Piet G. J. van Sterkenburg,
Institute for Dutch Lexicology and University of Leyden

Electronic onomasiology: Van Dale greater dictionary of synonyms

ABSTRACT: This paper serves a threefold purpose: (a) presentation of the concept, features and structural elements underlying a new Dutch dictionary of synonyms and other words with related meanings; (b) demonstration of the possibility to structure, based on a lexicological concept, a database according to strict rules. Special attention will be paid to how the whimsicality of the language can be made controllable by applying a structure in which a hyponym may only have one hyperonym; (c) pointing out the additional value of this electronic taxonomy of Dutch for other lexicological products.

1. Introduction

Most Dutch dictionaries marketed by Van Dale Lexicografie Publishers BV in Utrecht and Antwerp answer questions such as: "What is the definition of a specific word?". Of course, other questions can be asked of the lexicon. In the course of 1987 the then-Director of Van Dale Lexicografie, Prof. Bernard Al, asked us, therefore, to develop a prototype for a **"Greater Dictionary of Synonyms and other meaning-related words"**. See Van Sterkenburg 1991.

Apart from the fact that we argued that the concept should fit into the theoretical framework of cognitive semantics and, more particularly, a prototype framework, we also took the view that the dictionary envisaged should not be an exhaustive inventory of the Dutch language, but should be a current one (i.e. with words like *blooper, chemocar, intifada, lambada, pincode, train taxi,* etc.). However, the dictionary should in particular provide the answer to another type of question, such as: by means of what words or word groups is a given meaning expressed, and how are words expressing similar notions mutually related? Also questions like: "What is the name of a poem retracting what was said before?" (*palinode*), or: "What do you call a poem starting all lines with the same letter?" (*tautogram*) should be able to be answered with the help of this new dictionary. Other conditions to be met by the product to be developed were the following:

(a) The dictionary should contain a large collection of synonyms of the following type: *melody = tune; bicycle = bike; pastry = cake; fool = take in = hoax = trick = hoodwink,* etc. The differences in use between, for instance, *minister* and *excellency,* or differences in style level: *drunk: plastered,* should be unambiguously classified.

(b) Words are mutually related as to their hierarchical meaning. One word includes the other, so to say, has a wider sense, and belongs to a higher class. Therefore, a tight hierarchical network should constitute the basis for the total collection.

(c) Room should also be made for words with a narrower or more restricted meaning than the headword. The word with the wider sense is the hyperonym, the one with the narrower or more specific meaning is called the hyponym. So not only for the hyperonym *vehicle*, but also for the hyponym *motor vehicle*. Moreover, the subtle difference in meaning between the headword or hyperonym and the hyponym should be indicated by means of characteristic shades of meaning.

(d) In addition to synonymous, hyperonymous and hyponymous relations, antonymous and associative relations should also be included. An associative relation is a reference to an other conceptual meaning under which the word from which the reference is made could also be classified. *Servant* is hyperonym to *mate, houseman* and *valet*, but would do as well in a tree of *personnel*.

(e) Standing expressions should also be included, and they should, of course, come with the headwords to which they belong with respect to content, and not with one of the words from the expression. So *pull someone's leg* will be found under *fool* and not under *leg* or *pull*.

(f) In his search for variety the user should always be able to find more than a simple headword with a reference.

(g) Apart from these linguistic and typographic desiderata, both publisher and editors felt the need to be able to work with advanced electronic systems from the very start. In other words: it had to be possible to manipulate both the source file and the target file automatically.

2. Set-up

When comparing dictionaries of synonyms, you will soon discover that there are three different set-ups, roughly speaking. First of all you have the books which can be best described as the world in so many notions. In addition to these there is a substantial group of works having one alphabetical list as its starting-point, and the third group works with two alphabetical lists.

The idea behind a dictionary of synonyms dividing the world into a number of notions (for instance 1,000), is roughly the following: often people do not have a specific word in mind to which they want an equivalent, but only a notion or a concept for which they need the right word. The world can be divided into a number of areas, for instance abstract relations, and matter and the will. And those areas are again built up from sub-areas. The abstract relations, for instance, include such notions as *time* and *order*, among *will* we find mutual wanting in addition to deeds of will. Somebody who has a concept in mind does know whether that concept ranges among, say, *business* or *feelings*.

