Sanni Nimb, The Danish Dictionary, University of Copenhagen

Collocations of Nouns: How to Present Verb-noun Collocations in a Monolingual Dictionary

Abstract

The paper discusses how to present in a monolingual corpus-based dictionary of Danish a certain group of lexical collocations in the noun entry, namely verb-noun collocations that are transparent in meaning. These collocations can be divided into two groups: verb phrases where the noun for inexplicable reasons selects a certain verb, while synonyms are impossible, and verb phrases where the noun is just a typical object. One way of bringing these collocations is presented and certain problems are discussed.

1. Introduction

The Danish Dictionary is a comprehensive monolingual dictionary of modern Danish to be published in 1999. The dictionary is mainly based on a 40 million word corpus. One of the new things we introduce compared to other monolingual Danish dictionaries is information about the ability of words to combine with other words. One way of bringing such information is to present different types of typical collocations. Bringing typical collocations in a monolingual dictionary has two purposes: it is very useful in a situation where production of language is needed, both for native speakers and learners of the language, and it can supplement the semantic definition of the entry word.

The collocations I will concentrate on in this paper are verb-noun collocations that are to be mentioned without further explanation in the lexical entry of the noun. They are all semantically transparent and will therefore not be lexically defined.

With the corpus as the main source we have a very good chance of finding and passing on to the users of the dictionary different kinds of such verbs that typically occur to the left of the noun we want to describe. We have at our disposal special computer tools which measure the mutual attraction between two words in the corpus. The probability of meeting exactly those two words together is calculated, and the result is a list of words ranked by their probability of co-occurring with the entry word. For example, the result of a statistic analysis of the words to the left of the noun *konsekvens* ('consequence') tells us that we usually

265

combine the word with a number of different adjectives and with the verbs *overskue* ('to survey'), *drage* ('to draw'), *vurdere* ('to estimate') and *tage* ('to take)'.

A similar analysis of the noun *selvmord* ('a suicide') shows us that the verbs *begå* ('to commit') and *forsøge* ('to try') are verbs which typically appear to the left of the noun. And for the noun *gavn* ('benefit') a mutual analysis shows that the verbs *gøre* ('to do'), *have* ('to have') and *få* ('to get') are verbs which typically appear to the left of the noun. An example of such a list can be seen in table 1.

a)	(b)	(c)
uoverskuelige (unpredictable)	1242.39	[38]
vidtrækkende (farreaching)	1152.73	[26]
miljømæssige (environmental)	801.40	[43]
yderste (utmost, absolute)	358.91	[72]
overskue (to survey)	337.89	[25]
økologiske (ecological)	309.99	[33]
uheldige (unsuccessful)	286.25	[27]
samfundsmæssige (social)	257.39	[19]
negative (negative)	219.37	[31]
alvorlige (serious)	218.23	[60]
drage (to draw)	167.59	[25]
hvilke (which)	158.66	[113]
vurdere (to estimate)	100.65	[21]
økonomiske (economic)	79.98	[82]
mulige (possible)	68.30	[29]
dens(its)	27.91	[29]
politiske (political)	26.98	[33]
taget (taken)	26.68	[62]
tage (to take)	23.08	[87]

Table 1. A mutual analysis of *konsekvens* ("consequence"), second and first place to the left. Typical word (a), degree of probability (b), and total number of co-occurence(c):

Since The Danish Dictionary is corpus-based, it has been decided that we as a fundamental principle present statistically based collocations by listing them according to the order in which they appear in the statistic analysis, without further classification. For this purpose we have reserved a special element in our dictionary structure.

This principle of presenting collocations that are extracted by statistical analysis from the corpus without further classification by the

lexicographer, can, however, cause a variety of problems especially with regard to the verbs. The statistic analysis will often result in a mixed list of verbs with very different relations to the noun. By simply listing the verbs in the order in which they appear on the statistical list, we run the risk of mixing very different types of information.

