Where to Find Fixed Phrases in the New Van Dale Groot Woordenboek Hedendaags Nederlands

Jaap Parqui
Van Dale Lexicografie bv
P.O. Box 19232
3501 DE UTRECHT
THE NETHERLANDS
jaap.parqui@vandale.nl

Abstract

VLIS, one of the data structures in use at Van Dale, allows us to use rules to generate dictionary articles. For the new edition of the *Groot woordenboek hedendaags Nederlands*, we developed a rule that decided for each fixed phrase where in the dictionary, i.e. under which headword, it would be placed. The rule says that a fixed phrase will be placed under the first noun in the fixed phrase. If there is no noun, it will be placed under the first adjective; if there is no adjective, under the first verb. In this article I will discuss why Van Dale created this set of rules, and which exceptions and alterations were required.

Introduction

Several years ago, Van Dale developed a data structure called VLIS [Van Dale Lexicografisch Informatie Systeem]. VLIS contains the data of several comprehensive dictionaries, such as Dutch-English, Dutch-French and the *Groot woordenboek hedendaags Nederlands* (contemporary Dutch). These dictionaries all use the same Dutch side, so that Dutch information only needs to be edited once, instead of individually for each dictionary.

The VLIS system distinguishes between two kinds of phrases: so-called free phrases (examples, textual illustrations) and fixed phrases. Free phrases are illustrations of the use of a particular meaning of the headword, for example *ze zingt maar middelmatig* [her singing is pretty average] under the headword **middelmatig** [average]. Fixed phrases are more or less fixed combinations of words that occur frequently together. They include collocations, in which the individual words are used in their literal sense and for which no explanation is required. For example: *een verzilverde lepel* [a silver-plated spoon], *iemand onder narcose brengen* [give someone an anaesthetic]. Also included under fixed phrases are idiomatic expressions, in which one or more of the words are not used in their literal sense. In these cases an explanation is required. For example: *de kat op het spek binden* = 'iemand in verleiding brengen' [tie the cat to the bacon = 'tempt someone']. You might say that fixed phrases with explanation are simply entries in the dictionary, and that all that separates them from headwords is the fact they consist of more than one word.

Phrases in a dictionary serve several purposes: they help clarify a certain meaning of a word, but sometimes the user of the dictionary is actively looking for a certain phrase, for example if he wants to know its precise form or exact meaning.

The information in VLIS, including these phrases, has been stored in as flexible a way as possible. This means we can use rules to generate dictionary articles. In this paper, I will

describe how we developed a rule that decided for the third edition of the *Groot woordenboek hedendaags Nederlands* under which headword a fixed phrase should be placed.

Limitations and Choices

When we made the *Groot woordenboek hedendaags Nederlands*, we were forced to make choices, at least for the paper version. On the one hand we had an abundance of material and we wanted to place phrases under as many headwords as possible. On the other hand we were faced with limitations as to the volume of the dictionary. That meant we had to make some choices: what to include, what not to include, and especially, where to include it.

For most headwords, fixed phrases explain the meaning of the headword sufficiently. We therefore decided not to include the free phrases (examples and textual illustrations). Due to the limitation in volume, the fixed phrases could not however be included under *all* the headwords appearing in the phrase. We therefore needed a rule that would determine under which headword the user of dictionary would find a certain fixed phrase.

Which Principle to Apply?

Not much research has been done into the question under which headword the user expects to find a fixed phrase. If we want to look up *dat is een pak van mijn hart* [*that is a load off my chest*], do we look under **pak** or under **hart**? Will we find *honger is de beste saus* [hunger is the best sauce] under **honger, best** or **saus**? Or, more difficult, where to look for *een zielig hoopje mens* [a sorry heap of a person] or *pappen en nat houden* [dress and keep wet]? It would be very useful if we created a rule to tell users under which headword in the dictionary they should look for a certain phrase. It would be even better if that rule were to correspond closely with our intuition, so that it would also work for people who don't actually know the rule.

We needed to decide what the underlying principle of our rule should be. Should it be word frequency? Grammatical category? The syntactic structure of the phrase? Idiomaticity? Paul Bogaards (1990) published some interesting research into this issue. His main conclusions were that test subjects do indeed have a systematic strategy for looking up fixed phrases, that this strategy differs per language, and that for Dutch the look-up strategy is based on grammatical category. The test subjects first try to look under the noun, if there is one in the phrase, then under the adjective, if present, and lastly under the verb. The frequency of the individual elements turned out not to be particularly relevant.

The Set of Rules

It looks therefore as if people use an intuitive strategy when looking up fixed phrases. It seems logical to use precisely that strategy to formulate the rule for our dictionary. An additional benefit to this rule is that it is clear: frequency and idiomaticity are rather vague notions, but it is always obvious which word in a phrase is the first noun.

So we developed a set of rules that met these conditions. Wherever possible, the *Groot woordenboek hedendaags Nederlands* would place a phrase under two headwords. If there were two nouns in the phrase, it would be placed under both. If there were no two nouns, it would be placed under the adjective; if there was no adjective, under the verb, so

for example under noun-noun, noun-adjective, adjective-adjective, adjective-verb. *Dat is een pak van mijn hart* will therefore be found under both **pak** and **hart**. *Honger is de beste saus* will be found under **honger** and **saus**, *een zielig hoopje mens* under **hoop** and **mens** and *pappen en nat houden* under **nat** and **pappen**.

