Lars Trap-Jensen, The Danish Dictionary

Word Relations: Two Kinds of Typicality and their Place in the Dictionary

Abstract

The paper focuses on some aspects of semantic-lexical collocations, viz. word relations given in the dictionary as illustrative, typical chunks of real language. They appear as a supplement to the definition, thereby facilitating an accurate understanding of the entry word, and with no further explanation required as they are semantically transparent. I point out that the notion of typicality can be conceived of in at least two different ways: as an empirical measure of the degree of mutual attraction between two or more words, or as a psychological fact about cognitive salience in the speaker's mind. As the two conceptions may have only little in common, or even be inversely related, there is all the more reason to consider their implications thoroughly. Finally, I discuss which of the two conceptions should be given priority in the dictionary.

1. Introduction

For a given entry word in the dictionary, one may wish to supplement the definition with a list of non-fixed word combinations serving as examples of illustrative, idiomatic and/or culturally or pragmatically important language usage. For example, the definition of one meaning of *house* ('building designed for people to live in') may be followed by word combinations such as: *build a house, renovate a house, let a house, one-family house, detached house, summer house* to suggest a few possible candidates (extracted from The BBI Combinatory Dictionary of English). It is word relations of this kind that I want to discuss here. They may be distinguished from, on one hand, gram-matical collocations whose combinatorial constraints originate in grammatical selectional characteristics, e.g. phrasal verbs, and on the other hand, fixed expressions of meaning or idioms where the semantic dependencies among the constituting elements are non-transparent, and the expressions thus only understandable as lexical wholes that require their own definition. In other words, I am concerned here with word relations which are neither inherently part of the grammar nor the lexicon of the language. They are combinations of words that, as it were, happen to be more telling than others, perhaps realized more often, and they are often experienced as typical by native speakers.

Modern computerbased dictionary-making has made it possible to detect such word relations much more precisely than hitherto, and there are sound arguments for presenting the findings in a dictionary. There are, however, also good reasons for caution, and in the next sections I shall try comparing and evaluating computerbased statistical findings with traditional introspective methods. The account springs from the considerations underlying the format of The Danish Dictionary and therefore applies principally to the comprehensive monolingual dictionary.

2. Types of word relations

As a class, the word relations in question are more or less open-ended and have no clear-cut borderline towards collocations determined by either grammatical selectional properties (valency and constructional information) or idioms, idioterms and proverbs. In practice, they are distinguished from collocations requiring their own definition by the criterion of semantic transparency, and from grammatical collocations by their lower degree of selectional constraint.

Svensén (1993:101) provides a list of the most common types of word relations. He includes: possible adjectival adjuncts for a given noun, possible subject nouns for a given verb, possible adverbial adjuncts for a given adjective, possible verbs governing an object noun, and possible object nouns for a given verb. The list could be continued: possible adjectives functioning as heads of an adjunct adverb, possible prepositions modifying a governor noun in a prepositional phrase, possible adjunct nouns modifying another noun in a genitive construction, and so on.

Information on word relations may serve more than a single purpose, and often these purposes will mutually reinforce the reason for bringing it, but as they may also be competing, there are good reasons for being explicit:

a) for the specialist or the lay person with a specialized interest in language, it is in itself valuable to be able to seek information on word relations that occur with a significantly higher relative frequency than other possible relations. The dictionary containing this type of information imparts, so to speak, a state-of-the-art picture of actual language usage at a particular point in time. Needless to say, the information needed to fulfil this function can only be made accessible by the aid of a relatively large language

corpus in combination with software programs capable of making fairly sophisticated statistical analyses.

b) word relations function complementary to the definition given for the entry word by placing the entry word in, admittedly, minimal cultural and linguistic contexts. This is valuable for both the productive and the receptive use of the dictionary.

c) since word relations often have the structure of a phrase or a constituent, it will serve as a place to look for the productively oriented user, e.g. the advanced L2 learner who wants to improve her command of idiomatic expressions that are neither constructionally nor lexically constrained.[1] Inevitably, word relations will also serve as a model which can be used productively by analogy whether one, in theory, likes it or not.

