Editing multiple idiomatic phraseological units for book and CD-ROM

Lineke OPPENTOCHT, Utrecht, The Netherlands

Abstract

In dictionaries dating from pre-computerized times it is not uncommon that idiomatic phraseological units are given under more than one entry, in different forms and with different explanations. Once these units can be extracted automatically from structured text files you, of course, want to give such sets of phraseological units the same form and explanation everywhere. This is what we wish to do for future editions of the Dutch *Grote Van Dale*. At this moment, we are editing the file of the 13th edition of the *Grote Van Dale* (1999) in order to make the 1st edition of the *Grote Van Dale on CD-ROM* (september 2000). The CD will offer a tool to retrieve a list of all idiomatic phraseological units in the dictionary. This application demands a medium-specific approach to the editing of multiple idiomatic phraseological units. The question arises how to maintain the dictionary file in order to keep it suitable as the basis for book versions of the dictionary as well as for electronic versions (CD's).

1 Introduction

As in most larger monolingual dictionaries, many phraseological units can be found under more than one entry in the Dutch *Groot Woordenboek der Nederlandse Taal* (usually called: *Grote* or *Dikke Van Dale*), 13th edition (1999). A phraseological unit may be given under several entries for good reasons, e.g. to accommodate the user of the dictionary, or because the unit is a good illustration of the meaning of all the entry words where it is found.

In dictionaries dating back to pre-computerized times[1] however, one may find such *multiple phraseological units* in different forms and even with different explanations. For example, in the *Grote Van Dale* (from now on the GVD13) the phrase *niet meer op zijn benen kunnen staan* ('not be able to keep on one's feet any longer') is given in three entries, as follows:

op	(uitdr.) *niet (meer) op zijn benen kunnen staan* a) erg moe zijn b) erg dronken zijn
on	(phrase) *not be able to keep on one's feet (any longer)* a) be very tired b) be very drunk
been	(uitdr.) *ik kan niet meer op mijn benen staan*, zo zwak, zo vermoeid ben ik
leg	(phrase) *I can't keep on my feet any longer*, I'm that weak, that tired
staan	*de man kan nauwelijks op zijn benen staan*
stand	*the man could hardly keep on his feet*

Once these book versions of 'old' dictionaries are converted into a structured text file and all phraseological units can automatically be retrieved from the dictionary, the wish to give phraseological units occurring more than once in the dictionary the same form and explanation everywhere, is obvious.

2 Editing multiple idiomatic phraseological units for book and CD-ROM

For the first release of the *Grote Van Dale on CD-ROM*, we have focused our attention on the editing of form and meaning of *idiomatic* phraseological units. These are what [Van der Meer 1998] defines as *idioms*: preconstructed, not semantically literal word combinations (i.e. the words, or at least one of them, have not retained their conventional literal meanings, or at least cannot be analysed as such). We have included sayings, proverbs, similes, formulae and metaphors in this category.

Van der Meer opposes *idioms* to *free combinations* (not preconstructed and semantically literal) and *collocations* (preconstructed and semantically literal[2]). We have not edited the form and explanation of multiple occurring free combinations and collocations. One of the characteristics of these types of word combination is that they usually do not have an explanation in the dictionary. They need no other explanation than the general definition under which they are given in the dictionary article. Furthermore, it is quite natural that collocations in monolingual dictionaries are presented in different strings (forms) under different entries since they are given to illustrate the valency of entry words. Under the entry *commit* in a monolingual English dictionary, for example, we expect to find a string such as 'to commit a crime, a sin, theft, burglary' while under *crime* we would like to find something like 'to commit, perpetrate a crime'.

For idiomatic units, different principles apply. We take the position that for both the proper dictionary and the CD-ROM it is advantageous that an idiomatic phraseological unit can be found under different entries. There are, however, some medium-related differences:

- Book:
  - An idiomatic phraseological unit may be given under different entries but in order to prevent the dictionary from becoming unmanageable, the unit should get an explanation under one entry only. At the other entries the user should be referred to the entry containing the explanation. The idiomatic unit is ideally explained at the entry where the user of the dictionary is most likely to look it up. Up until now, however, it is not exactly clear how the dictionary users proceeds when looking up an expression (see [Boogaards 1990] for some suggestions). In principle, an idiomatic unit is only explained in the *Grote Van Dale* under, respectively, the first noun, adjective or verb it contains but for pragmatic reasons exceptions to this rule are allowed. For the GVD13, the editing of all idiomatic units according to these rules was started. This editing, however, was not completed. Therefore, a multiple idiomatic expression may still have an explanation under more than one entry in the book (cf. the example in section 1).
  - The form of idiomatic multiple phraseological units is ideally, but not necessarily, identical under all entries. The user of a book version of the dictionary will most probably not notice formal differences because once he has found a unit with an explanation under one entry he will not look up the same unit under another. Once multiple idiomatic units are only explained under one entry and cross-references to this entry are given under the other entries, their form should of course be identical.