Some of the verb-noun collocations will be semantically predictable, for instance if the noun is just a typical object of the verb for semantic reasons (collocations like *to build a house, to cure a disease* etc.). Information in the dictionary on these free lexical combinations mainly serves to underline the semantic definition of the noun. Two examples of this type of collocation from the list in table 1 are the collocations *overskue/vurdere konsekvenserne* ('to survey/to estimate the consequences').

Other verb-noun collocations found by the statistic analysis of the corpus will be more fixed verb-noun collocations that are impossible for non-native speakers to predict and where the noun selects the verb for unexplainable syntactic reasons. These are collocations like *take a drag (on a cigarette), pay attention, deliver a speech* where *take, pay* and *deliver* cannot be replaced by synonyms. Some charateristics of the verbs in these fixed lexical combinations are that they are loosing their concrete semantic meaning and that they contribute very little to the meaning of the phrase, acting almost like an auxiliary verb. Such verbs are called support verbs. Information on support verbs in the dictionary serves to teach mainly non-native speakers how to construct well-formed sentences with the noun. From the list in table 1 we have the following examples of collocations that consist of support verb + noun: *drage/tage konsekvenserne* ('to draw/to take the consequences').

By simply listing the free lexical combinations and the more fixed support-verb constructions in the same order as they appear on the statistic list, we do not make their difference clear to the user.

Another problem caused by the principle of simply listing the verb-noun collocations arises with nouns that simultaneously select a verb to the left and an obligatory prepositional phrase to the right. For example the noun *gavn* ('benefit') cannot stand alone in verb-noun collocations with the support verbs *have* ('to have') and *få* ('to get') without a prepositional phrase beginning with *af* ('from'). Consequently, in order to present the verb-noun collocation *få gavn* ('get benefit'), we would need to mention an incomplete phrase like *få gavn af* ('to get benefit from', which we do not find very satisfactory. In order to complete the phrase we would then hope to find a frequent and typical head of the prepositional phrase in the corpus, but this is rarely possible. Since the statistical element in the dictionary must only contain words which

frequently appear in the corpus, we are left with a presentational problem.

The procedure adopted by The Danish Dictionary in order to solve the two above-mentioned problems will be further elaborated in the next section.

2. How the Danish Dictionary classifies the different types of verb-noun collocations

As mentioned above the default method in the Danish Dictionary is simply to list verb-noun collocations in the same element in the order in which they appear on the statistical list. However, to avoid the mixing of support verbs and "free" verbs in the cases where more than one support verb figures on the list, it has been decided to deviate from this default method by grouping the support verbs irrespective of their statistic order:

> konsekvens sb.
> følge; virkning typisk: *overskue konsekvenserne,*
> *vurdere konsekvenserne, <u>drage/tage konsekvensen</u>*

> ('consequence' n.
> result; effect typical: *survey the consequences, estimate*
> *the consequences, <u>draw/take the consequence</u>*)

Table 2. (example of lexical entry)

Moreover, we have in two cases decided to move the support verbs out of the statistical element and present them in another element in the dictionary, reserved for formalized information on how the noun is construed with other words. Information in this element does not need to be based on a statistic analysis, but is meant to describe more valency-like information on the entry noun, as for instance certain prepositional phrases selected by it (e.g. *a key **to a door***).

For the cases where this element is already being used for this kind of information, we have decided also to place support verbs here. The verb is only mentioned here when the noun often occurs with the verb as well as with the prepositional phrase, though the prepositional phrase does not have to be obligatory.

For the presentation of the noun *konsekvens* ('consequence'), which optionally selects a prepositional phrase to the right: *af NGT/at..* ('of something/gerund..'), this means that we have the possiblity of mentioning the two support verbs, *drage* og *tage* ('to draw' and 'to take'), in

a formalized way instead of describing the noun as seen in the example in table 2. The two verbs for which *konsekvens* is just a typical object, *overskue* and *vurdere* ('to survey' and 'to estimate') will still be listed as good examples of language use in the statistical element.