3. Different types of typicality

As mentioned above, word relations represent small chunks of language that are somehow felt to be typical of the usage of a particular word. But as it is not at all clear what exactly is to be understood by 'typical', it is worthwhile discussing the notion in greater detail.

One possible way of defining a typical word relation is in terms of *frequency*. For example, for a given noun one could argue that the most frequent adjective occurring immediately to its left would be the most typical modifier of that noun. Obviously, this is a very poor definition of typicality as it does not take into account the fact that some adjectives are in themselves much more frequent than others and therefore tend to occur correspondingly more often with almost any noun. It is, however, quite possible to allow for the absolute frequencies of the respective collocates in statistical analysis, thereby obtaining a figure that reflects actually realized occurrences in relation to possible occurrences. This measure may be interpreted as expressing the degree to which two words mutually attract each other. The measure can be used by the computational lexicographer as an operational definition of typicality, and as such it is known as the *mutual information index*.[2]

On the other hand, there is an altogether different conception of typicality which has been widely used in cognitive psychology and which in recent years has been adopted into linguistics by cognitive semanticists. According to this view, typicality should be conceived of as equivalent to *prototypicality* in the psychological sense. Prototypical meaning represents the meaning first learnt by children, and it corresponds to the answers given by people when asked to give a good

example of the meaning of a particular word. Naturally, this conception is also more in accordance with people's own introspective judgments about typicality.

Many people would perhaps expect that the two conceptions of typicality are more or less congruent so that the word relations resulting from statistical analysis would correspond roughly to our own judgment as native speakers about typical relations. For these people it may be surprising to learn that analysis of actual language data clearly shows that this is not at all the case. Indeed, if anything, the reverse is true: "The majority of examples are found in the skirt and at the periphery", but not in the core of a fuzzy set or meaning continuum (Coates 1983:13).

At this point, it is appropriate to be somewhat more concrete and illustrate with a few examples. For the sake of space and clarity, I shall confine myself to only one type of word relations, viz. adjectives and nouns in adjunct-head relations.

In the tables 1–4 typical word relations for two nouns and two adjectives are listed. The tables summarize the mutual information scores resulting from statistical analyses of the 40 million word corpus of The Danish Dictionary. In table 1, the top 15 adjectives are given in descending order of statistical typicality (with the mutual information score appearing in the left column) for the Danish word *træ* (equivalent to both 'tree' and 'wood' in English) as head word including all inflected forms and with one place to the left of the node word as contextual constraint.[3] The figures in the right column show the number of absolute co-occurrences. In table 2, the same information is given for the word *dyr* ('animal'), whereas tables 3 and 4 list the most typical (by the same definition) nouns occurring immediately to the right of the adjectives *rød* ('red') and *hvid* ('white') as node words.

dry-rotten	4597.74	7
fairest	4105.13	5
compregnated	3694.62	9
newly planted	1768.36	7
leafless	1768.36	7
shady	1728.48	10
evergreen	1642.05	10
deciduous	1420.15	8
centuries old	1172.89	5
dicotyledonous	804.27	12
lacquered	665.13	8
hollow	603.97	8
felled	500.96	9
liquid	443.80	5
naked	300.60	40

Table 1
mutu: *træ* (infl), interval
[–1, –1], co-occ >=5

invertebrate	5171.24	27
poikilothermic	2585.62	5
transgenic	1292.81	5
wild ('vildtlevende')	1175.28	5
higher	1170.85	12
stuffed	1149.16	10
monocellular	1108.12	12
full-grown	1077.34	5
grazing	904.97	7
dumb	718.23	5
wild ('vild')	471.00	194
tame	373.18	7
journalistic	239.73	7
intelligent	159.61	5
threatened	125.87	24

Table 2
mutu: *dyr* (infl), interval
[–1, –1],co-occ >=5

Cross packages	4314.02	10
Cross committee	4314.02	5
Khmer	4086.97	64
tiled roofs	2986.63	9
brigades	2876.01	16
giant star	2696.26	5
lantern/light	2426.64	9
blood cells	2396.68	5
amanita (=fly agaric)	1764.83	9
Cabinet	1403.36	27
tights	1232.58	6
flunkeys	1198.34	5
banners	1190.07	32
pepper	1186.36	44
shrimp	995.54	6