Contrary to the CD-ROM, the book only gives the explanation, if there is one, under the first of the two headwords. The phrase under the second headword refers to the first headword by means of an asterisk. So, only under **nat** will we find 'op de huidige manier doorgaan zonder de benodigde duidelijkheid te verschaffen' ['continue in the current manner without much-needed clarification'] as the explanation of the meaning of *pappen en nat houden*. Under **pappen**, an asterisk in *pappen en nat houden* refers us to **nat**.

Refining

By applying our rule to the content of the second edition of the *Groot woordenboek hedendaags Nederlands*, we essentially gave it a complete overhaul. Because the majority of phrases would be placed automatically, without direct intervention of an editor, we had to refine our rule to avoid certain problems.

Support Verbs Excepted from the Rule

The rule moved many fixed phrases to a different headword. As a consequence, some headwords collected huge numbers of phrases. This was especially true for support verbs such as *doen, worden, maken* [do, become, make]. They were flooded with phrases, even though according to the rule, verbs are only in third place in the headword priority.

Articles that run over several columns are not very user-friendly. Moreover, because the actual meaning of support verbs is rather vague, it did not make sense for them to contain so many phrases. We therefore decided that these verbs would be excluded from our rule. They would not be a valid option for the placement of phrases.

On the other hand, we didn't want this type of verbs not having any fixed phrases at all. We therefore manually selected a number of phrases that were placed under support verbs, despite their exception to the rule.

Closed Class Words Explained

The new distribution of fixed phrases also led to articles not containing any phrases at all. Because fixed phrases are placed under nouns, adjectives and/or verbs, closed class words in particular (such as prepositions and conjunctions) would be devoid of any phrases. But because the meaning of closed class words is difficult to capture in definitions, example phrases are essential. We therefore decided that closed class words were also excepted from the rule and that they were allowed to have a number of phrases.

More Explanation Needed

According to the rule, our dictionary provides the explanation of a fixed phrase under the first headword, and under the second headword it refers to the first. There were two groups of phrases for which it was necessary to provide an explanation of the phrase under the second headword as well as under the first.

The first group is made up of phrases that are idiomatic: they do not correspond to the meaning of the headword, for example *een morganatisch huwelijk* [a morganatic marriage] under the headword **morganatisch**. Under the rule, the explanation would be given under the first noun, i.e. **huwelijk**. The headword **morganatisch** does not have any meanings, but contains only this phrase. If a user were to look up the headword **morganatisch**, she wouldn't find any explanation of the word at all, which is not very user-friendly. We therefore decided to provide explanations of all phrases not corresponding to the meaning or one of the meanings of the headword. So, in the case of *een morganatisch huwelijk*, the headword **morganatisch** *does* give the explanation 'huwelijk waarbij de ene partner en eventuele kinderen niet delen in al de rechten, aan de adellijke stand van de ander verbonden' ['marriage in which one of the partners and any children do not share the rights connected to the title of the other partner'].

The second group for which an additional explanation was needed, is made up of metaphors and similes. The rule dictates that the explanation of *zo gek als een deur* [as mad as a hatter] is placed under **deur**. But the phrase says something about the notion **gek** [mad]; in this case reinforcing it. For that reason, it will be to the advantage of the user to give the explanation under the headword **gek** as well. Metaphors and similes – tagged as such in the database – are therefore explained under all substance words present in the phrase.

Duplicate and Semi-Duplicate Phrases

In the previous edition of the *Groot woordenboek hedendaags Nederlands*, fixed phrases were often given under different headwords, sometimes with slight differences in form. If we would redistribute the phrases in accordance with our rule, duplicate and semi-duplicate phrases would be thrown together, seeing as they would be placed under the same headword. Naturally, we did not want this to happen. We therefore made good use of a bad situation and picked the standard form of each phrase, at the same time making the dictionary more consistent.

Arranging Phrases within Articles

The principle described above was not only useful for deciding under which headword phrases should be placed, but it also served to arrange phrases within articles. We have seen that, in principle, phrases are placed under two headwords. If a phrase is listed under a certain headword, the *other* word has been given a blue colour: it serves as a guideword. So the headword **pak** lists the phrase *dat is een pak van mijn hart* with **hart** as the guideword and printed in blue. Under the headword **hart**, the phrase is given with **pak** as the guideword and printed in blue.

Fixed phrases are listed at the end of the article, in alphabetical order of guideword. Because all guidewords are blue, it is easy to find the right phrase by quickly scanning the article.

Figure 1: Article 'figuur' with guidewords in blue (circled)

Conclusions

A newly developed rule can be applied by automatic process, provided a few preconditions are met. The database should be flexible: phrases should be linked to several headwords. The collection of phrases in the dictionary should be consistent, and some aspects of our rule had to be refined.

At the time of writing, the new edition of the Van Dale *Groot woordenboek hedendaags Nederlands* has been on the market for almost a year. The system described above, with its exceptions and alterations, seems to be very workable. The limited volume of a paper dictionary has been utilised in such a way that in the majority of cases, the user will immediately find what he is looking for.

References

Bogaards, P. 1990. 'Où cherche-t-on dans le dictionnaire?' in *International Journal of Lexicography*, vol. 3, no. 2. Oxford: Oxford University Press.

Van Sterkenburg, P. G. J. et al. 1991. *Groot woordenboek van hedendaags Nederlands*, 2nd ed. Utrecht/Antwerpen: Van Dale Lexicografie bv.

Van Sterkenburg, P. G. J. et al. 2002. *Groot woordenboek hedendaags Nederlands*, 3rd ed. Utrecht/Antwerpen: Van Dale Lexicografie bv.