It is rather arbitrary to divide the world into a thousand notions; more or fewer would also have been possible. The world in a thousand pieces creates a coarse-mesh image where it is difficult to predict under which one the concept you have in mind is hidden. If there are many more pieces, the mesh will be too fine as a result, and you cannot at all find the way. It is so difficult to predict and find the way because the classification depends largely on the personal taste and world view of the makers. This is a problem which cannot be solved. For this reason we deemed it advisable to abandon the idea of a classification of the world in making a dictionary of synonyms. For the user is not inter-

ested in the whole world, is he? If the notion synonym is not interpreted too strictly he must be able to find what he is looking for through an alphabetical list in a much easier way.

By far, most dictionaries of synonyms are composed as a normal dictionary: one long series of entries which have been alphabetically arranged. It goes almost without saying that how long the entries are, and how much and what extra information they contain is dependent on the target and volume of the book. Information on parts of speech, sample sentences, stylistic indications, expressions and antonyms can all be found in this second type of dictionary of synonyms. The user knows immediately whether it contains a direct synonym of what he is looking for, and, if he is lucky, he will get quite a few extras offering him more insight or additional options. If he does not know exactly what he is looking for, however, the odds are that he will have to leaf and read through a lot of entries before he will be where he wants to be.

The broad set-up of dictionaries of synonyms with two alphabetical lists as their starting-point is simple: one list contains all words occurring in the book; the second list consists of the alphabetically arranged lemmata. The order of these two can vary, and so can length and content of the lemmata, as well as the number of words included. Often the lemmata have in common that a hyperonym constitutes the entry.

In describing meaning-related words we have tried to arrange the (hierarchical) relations between words and word groups for various conceptual meanings on the basis of family likeness. With the help of synonyms, hyperonyms, hyponyms and other relations we have composed a kind of tree structure of conceptual meanings intended to visualize the way people classify their thoughts.

The theoretic model used was that of cognitive semantics, in particular that of the prototype theory, as the following example may illustrate. The conceptual meaning of "sleeping-place" is, for instance, given by the lexemes *three-quarter bed, crib, plank, camp bed, bedstead, berth, couchette, box spring, box bed, bed, bunk, couch* and *bunkbed*. Of this series *bed* was regarded as a central synonym or as a prototype for our culture. In the framework of the above-mentioned theory the presentation of the semantics existing in the lexicon of a language, and as described by us, is one which can be imagined as a kind of continuum from typical to less typical representatives of a conceptual meaning, or from core to periphery.

3. Source file

One of the possibilities on the basis of which a skeleton of hierarchical relations could be built up is to make use of electronic files of general Dutch dictionaries. Particularly because analytic definitions have been frequently used in them, these dictionaries can be alphabetized on the family from those definitions. In this way an interesting collection of hyperonyms can be made available relatively quickly. We demonstrate that, though not exhaustively, again by the gender and hyperonym *poem* which was mentioned above:

poem containing a denouncing, usually jesting idea: epigram
poem in which somebody or something is lamented: lament
poem of a tender and intimate feeling: madrigal
poem singing the praises of a sublime subject: ode
poem written by somebody shortly before his death: swan song

Van Dale, however, searched the basic material for this dictionary in the series of large electronically stored multilingual dictionaries for present-day use (German-Dutch, English-Dutch, French-Dutch). In Al (1988, 10-13) we read how all Dutch words occuring as synonyms in the foreign language-Dutch volumes of that series have been selected with the help of the computer. The French-Dutch dictionary, for instance, gives:

grâce
0.1 gracefulness → sweetness, grace, loveliness, charm, elegance
0.2 favour → graciousness
0.3 mercy
0.4 grace → forgiveness, mercy
0.5 thanks → gratitude

In German-Dutch you will find:

Anmut
0.1 gracefulness → sweetness, grace, charm
Gnade
0.1 mercy → favour, benevolence, grace, charity

