meaning. By adopting a less strict definition. we can avoid Makkai's handicap: the belief that compositionality must be accounted for totally by lexical means.
Nothing is gained by the usual tactic of treating as idiomatic every phrase that is strange to conscious intuition. As a consequence of my definition, idiomaticity cannot be directly, immediately and obviously judged; rather, it should be concluded only after an exhaustive, and finally futile, investigation that finds no linguistic unities. The defects of Makkai's view of compositionality are revealed by Makkai himself; he is his own inadvertent and conscientious critic. At the end of his book he organizes idioms by frequent meanings they exhibit, and notes (p. 308) that four idioms with *up* (*build up* 'increase', *build up* 'exaggerate', *lay up* 'accumulate', and *mark up* 'increase prices') share a sense of 'increase', a sense that *up* exhibits in other expressions also. Makkai must conclude that "this is not sufficient reason to regard *up* as literal here and disqualify the idioms *qua*

idioms, because the total paraphrases remain nondeducible from the constituent parts" (p. 310).

Because of misconceived (and thus misapplied) compositionality, it must be denied that the *up* here is the same lexeme that occurs elsewhere, though its semantic role is evident even to the linguist who denies it. Also, a number of idioms with the meaning of 'decrease' and 'diminish' contain the word *down* (p. 308: *die down* 'decrease', *cut someone down* 'deflate ego', *mark down* 'decrease price of', *play down* 'deprecate', *talk down* 'minimize importance of'), yet these are unrelated to each other or to other uses of *down*. Seven different idioms with *get* share a sense of 'success, attainment' (p. 308: *get along* 'succeed', *get along with* 'have successful relation', *get by* 'barely succeed', *get away with* 'succeed in perpetuating illegal act or mischief without punishment or repercussions', *get something over with* 'render accomplished', *get on* 'succeed', *get in with/on* 'succeed in obtaining desirable position or association'); a number of idioms referring to speech have words such as *speak, tell, talk* and *answer* (p. 307: *speak up* 'speak louder', *tell someone off* 'blast, tell honest opinion in anger', *talk over* 'discuss', *answer back* 'retort disrespectfully', *talk back* 'reply sassily'); and yet, because compositionality is lacking, none of these words can be related to other uses. Although Makkai's appendices diligently show these semantic links, there is no place in his theory to accommodate them. Such conclusions should indicate that something is wrong with his definition of *idiom*.

A startling aspect of Makkai's analysis is that he ignores (solely on the strength of principle) even the most transparent relationships. Equally startling is the total faith he has in his glosses; they sometimes create the differences he finds. The examples just quoted give evidence. *Play down* and *talk down* are closely related, yet the glosses 'deprecate' and 'minimize importance of' needlessly blur the relationship. *Get along* and *get along with* differ in parallel to *succeed* and *succeed with*, yet Makkai obscures the use of *with* by glossing the latter as 'have successful relation'. *Answer back* and *talk back* are obviously linked, but the glosses 'retort disrespectfully' and 'reply sassily' seem to imply greater distance.

These examples also show the critical role of glosses in overconscious treatments. Although glosses are presented as evidence, faithfully and accurately representing meaning that form obscures, they are rather akin to propaganda, serving external judgments that have been made in advance. We have a typical process of conscious distortion. First, an expression appears puzzling to the conscious mind. Then, instead of researchers admitting they are puzzled, and thus suspending judgment while they gather a wider range of data, they rush to a conclusion, based on paraphrase and compositionality. The conclusion is disguised because it is formulated as a gloss. The researchers then proceed to analyze, not the expression, but the gloss. There are no established guidelines for glosses, and so they can be slanted or subtly rephrased to support any prior theoretical claim.

While it is assumed that the glossed expression is misleading, the gloss is taken as accurate, at least to the degree that it makes no difference in the analysis. The researchers then draw the conclusions that are inherent in the gloss. Whatever results is an irrelevancy, because the original data have been eliminated from the proceedings. English is not Makkai's first language, so he relied on informants rather than his own intuitions; but informant judgments are also intuitions, and suffer from the selective awareness of consciousness, which has little insight into paradigms. A number of other idioms Makkai cites also have closely related expressions. He gives *look back on* 'reminisce about' (p. 222); there is also *think back on.* He gives *fly in* 'arrive by airplane' (p. 230); there is also *fly out, fly by, fly over*, as well as *come in, drive in, motor in, ski in, jet in*, and others. He gives *come again* 'repeat what you said' (p. 216); there are also *send that by me again, run that by me again, let me have that again, give me that again, give that to me again,* and *put that to me again*, data that not only establish the non-idiomaticity of *again,* but provide evidence for possible relationships between *come, send, run, have, give* and *put*. These expressions do not allow immediate, obvious conclusions, for they are all problematic; but they are data, not glosses,
and they demonstrate sufficiently well that quick judgments of idiomaticity are highly suspect.