Table 3
mutu: *rød* (infl), interval
[+1,+1], co-occ > =5

minority rule	3278.50	15
cotton panties	2723.67	9
blood cells	2377.50	84
oxeye (=oxeye daisy)	1967.10	8
coat	1650.71	60
slave traffic	1416.31	9
ankle socks	1124.06	6
poodle	1049.12	16
tornado	786.84	8
box	786.84	12
sandy beach	786.84	8
South Africans	737.66	6
shirt blouse	731.94	8
latex	715.31	6
sheets	715.31	6

Table 4
mutu: *hvid* (infl), interval
[+1, +1], co-occ >=5

In order to compare these findings with word relations of more psychological salience, I carried out a small, informal test. Four test persons (with no relation to lexicography) were asked to write down the first word relations that spontaneously came to their minds when presented with a number of nouns and adjectives, including the ones

appearing in the tables. For the four words, the test persons suggested relational words such as the following (rendered in English translation, and in random order):

træ ('tree'/'wood'): scented, strong, green, split, sawn-off, overturned, wilted, big, untreated, planed, small
dyr ('animal'): dangerous, small, wild, big, dead, threatened, tamed, wounded, encaged, hunted, soft, hungry, cut-up, furry
rød ('red'): colour, flag, pepper, apple, rag, brick, tiled roof, communist, heart, house, sports car, rose, banner, stoplight
hvid ('white'): snow, teeth, swan, cross, lily, sheet, colour, dove, handkerchief, china, chalk, mice, lime

4. Discussion

If both methods are considered legitimate ways of obtaining candidates for typical word relations, it is striking how little they have in common. Out of 60 possible candidates derived from the corpus analyses, only seven show up on the test persons' lists. This is, of course, a confirmation of the above-mentioned observation that prototypical meaning is rarely statistically prominent. A cursory glance at the two lists also indicates an explanation as to why this might be so. It seems that we can recognize three continuous, broad descriptive levels for adjective-noun relations which at their centre are clearly separable, ranging from very general to very specific characterization with a neutral or basic level between the two. The general level includes content-weak adjectives whose function semantically approaches that of grammatical words, e.g. pronouns (*different, other, similar, various* etc.). The basic level includes common simplex words belonging to the core of the vocabulary (*big, nice, man, dog, chair, round* etc.), whereas the specific level comprises informatively heavy, often compound words with peripheral status in the vocabulary and often of low absolute frequency (*membrane-winged, lanceolate, bipinnatisect, multinuclear* etc.).

It is obvious that the psychologically typical word relations are selected from the basic level of description, whereas the candidates showing up in the statistics tend to belong towards the specific end of the continuum. On further reflection, this also makes sense: firstly, if an adjective is very central in meaning to the noun that it modifies, it becomes an almost inherent property of that noun, and it is thus semantically or pragmatically redundant to use the adjective explicitly. For example, to say of a roof that it is sloping is trivial and non-

informative because roofs usually are (at least in this part of the world). Generally speaking, the shapes of roofs are most likely to be mentioned in natural language when they in fact deviate from the normal situation and therefore become thematized as topics of discussion. In our corpus *fladt tag* ('flat roof') is consequently a statistically more typical relation than *skråt tag* ('sloping roof'). On the whole, arguments along these lines would lead us to appreciate why prototypical word relations are not necessarily very frequent in performance data. Secondly, even if prototypical word relations should consist of frequent and common combinations of words, they tend to be disfavoured by the statistics, as the common collocates will inevitably also combine with a lot of other words. Conversely, rare words which are lexically more bound will show up in the statistics, however small the absolute frequency of occurrence.

These observations have important consequences for the way in which word relations are treated in the dictionary. It is obviously not the case that the one way of describing typicality represents the ultimate truth to the detriment of the other. Both are perfectly legitimate ways of describing what is typical. You could say, however, that they are truths about language on two different levels, but for the lexicographer the crucial question remains: which of the two contains the more valuable information and should therefore be given priority in the dictionary? As is often the case in these matters, the only reasonable answer is: it depends..., and, hardly surprising, it depends on the needs and interests of the user. If you are a non-native speaker seeking information about the meaning of the word *red*, you will probably be more satisfied with the prototypical relations *red apple*, *red rose* or *red flag* than with the statistically more prominent *red tights* or *red giant star*. On the other hand, to the hard-core descriptivist it might be more interesting to learn that the word *animal* is statistically more typical when combined with *invertebrate* and *poikilothermic*, rather than with some relatively non-informative words like *big* or *wild*.