In English-Dutch can be found:

grace III <n.count.noun>
0.1 gracefulness → grace, charm, loveliness, elegance
charm II <n.count.noun>
0.1 charm → attractiveness, appeal, enchantment

Series of synonyms from German-Dutch, English-Dutch and French-Dutch having at least one word in common have been linked, for instance, from English-Dutch the series of translations of *grace* and of *charm*. But also the translations of the French *grâce* and those of *Anmut* and *Gnade* have elements in common with the already mentioned series, and were consequently linked. As a result one large chain of words came into being, and provided the following picture:

(1) grace = allure, attractiveness, charm, appeal, enchantment, gracefulness, amiability, elegance, refinement, loveliness, sweetness, decoration.

In addition to this cluster there are other collections of which grace is a part, being:

(2) grace = charity, mercy, affection, benevolence, sympathy, goodness, favour, kindness.

(3) grace = amnesty, blessing, mercy, exoneration, pardon, commutation, remission, forgiveness, absolution.

Example (1) almost totally consists of earlier-mentioned elements. Series (3) is composed of the translations of *grâce* (0.4), *Gnadenakt* (0.1) and of translations of, for instance, the English words *grace, pardon* and *reprieve*.

But the series (1), (2) and (3) are also mutually related because the word *grace* occurs in all three of them; (2) and (3) even more because *mercy* is a part of both. And in this way more cross references could be mentioned. The fundamental cause of this crisscross is the fact that particularly the most frequently used part of the lexicon is characterized by polysemy (Al 1988, 12).

4. Processing phases

Two main methods of processing were planned for the source file. First of all, preliminary editorial processing would take place, including the selection of the headwords. Then the canonical form would be determined, the conceptual meaning registered, and the relations between meanings established in terms of synonyms and hyponyms, or rather in terms of horizontal and vertical relations. The horizontal ones, the brothers and the sisters, have an almost similar meaning; whereas the vertical ones have a specific meaning. In the actual editorial phase the specifications would be introduced with the help of more explicit differentiations and indications.

In cooperation with a registered information scientist a functional design was made. A functional design describes how a static underlying abstract structure (derived from the editorial concept) leads to concrete interpretation of the data model with the help of rules or functions which structure the data. In our case this meant that the editors could assign editorial characteristics to words, and define relations between the so named words using the functions. Moreover, the functions monitor agreement with the model. Anything contradictory to the underlying data model was rejected.

5. Editorial process

Let us return to the above-mentioned source file. After the somewhat more abstract considerations on data model and functional design we are in a better position to explain how the editorial work was done on the computer system.

The source file consisted of 80,000 words. From them the editors selected 25,000 words and split these words into 42,000 word meanings. Through the computer system, a Vax which was operated online, the editors linked each word meaning to another word meaning, i.e. either to the preferred synonym or, if the word itself was a preferred synonym, to the hyperonym. If there was no further hyperonym to be found, the word was marked as the tree top. Thus language trees were grown, tops with downward branches.

Now let us examine step by step what all this meant in practice:

(1) The source file offers the following series of words in alphabetical order: *accurate, astute, attentive, careful, concentrated, conscientious, eager, intent, modest, moved, observant, perceptive, profound, quiet, sedate, solemn, tense* and *watchful.*

(2) For the time being *attentive* is taken to be the most general word of this series.

(3) The prototypical meaning of *attentive* is established and consequently documented by linking it to the relevant meaning number in Van Sterkenburg 1984. The meaning number in question is entered behind *attentive.*

(4) Then the other words are treated accordingly. It is determined what meaning of moved is similar to *attentive* or what meaning can be linked up with the prototypical meaning of *attentive,* on the basis of family likeness. The outcome is that there is no connection whatsoever between moved and the present prototypical meaning of *attentive.* The prototypical meaning of *moved* bears close relation with that of *touched.* For that reason the editors have made a link between the two words. Here the meaning numbers from the above-mentioned dictionary are also linked to the words to be connected.