> konsekvens sb.
>
> følge; virkning [drage/tage konsekvensen af NGT/at..];
> typisk: *overskue konsekvenserne, vurdere*
> *konsekvenserne*
>
> ('consequence n.
>
> result; effect [draw/take the consequence of something/gerund];
> typical: *survey the consequences, estimate the*
> *consequences*')

Table 3. (example of lexical entry)

For nouns selecting obligatory prepositional phrases this method also gives us the possibility of solving the presentational problem mentioned above. Both verb, noun and prepositional phrase are described in a for-malized way in the construction element: *[få gavn af NGT/at..]* ('[to get the benefit from something/gerund..]').

The other case where we place a support verb in the construction element is when a noun can combine only with one support verb. This is often the case with nouns that are not very frequent in the corpus, and where the statistic analysis is not informative.

An example of such a case is the noun *helligbrøde* ('sacriledge'), which appears only 24 times in the corpus and which only selects one support verb *begå* ('to commit'). We will therefore place this verb-noun collocation in the construction element.

For all other nouns (those which do not select a prepositional phrase, but which selects more than one support verb), we only distinguish between the cases of free lexical combinations and the cases where a verb from the statistic list is a support verb by grouping the latter in the statistic element. This means that if only one support verb figures on the statistic list there will be no notable difference between the "free" verb and the support verb. The noun *selvmord* ('a suicide') is an example of this. The statistical result from the mutual analysis tells us that both the support verb *begå* (= 'to commit') and the "free" verb *forsøge* are very frequent immediately to the left, but since the noun has more than one possible support verb (also *gøre*, 'to do', is possible but just not very frequent), and since the noun does not select a valency-like prepositional

269

phrase, we simply list the two typical verbs *begå* and *forsøge* from the statistic analysis as good examples of language use.

3. Problems regarding the presentation of verb-noun collocations

The presentation of verb-noun collocations described above can of course be discussed. It might seem inconsistent that we treat verb-noun collocations differently, depending on the capability of the noun to select prepositional phrases or more than one support verb.

This treatment, however, is mainly due to practical circumstances. As long as the noun does not need to be described in a formalized way, not selecting valency bound prepositional phrases, we prefer not to complicate the lexicographer's analysis more than necessary by simply mentioning the results from the statistical analysis. The Danish Dictionary is mainly corpus-based and not meant to be complete in its information on support verbs – this would need a much more detailed analysis and description of each noun.

The disadvantage is of course that by simply listing different types of verbs in the same element, the user will not know when the collocation mentioned is a model to be strictly followed, or just an example of language use.

Therefore we have chosen to take the step of distinguishing between the verbs in the lexical description of some nouns, because we in these cases already need to introduce a more formalized presentation of the noun in order to describe its valency. In these cases we want to underline that certain verbs also play a role in the construction of sentences with the entry word, in the hope that a more formalized presentation provides more precise guidance to the user.

In the cases where only one support verb is possible we also hope that the user will perceive the formalized pattern as a model to be followed when producing sentences.

References

Benson, Morton 1986. *The BBI combinatory dictionary of English.* Amsterdam, John Benjamins Publishing Company

Heid, Ulrich 1994. "On Ways Words Work Together – Topics in Lecical Combinatorics" in: *Euralex '94, Proceedings*

Lyly, Erika 1993. "Halvfasta fraser – ett lexikografiskt problem" in: *Nordiske Studier i Leksikografi II. Skrift nr.2*, Copenhagen: Nordisk Forening for Leksikografi

Svensén, Bo 1993. *Practical Lexicography – Principles and Methods of Dictionary Making*. Oxford, Oxford University Press

Feil, Ruth 1995. "Funktionsverber i det danske sprog" in: *Nordiske Studier i Leksikografi 3,* Reykjavik: Nordisk Forening for Leksikografi