Henrik Lorentzen, The Danish Dictionary, Copenhagen

Lemmatization of Multi-word Lexical Units: In which Entry?

Abstract

The paper examines and discusses the difficulties involved in lemmatizing[1] multi-word lexical units. The principles used in a large monolingual dictionary are presented, namely the establishment of an order of word classes to be followed when deciding in which entry a particular lexical unit of this kind should be placed. The results of user tests are introduced, and the advantages and drawbacks for both the user and the lexicographer are discussed and illustrated.

1. Introduction

The aim of this paper is to discuss a rather straightforward problem which is relevant to everyone involved in the use and the making of dictionaries: in which of the possible entries should lexical units consisting of more than one word be placed – and preferably – found? The principles applied in existing dictionaries seem to vary a great deal, and many dictionaries do not seem to have decided on a guiding principle. The result, of course, is a confused user. Fortunately, user-orientation and user-friendliness have now become more important issues in lexicography. Martin and Al (1990) mention as a crucial point finding a consistent procedure and then living up to it in order for the user to know where he should expect a certain type of information to be placed in the dictionary. In the case of multi-word lexical units, the procedure that many dictionaries follow, often without being explicit or even conscious about it, is that of the 'semantically heaviest word'. The only problem is knowing if this coincides with the user's expectations and/or intuition.

I agree with Martin and Al that a consistent procedure is necessary in order to make it easier for the user to consult the dictionary, once the principles of lemmatization are understood. This, of course, means that the principles have to be rather simple. In this paper I would like to present and discuss the principles used in The Danish Dictionary (a comprehensive dictionary of modern Danish to be published in 1999). The discussion will include the results of a test very similar to the ones presented by Bogaards (1990, 1992).

The types of lexical units relevant for this paper are:

- idioms: *stå på pinde for nogen* ('be at someone's beck and call')
- (technical) terms: *galvanisk element* ('galvanic cell')
- phrasal verbs: *bryde ud* ('break out')
- reflexive verbs: *more sig* ('enjoy oneself')

In the paper the term 'word group' will be used synonymously with 'multi-word lexical unit'.

2. Principles used in The Danish Dictionary

2.1 Place within the entry

First, I should like to point out that the problem to be discussed in this paper is which entry a multi-word lexical unit should be placed in. The exact place within the entry can also be discussed, but in recent years many dictionaries have preferred to place these items in a special section of the entry and list them alphabetically in order to make it easier for the user to find them. This solution has also been adopted for The Danish Dictionary where these items will be placed in a section of its own, as subentries after the main meanings of the headword, and before information about word formation and etymological information.

2.2 Word groups containing nouns or words used as nouns

When confronted with a multi-word lexical unit the lexicographer has to decide in which entry this particular item belongs. A simple solution would be to make an entry for each item of this kind, but most probably this would create more problems than it would solve, especially if the items were to be listed alphabetically according to the first word since these words tend either to be subject to variation or to be grammatical words with little semantic weight (cf. Botha 1992). Making a word group into a headword is necessary, however, if none of the words in the group can be recognized as a Danish word, e.g. *a cappella, da capo*. These are word groups that have been borrowed from a foreign language as they are, and the meanings expressed by the individual words outside these combinations do not exist in Danish; accordingly, they should not be described in our dictionary.

Apart from these special cases the procedure in The Danish Dictionary is to sublemmatize word groups under the first noun or, if there is no noun, the first word used as a noun:

1 *i nøden skal man kende sine venner* (lit. 'when in need you will know your friends' = 'a friend in need is a friend indeed') – is lemmatized under the noun *nød* ('need')
2 *for meget af det gode* (lit. 'too much of the good' = 'too much of a good thing') – is lemmatized under the adjective *god* ('good'), here used as a noun

The noun under which we sublemmatize the word group must be semantically 'heavy', cf. the expression *vide hvad vej vinden blæser* ('know which way the wind is blowing'), where the first noun *vej* does not carry any meaning in itself, but is a part of an adverbial group (*hvad vej*) that is often replaced by the adverb *hvorhen* ('where'). In this case the first noun that carries meaning is *vind* ('wind') in which entry the expression is sublemmatized.