(5) *Accurate, concentrated, intensive, observant, perceptive, tense* and *watchful* are directly related to *scrupulous.*

(6) In determining the preferred synonym, a definite choice is made for *attentive,* which ranks as prototypical with the conceptual meaning mentioned. *Accurate* and *watchful* do not seem synonyms. They appear not to have a horizontal but a vertical relation with *attentive.* Both words are included in *attentive. Accurate* is "very attentive" and *watchful* is "attentive with regard to danger". These are clearly hyponyms of *attentive* here. Also *tense* is a more extensive form of *attentive.* The notion [+ anxious] is a part of it. That notion, however, is expressed by other words such as *breathless, under the spell of, fascinated,* etc. After editing the entire file *fascinated* appeared to qualify best to make itself known as prototypical for this horizontal relation.

(7) *Withdrawn in oneself, intent, eager, solemn, careful* were put aside for reasons into which we cannot go here in more detail.

(8) *Modest* went to *unassuming, conscientious* to *careful, quiet* to *silent, astute* to *smart* and *sedate.*

(9) Elsewhere in the file the link was made between *considerate* and *attentive,* words which were deemed similar at the horizontal level. At the vertical level *attentive* forms a link with the words *alert, fixed on, intensive, sympathetic* and *careful.*

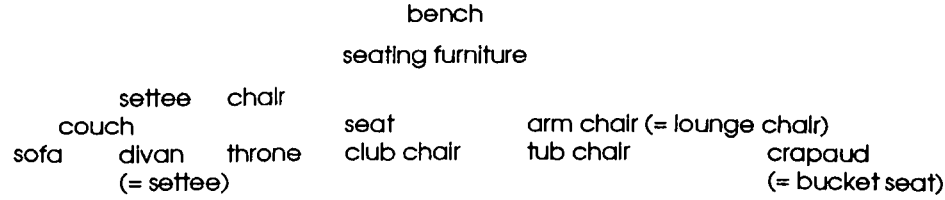
For all words selected from the source file and many supplements from other sources, the actions mentioned above were performed. The same went for more than 10,000 standing combinations.

We would like to demonstrate what happened further in this phase, with the help of an "evergreen" from structuralism. The equal sign stands for synonyms, the arrow denotes hyponomy and the colon symbolizes the top of the tree:

couch → seating furniture (on which two or more persons can be seated)
canapé = sofa
club chair → arm chair (low and round)
settee = divan
tub chair → arm chair (low and round)
divan → settee (low without back)
lounge chair = arm chair
bucket seat = tub chair
arm chair → chair (easy, with armrests)
bench: top
sofa → couch (upholstered, with back)
chair → seating furniture (with back and legs)
throne → seat (of a monarch)
seat → chair (large and impressive)
seating furniture → bench (to be seated on).

Apart from making synonymous and hyponymous links, shades of meaning were assigned to those hyponyms. In our example they are placed in brackets.

As a result, a computer program generated the following schematic tree:



Besides the nuances innumerable indications have been introduced manually, like stylistic indications (formal, informal, vulgar and archaic), group language indications (children's language, slang, biblical language, youth language, conference language and r.c.), attitudinal indications (jesting, ironic, pejorative, insulting and euphemistic), and finally, words restricted to a certain region within the language area, or words revealing the professional jargon from which they originate. Because I cannot demonstrate all this within the framework of this paper, I will give only one expressive example:

brothel
= sex-club
= house of ill fame
= whorehouse (inf.)
= bawdy house (inf.)
= cathouse (inf.)
= massage salon (euph.)
= indulgence house (euph.)

Furthermore the following is fundamental in our concept: a hyperonym can have more hyponyms: next to *aperitif* you will find *beer, beverage, boilermaker, cocktail, long drink, refreshment, sherry with orange juice* and *shot*, all hyponyms of *drink*. A hyponym, however, can only have one hyperonym. Thus in our example *beer* and *shot* cannot also belong to the hyperonym *alcoholic drinks*. Hyponyms can all, of course, in their turn, have horizontal relations again. Thus *pint, half-pint, Pilsner beer* and *draught* function as hyponyms of *beer*, whereas *beer* has as synonyms *ale, porter* and *stout*. We have selected this option because for the time being we do not deem ourselves capable of making a multi-dimensional distinction between salient and non-salient conceptual data in a formal model of the mental lexicon. Partial overlap of conceptual meanings cannot (yet) be found in this dictionary. In the concept of this dictionary the evident agreement of meaning between for instance *bolt* and *sledge* (both objects slide) cannot be expressed (Verkuyl 1991).