- CD-ROM:
  - An idiomatic phraseological unit may be given under more than one entry. No space saving devices have to be used. In principle, it doesn't matter under which entry it is given because the user of the CD-ROM is supposed to use search mechanisms. The user doesn't have to decide under which entry to look for a unit. All he has to do is enter a form and the entry appears on screen.
  - Since no space saving devices have to be used, idiomatic phraseological units can be explained under each entry. A good CD-ROM should not force its user to jump from one entry to another to find a meaning.

The *Grote Van Dale on CD-ROM* offers the possibility to retrieve a list of all idiomatic phraseological units in de dictionary. When the user selects one of the units in this list, the entry under which he can find the meaning of this unit is shown. Considering all the above, the following question has to be answered: when an idiomatic phraseological unit is given under more than one entry in the dictionary file, which of those units should then be given in the list of idiomatic units? It would not be elegant to give them all. If, for example, we would ask a list of all idiomatic units in the GVD13 containing the word *tong* ('tongue'), or *hart* ('heart') without editing them first, part of the result (ordered alphabetically by the first word) would be as follows:

tong	*een zalfje voor de tong*
	een zware tong hebben
	een zware tong hebben
	goed van de tongriem gesneden zijn
	haar radde tong wist overal antwoord op
	haar tong is niet van (schapen)leer
	heb je je tong verloren?
	heb je je tong verloren?
	het hart ligt hem op de tong
	het hart op de tong hebben
hart	*waar het hart vol van is, loopt de mond van over*
	waar het hart van vol is, vloeit de mond van over
	waar het hart vol van is, daar loopt de mond van over
	waar het hart vol van is, loopt de mond van over
	waar het hart vol van is, loopt (of vloeit) de mond van over

Figure 1: Part of the list of all idiomatic units containing the string tong and hart respectively

In the list under *tong*, we find formally identical multiple idiomatic units (*een zware tong hebben* (2x, given under the entry *zwaar* and under *tong*) and *heb je je tong verloren*? (2*, given under *tong* and under *verliezen*)). There are also two formally different, but strongly related, idiomatic units which have the same meaning but are given under different entries*: het hart ligt hem op de tong* (given under *tong*) and *het hart op de tong hebben* (given under *hebben*). The unwanted

result is even more clear in the list of idiomatic units containing the word *hart*. One proverb is given under five entries (*vol*, *overvloeien*, *overlopen*, *mond* and *hart*), each time in a slightly different form, sometimes with and sometimes without an explanation.

We take the position that in the list of idiomatic phraseological units the user of the CD-ROM should be confronted with only one of a set of similar multiple idiomatic units. But how do we choose which one this should be?

We wish to work with one text file for both book and CD-ROM. However, as we stated above, on the CD the meaning of a multiple idiomatic unit should be given under each entry. In the book, on the other hand, the explanation should be given under one entry only and cross-references should be given at the other entries. This difference between CD-ROM and book is exactly what we have made use of to decide which idiomatic unit of a set will be presented in the list on the CD.

In the structured text file only one idiomatic unit of a set gets an explanation. The other units of the same set are marked with a double asterisk (**). Additionally, the latter units are followed by a field called <ster..> containing the entry word under which they are explained.

For example:

tong	<idvb..>	kind of expression	uitdr. phrase
tongue	<vbzi..>	expression	heb je je tong verloren? did you loose your tongue?
	<vbvk..>	explanation	kun je niet spreken, niet antwoorden? can’t you speak, answer?
verliezen	<idvb..>		uitdr.
loose	<vbzi..>		heb je je tong** verloren?
	<ster..>		tong

For the book, the field <ster..> will be replaced by a cross-reference (*zie*...(*see*...)) to the entry with the explanation (the entry *tong* in this example). For the CD the field <ster..> will not be replaced by a cross-reference but by the explanation given in the field <vbvk..> following the unit which does not contain the asteriks. When the user of the CD-ROM will not make use of the list of idiomatic units to look up the meaning of a unit but will use the CD the way he would use the book, he will find an explanation of the idiomatic unit under each entry under which this unit is given.

For the list of idiomatic units presented on the CD-ROM, only those units will be selected which do not contain the double asteriks, i.e. those units which are followed by an explanation instead of a cross-reference in the book.