The decision to place word groups under the first noun is due to the fact that we had an intuitive feeling that nouns often contribute heavily to the meaning of phrases. This corresponds to the findings of Bogaards (1990): that dictionary users (in Bogaards' case Dutch users) tend to look for word groups under the noun. The tendency first to look under the noun was partly confirmed in my own test (Lorentzen 1994) in which I asked 49 native speakers of Danish to indicate in which entries they would search for a certain number of lexicalized word groups. In many cases a majority of the subjects chose the first noun in the word group, but this tendency is partly in opposition to two other tendencies: **a)** a tendency to choose the first 'real' word of the group, i.e. noun, adjective or verb (cf. *Oxford Advanced Learner's Dictionary*: 1577); **b)** a tendency to choose an infrequent word, which could be a foreign word or a Danish word with low frequency:

3 *brændt barn skyr ilden* (lit. 'a burnt child avoids the fire' = 'once bitten, twice shy') – a majority of the subjects choose the first 'real' word *brænde* ('burn')
4 *alpin kombination* (lit. 'Alpine combination' = 'Alpine combined') – a large majority of the subjects prefer the infrequent and foreign word *alpin* instead of the frequent – and here semantically 'light' – word *kombination*

A solution to the problem of not knowing where the user will look for a particular expression would be to sublemmatize it under each word in the group. This would of course be very uneconomical. We prefer to give the expression a subentry under only one of the words involved, but with the possibility of making cross-references from one of the other (heavy) words of the group. This should also eliminate the risk of giving different explanations to the same expression, which is often found in existing dictionaries.

2.3 Word groups without nouns

If the word group contains no nouns or other words used as nouns, we use the following order to decide in which entry the word group should be sublemmatized:

- first verb (excluding auxiliary and modal verbs)
- first adjective
- first adverb

5 *det skal du få betalt!* ('you'll pay for this') – the modal verb *skal* ('shall') and the auxiliary verb *få* (approx. 'get') are ignored, and the word group is lemmatized under the first 'full' verb *betale* ('pay')
6 *nu og da* ('now and then') is lemmatized under the first adverb of the group: *nu*

This is very similar to the principles of *Longman Dictionary of the English Language* except that they use the order: adjective, adverb and then verb.

We may, however, deviate from the above-mentioned order on account of two important factors: **stress** and **variation**.

2.4 Stress

An important feature of Danish, that speaks in favour of *Longman*'s order, is the fact that stress and semantic weight often go together. This means that in many word groups the stress is not on the verb, but on e.g. an adjective:

7 *stirre sig 'blind på* (lit. 'stare until you become blind' = 'become obsessed by something')
8 *ligge 'brak* ('lie fallow' = 'be inactive')

In these cases we suspend the order mentioned above and lemmatize the word group under the adjective, preferably with a cross-reference from the verb. A general exception to this rule are the phrasal verbs in which the adverb particle always carries the stress: *bryde 'ud* ('break out'). We doubt that any users would try to find this kind of word group in the entry for the adverb particle. This corresponds to the principle mentioned by Svensén (1993:216) stating that word groups "should not be placed under the function words involved, unless it is these whose meaning is to be illustrated". That this principle corresponds with the expectations of the user was confirmed in my test, cf. the following two expressions:

9 *spille fandango* (lit. 'play fandango' = 'fool around')
10 *spille 'op* ('play up')

in which 86 % of the subjects preferred the noun (*fandango*) to the verb and 95 % preferred the verb (*spille*) to the adverb particle.

2.5 Variation

Idiomatic expressions are part of what is known as 'fixed phrases'. The question is how fixed are they? In most dictionaries – and presumably in the minds of many dictionary users – the variation within idioms is very little: the idea of a canonical form of an idiom seems to prevail. But once you look at so-called 'fixed' phrases in a large text corpus, you will find that the fixed part of them can be very limited.

As an instance of this I should like to examine the idiom *svaret blæser i vinden* ('the answer is blowing in the wind') and the examples found in our text corpus. This expression including variations of different kinds is found 20 times in the corpus, but only 7 times in the canonical form cited above. The other examples show different types of variation, most of them predictable on a morphological and syntactic level:

- variation in number: *svarene* ('the answers') (2 ex.)
- variation in tense: *blæste* ('was blowing') (1 ex.)
- addition of adverbs: *ikke* ('not'), *endnu* ('still') (3 ex.)
- modalization: *svaret må blæse* .. ('the answer must be blowing ..') (4 ex.)