When clusters of meaning are structured as described above, tree structures will be formed as mentioned before. In both the horizontal and the vertical relations the prototypical terms function as junctions or branches.

The method used and the formal conditions fulfilled in making these links could be described in a number of logical rules. This enabled input into and control by a relational databank.

From the very start it was clear that it would be impossible to present the trees in their entirety in a dictionary. There are trees with 9 or 10 layers. This was not only graphically impossible, but it would also become very complex for the user who is looking for direct information: if a tree is 10 layers deep, all those layers will have to be examined in order to be able to find the right word. Therefore it was decided to present only two layers at a time in the book. The headword with its synonyms, the hyponyms and their synonyms. Realization of this choice was, of course, a piece of cake for the computer.

Hyponyms of hyponyms are not found there, but with the hyponym as headword. Thus we find the hyponyms *hockey field* and *football field* for *sports field* (which itself is a hyponym of *sports centre*) not under the latter headword, but under *sports field*. If a word itself is not a top, but a hyponym, arrows make clear at the foot of the headword what higher layers there are in the semantic tree, of which the headword consulted by him is a pruned branch. The user can see with *sports centre* that *centre* and *field* constitute the next higher layers. It may happen that the path to the top is the only information given with a headword. This is the case if a hyponym itself has no synonyms or hyponyms. Example: *antiquities room*, where references are made to the layers *museum, building* and *edifice*.

6. Conclusion

Bertrand Russell made the following statement: "the most essential characteristic of mind is memory". Thinking of this statement upon completion of this dictionary of synonyms, the publisher should keep in mind that the external memory storing the electronic file is a vital source for the development of new lexical products. Moreover, that file is of great significance as well for the existing series of explanatory and multi-lingual dictionaries. In order to improve the quality of those dictionaries, the conceptual meanings from the dictionary of synonyms can be taken as a starting-point. Greater consistency of the semantic profile in those dictionaries can be reached by editing words by concept and not alphabetically. Such a method will lead, in passing, to a definition grammar.

As an electronic product the synonyms file can, of course, also come in useful in the field of word processing or in synonyms banks. Moreover, the possibility to open up archives which are inaccessible because questions are not put unambiguously could be considered. If, for instance, one is looking for the word *bicycle* in documents, and it does not occur, no information will be obtained, but if the synonyms of *bicycle* can also be used as search words, accessibility comes much closer.

The editors will, however, find deepening and formalizing the network structure between the conceptual meanings by far the most challenging task. The challenge of approaching a description of the structure of the mental lexicon and of the organization of the human mind casts an irresistible spell on this linguistic project.

7. Bibliography

AL, Bernard P. F. (1988): "Op zoek naar Synoniemen". In: Jaarboek 1987-1988 Corpusgebaseerde Woordanalyse 7-14.

STERKENBURG, Piet G. J. van, e.a. (1984): Van Dale Groot woordenboek hedendaags Nederlands. Van Dale Lexicografie. Utrecht/Antwerpen. Second edition 1991.

STERKENBURG, Piet G. J. van, e.a. (1991): Groot woordenboek van Synoniemen en andere betekenisverwante woorden. Van Dale Lexicografie. Utrecht/Antwerpen.

VERKUYL, Henk (1991): "Groot woordenboek van Synoniemen en andere betekenisverwante woorden". In: Forum der Letteren 32:312-315.

KEYWORDS: Computer-aided lexicography, conceptual meaning, dictionaries, dictionary of synonyms, hyponyms, hyperonyms, lexical taxonomy, lexicology, lexicography, onomasiology, synonyms