3 Variants of idiomatic phraseological units

The field <ster..> can only be automatically replaced with the right explanation when the idiomatic units have exactly the same form throughout the dictionary[3]. Therefore, formal varia-

tions have to be edited first. We have, for example, to choose one form from the three variations on the expression *niet (meer) op zijn benen kunnen staan* etc., given under *been*, *op* and *staan* (see section 1). We prefer the infinitive form. Therefore, we choose the form *niet meer, nauwelijks op zijn benen kunnen staan* (the parentheses of *meer* have been deleted because corpus analysis indicated that the idiomatic unit doesn't occur without an adverb such as *meer* ('any longer') or *nauwelijks* ('hardly')). After editing the form and explanation of the expression the result is as follows:

been	<vbzi..>	expression	niet meer, nauwelijks op zijn benen kunnen staan
	<vbvk..>	meaning *a*	erg moe, zwak zijn;
			be very tired, weak
	<vbvk..>	meaning *b*	erg dronken zijn
			be drunk
op	<vbzi..>		niet meer, nauwelijks op zijn benen** kunnen staan
	<ster..>		been
staan	<vbzi..>		niet meer, nauwelijks op zijn benen** kunnen staan
	<ster..>		been

The example above illustrates what we did when one phraseological unit was given in different forms throughout the dictionary. The same principle can be applied when there are two different but related phraseological units having the same meaning. For example *dat is koren op zijn molen* ('that is grist to his mill'), explained (differently) under *koren* and under *molen* in the GVD13, and *dat is water op zijn molen* ('that is water to his mill'), explained under *water* and *molen*.

In our approach these expressions will be explained under *molen* ('mill') . It is not certain that this is the entry under which the user is most likely to look first. However, the advantage of choosing this entry is that under a single entry (*molen*) information on the possible variants of the expression can be given. Moreover, this approach allows us to save a lot of space in the book. The result is as follows:

molen	<vbzi..>		dat is koren op zijn molen
	<frla..>	frequency	niet alg.
			not common
	<vbzz..>	alternative	dat is water op zijn molen
	<vbvk..>	meaning *a*	dat komt hem goed van pas
			it's a great opportunity for him
	<vbvk..>	meaning *b*	dat is naar zijn zin
			he likes it very much
koren	<vbzi..>		dat is koren op zijn molen**
	<ster..>		molen
water	<vbzi..>		dat is water op zijn molen**
	<ster..>		molen

Both the <vbzi..> and the <vbzz..> under *molen* will be given in the CD's list of idiomatic units because they do not contain the double asterisks. So, when the user selects the expression

dat is koren op zijn molen or *dat is water op zijn molen*, the entry *molen* will be presented on screen. For the CD, an explanation will be given under *molen*, *koren* and *water*. In the book, an explanation can be found under *molen* only. Under *koren* and *water* the user will be referred to *molen*. The alternative, and for the book space demanding, situation would be as follows:

molen	<vbzi..>		dat is koren** op zijn molen
	<ster..>		koren
	<vbzi..>		dat is water** op zijn molen
	<ster..>		water
koren	<vbzi..>		dat is koren op zijn molen
	<vbvk..>		dat komt hem goed van pas
	<vbvk..>		dat is naar zijn zin
water	<frla..>		niet alg.
	<vbzi..>		dat is water op zijn molen
	<vbvk..>		dat komt hem goed van pas
	<vbvk..>		dat is naar zijn zin
	<vbsy..>	neutral syno	dat is koren op zijn molen
	<vgvw..>	compare	koren

Alternatively, to save space in the book, the explanations under *water* could be omitted; we could confine ourselves to giving only the neutral synonym. Another possibility is to omit the expressions in the entry *molen*. Both approaches, however, would not be very user-friendly.

4 Conclusion

In this paper we have explained how we edit multiple idiomatic phraseological units for the Dutch *Grote Van Dale* so as to obtain a good basis for publishing future book versions of the dictionary and for making the first edition of the Grote Van Dale on CD-ROM (2000). We showed how the dictionary file can be edited in such a way as to offer the user of the CD-ROM the possibility to retrieve a non-redundant list of all idiomatic units in the dictionary. We also showed that the described approach has advantages for future book versions of the dictionary: the form of multiple idiomatic units will be identical under each entry, the user will no longer be confronted with different explanations, variants of idiomatic units will be given under a single entry while they were often given in different entries in previous editions, and, last but not least, a lot of space will be saved.

Notes

[1] The first edition of the *Grote Van Dale* under the name Van Dale dates back to 1872.

[2] See [Cowie 1998] for discussions on the demarcation between types of word combinations.

[3] This will not be necessary once hard links will be established between phraseological units in the database.

References

Boogaards, Paul (1990). Où cherche-t-on dans le dictionnaire? in: *International Journal of Lexicography*. Vol. 3/2, pp. 79-102.

Cowie, A.P. (ed.) (1998). *Phraseology. Theory, Analasys, and Applications.* Clarendon Press, Oxford.

Geeraerts, Dirk (2000). *Groot Woordenboek der Nederlandse Taal (CD-ROM)*. Van Dale Lexicografie, Utrecht/Antwerpen.

Geerts, Guido and Den Boon, Ton. (1999). *Groot Woordenboek der Nederlandse Taal*, 13th edition. Van Dale Lexicografie, Utrecht/Antwerpen.

Meer, Geart van der (1998), Collocations as one particular type of conventional word combinations. Their definition and character, in Fontenelle, T., Hiligsmann, A., Moulin, A., Theissen. S. (eds.), *EURALEX Proceedings I-II, Vol. I*, Liège, Belgium, pp. 313-322.