This type of variation does not make it necessary to alter the form of the idiom, and the user should still be able to find the subentry for *svaret blæser i vinden* in the entry *svar* (first noun in the group) since this element seems only to vary in number. The situation gets more difficult in the examples where the word *svar* is replaced by other (more or less synonymous) elements: *myten* ('the myth'), *spørgsmålet* ('the question'), an indirect question: *hvad de mange penge er brugt til må blæse i vinden* ('what all that money has been spent on must be blowing in the wind'). We prefer to regard this type of variation as a creative use of the language (cf. Clausén 1994) that may be shown in the dictionary by means of additional examples if the variation is common; but if it is a well-established variation, the lemmatized form of the idiom must reflect the variation.

Another example of variation within idioms is the expression *have gelé i knæene* (lit. 'have jelly in your knees' = approx. 'your legs feel like jelly'). Our corpus contains 13 examples of which only one is in the form mentioned above. A number of the examples show that the verb can change and thus express different aspects: the verbs *få* ('get') and *give* ('give') both express the process instead of *have* ('have'), which expresses a situation. On the other hand, *få* focuses on the person whose legs feel like jelly, and *give* focuses on the person or phenomenon which causes somebody to feel that way. In other instances the expression has the form of a comparison: *hendes knæ er som gelé* ('her knees are like jelly'). It seems that the constant components of this idiomatic expression are *gelé* and *knæ*. Our solution is to decide on a standard form (though not very frequent in the corpus): *have gelé i knæene*, inform the user of the possibilities of variation and then give a citation that shows this.

3. Conclusion: advantages and drawbacks

After having presented the principles of lemmatization of word groups in The Danish Dictionary, I should like to sum up a few of the advantages and drawbacks of this model.

It seems that in some cases the user's first try will be in vain, especially in the case of word groups where the first noun is frequent and other words in the group attract the user's attention. Both the tests carried out by Bogaards (1990, 1992) and myself show a certain tendency to search under infrequent words: *merceriseret bomuld* ('mercerized cotton') under the (infrequent) adjective; *skifte heste i vadestedet* (lit. 'change horses in the ford' = 'midstream') under the infrequent noun

vadested. We try to help the user in this type of situation by giving a cross-reference to the correct entry if we suspect that many users will begin by looking up under other words than the first noun.

What is more important: by following a consistent and simple procedure, viz. lemmatizing under the first noun whenever the word group contains one, we expect that the user will soon learn the principle so that the number of times he looks up an expression in our dictionary in vain will be reduced to a minimum.

Notes

1. *Lemmatize* is used in the sense of 'make into a lemma', cf. Botha (1992).

References

Bogaards, P. 1990. "Où cherche-t-on dans le dictionnaire?" in: *International Journal of Lexicography.* Oxford, Oxford University Press, Vol. 2, No. 3, pp. 79–102.

Bogaards, P. 1992. "French dictionary users and word frequency" in: Tommola et al. (eds.): *Euralex '92, Proceedings I–II*, pp. 51–61.

Botha, W. 1992. "The Lemmatization of Expressions in Descriptive Dictionaries" in: Tommola et al. (eds.): *Euralex '92, Proceedings I–II*, pp. 465–471.

Clausén, U. 1994. "Idiom och variation" in: *Nordiske Studier i Leksikografi II.* Skrifter udgivet af Nordisk Forening for Leksikografi. Skrift nr. 2. Copenhagen, Gads forlag, pp. 47–52.

Longman Dictionary of the English Language. 1991. 2nd ed., London, Longman.

Lorentzen, H. 1994. "On Lemmatization of Word Groups". A user test and reflections on user behaviour (not published).

Martin, W. and B.P.F. Al 1990. "User-Orientation in Dictionaries: 9 Propositions" in: Magay and Zigány (eds.): *BudaLEX '88 Proceedings.* Budapest, Akadémiai Kiádo, pp. 393–399.

Oxford Advanced Learner's Dictionary of Current English. 1989. 4th ed., Oxford, Oxford University Press.

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