In principle, the two types of information are equally valuable, but belong in different dictionaries aimed at the respective target groups. In practice, the lexicographer must make a choice, and here the most reasonable solution will be to take the position between the two. In a multi-purpose comprehensive dictionary, the most illustrative word relations are to be found in the transitional zone between the basic and specific levels of description. At this level, one receives information on both conceptions of typicality and at the same time gets rid of the too general, non-informative combinations on one hand, and of the too specific or over-informative on the other, i.e. psychological salience and semantic specificity are brought together, to the satisfaction of the

greatest number of users. The practical way to handle the solution involves running the statistical analyses and choosing from the list of candidates generated *not* the words with the highest mutual information score, but selecting among the words of a somewhat lower score those that fit the level of description between basic and specific. A good indicator of this is cooccurrence in absolute numbers. In fact, a relatively high number of absolute cooccurrences in combination with a relatively high score on mutual information are in most cases precisely the characteristic features of the most illustrative word relations.

In conclusion, some reservations ought to be made about the scope of these observations. The discussion has focused on adjective-noun relations exclusively, and it may well be that this type of relation has a peculiar status that cannot readily be generalized to other kinds of word relations. Adjective-noun relations are perhaps better suited than other kinds of relations to illustrate the essence of prototypicality, or at least one central aspect of it, viz. the attribution of one or more characteristics to an entity, either explicitly as a statement with the adjective in predicative position or implicitly by means of an attributive adjective (in which case a predication may be presupposed). For other types of relations, it is quite likely that the general picture is altogether different. Finally, one may speculate if identical results could have been obtained from other corpora of differing size or composition. In spite of its considerable size, the corpus of The Danish Dictionary is still remarkably sensitive to the contents and topics of just a few or perhaps even a single text. Maybe it is the nature of performance data to be unstable and fluctuating, but similar investigations from other projects would indeed provide a valuable basis for comparison. However such investigations turn out, it remains of course true that in each case the selection of appropriate word relations can only be successful when the computer generated statistics are combined with the lexicographer's careful introspection.

Notes

1. For a relevant discussion of the pedagogic implications of authentic vs. introspective examples, see Laufer (1992). In this connection pedagogic value constitutes of course only one (but indeed important) of several aspects to be considered.
2. For a more detailed exposition of the technicalities involved, see e.g. Heid (1994: 247ff.), Church et al. (1991).

3. The list has been cleared of all non-adjectival forms. The words are rendered in English translation, but remain, of course, facts about Danish. Inflectional forms of the collocates have been lumped together.

References

Benson, M. 1989. "The Structure of the Collocational Dictionary", in: *International Journal of Lexicography.* Volume 2, Number 1, pp. 1–15.

Benson, M., Benson, E., and Ilson, R. 1986. *The BBI Combinatory Dictionary of English. A Guide to Word Combinations.* Amsterdam/ Philadelphia, John Benjamins Publishing Company.

Church, K. et al. 1991. "Using Statistics in Lexical Analysis" in Zernik, U. (ed.). *Lexical Acquisition.* Hillsdale, N.J., Lawrence Erlbaum Associates, Publishers.

Coates, J. 1983. *The Semantics of the Modal Auxiliaries.* Beckenham, Croom Helm Ltd.

Heid, U. 1994. "On Ways Words Work Together – Topics in Lexical Combinatorics" in Martin, W. et al. (eds.). *Euralex '94 - Proceedings.*

Laufer, B. 1992. "Corpus-Based versus Lexicographer Examples in Comprehension and Production of New Words" in Tommola, H. et al. (eds.): *Euralex '92 - Proceedings I–II,* pp. 71–77.

Svensén, B. 1991. *Practical Lexicography.* Oxford, Oxford University Press.