Frames and Semagrams. Meaning Description in the General Dutch Dictionary

Fons Moerdijk
Universiteit van Amsterdam

This paper discusses the semagram, an innovation in the way of describing meaning in lexicography, as used in the Algemeen Nederlands Woordenboek *(General Dutch Dictionary). A semagram is the representation of knowledge associated with a word in a frame of slots and fillers. Slots are conceptual structure elements which characterise the properties and relations of the semantic class of a word-e.g. colour, smell, taste, composition, components, preparation for the class of beverages. The abstract meaning frame for such a semantic class is called type template. After a motivation for the use of frames in lexicography we reveal how semantic classes are determined and how type templates are composed. We illustrate this with the type template of the animal names and show how the semagram of cow is based upon it. We conclude by summing up the main advantages of the use of semagrams.*

1. Introduction

This paper discusses a new way of describing meaning in lexicography, as used in the *Algemeen Nederlands Woordenboek* (General Dutch Dictionary), further abbreviated as ANW. The ANW is a comprehensive online scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, the Dutch speaking part of Belgium. The first results will be published on the web in 2008. One of its main innovations is a twofold meaning description: definitions will be accompanied by "semagrams".

A semagram is the representation of knowledge associated with a word in a frame of "slots" and "fillers". "Slots" are conceptual structure elements which characterise the properties and relations of the semantic class of a word (e.g. COLOUR, SMELL, TASTE, COMPOSITION, COMPONENTS, PREPARATION for the class of beverages). On the basis of these slots specific data is stored ("fillers") for the word in question. Frame and semagram refer to different things. A frame is an abstract structure schema, which is called a "type template" in ANW jargon. A semagram is such a frame populated with concrete word data (see also Moerdijk 2002 and Moerdijk 2004).

The use of frames for the representation of meaning in dictionaries, is a novelty, but of course does not appear out of the blue. In metalexicographic publications, the possibility and desirability of using frames has already been mentioned for a few decades. The initial impetus was given by Wierzbicka (1985), who provides exhaustive definitions in terms of conceptual structures according to certain semantic schemas, and Wegner (1985). This line of research is in particular continued in Germany with publications by Wiegand (1989 and 1992), Konerding (1993), Konerding and Wiegand (1994), Kammerer (1999) and Bublitz and Bednarek (2004). In South Africa the influence can be found in publications by Louw (2000) and Smit (2000).

Most mainstream linguists will immediately think of Fillmore and his FrameNet when talking about frames. Understandable, but the ANW frames cannot be identified with the FrameNet ones. The frames of FrameNet represent a conceptual structure of a stereotypical *stituation*. The participants, object, properties and events, that can be distinguished in such a situation, form the frame elements. These are found in sentences by establishing their semantic role with regard to the predicating word. Frame elements are "conceptual roles" which are used to annotate corpus sentences semantically. I would like to call this type of frames, borrowed from syntactic analysis and intended for the semantic annotation of sentences "syntagmatic frames".

The ANW, on the other hand, is more in line with Wierzbicka-Wiegand-Konerding: the frame represents a conceptual structure which is not a description of a situation, but of *the lexical concept*, of the bundle of knowledge which can be associated with a word form. The frame elements are not "conceptual roles", but conceptual features. They are not the result of a semantic-syntactic analysis and their goal is not the semantic analysis of sentences. Because of this clear distinction with FrameNet, I call this type of frames "paradigmatic frames".

Another famous lexical semantic resource is Wordnet. Wordnet is a large lexical database of English, developed under the direction of George A. Miller. Words are grouped into "synsets", sets of synonyms, each expressing a distinct concept (cf. *http://wordnet.princeton.edu/*). These synsets are interlinked by means of conceptual-semantic and lexical relations, especially relations of hyponymy and hyperonymy (in Wordnet's own terminology: hypernymy). The English Wordnet inspired dozens of similar wordnets in other languages. Their aim is to fix linguistic ontologies based on (mainly taxonomical) relations between word senses, but they do not give analyses of these word senses themselves. That is precisely what the ANW aims at in the semagram. So the ANW-semantics differs not only from Framenet, but also, and even stronger, from Wordnet.

2. Motivation for the use of frames

The need to include semagrams in addition to definitions in dictionary entries stems in first instance from the consideration that definitions alone cannot explain meaning. There is often a lot more semantically relevant knowledge associated with a word than can be shown in a definition. This is of course mainly a problem for printed dictionaries. Furthermore, definitions often suffer from arbitrariness and inconsistency as well with regard to the choice of the "keywords" for the "genus proximum" as the nature and the amount of features included. The same meaning is often expressed in different ways in different dictionaries, and within one dictionary the definitions of instances of the same semantic class can differ significantly. Frames should guarantee more uniformity and consistency.

The application of frames in lexicography can also be motivated from the rise of cognitive semantics in the last quarter of the previous century and the theory about word meaning associated with it. Up to then the structural-semantic vision predominated, i.e. the meaning of a word can only be determined in relation to other words in the same semantic field. Only those features should be included in the description of the meaning of the word in question that clearly define it from other words. Only necessary and distinctive features should be taken into consideration. These are the pure semantic features. Encyclopedic features do not belong in a definition. The cognitive-semantic vision assumes that the meaning of a word is a lexical concept which can be described separately from its relation to other words. In this description not only necessary and sufficiently distinctive features matter, but also those features which are (proto)typical and not necessarily distinctive. Thus the boundary between semantics and encyclopedia is removed.

Lexicographers describe language use. Frames are a means to record how our knowledge and our perspective on reality are reflected in language use. The cognitive-semantic notion of meaning is closest to the kind of lexicography which determines and describes meaning on the basis of analysis of real language utterances (quotations or corpus concordances). The ANW practises this kind of lexicography and takes with the introduction of frames in dictionary writing a step which logically follows from the influence of cognitive semantics.

A third and very important reason for using frames in lexicography follows from the nature of the ANW dictionary being an electronic dictionary. Frames provide an increase in search and query facilities. This is particularly the case for queries guiding the user from content to form.

3. The development of frames as "type templates"

The ANW adopted its own method to develop the frames or "type templates" which roughly consisted of the following tasks:

- content analysis of existing dictionary definitions;
- definition and clustering of semantic classes on the basis of keywords from those definitions;
- determination of the features which were indicated in the analyses of those classes.

The choice to work "bottom up", i.e. starting from analyses of definition contents, followed from the wish to base the determination and classification of the semantic classes and subclasses as far as possible on linguistic foundations. There are numerous ontologies available, but they are, in most cases, classifications of reality which are based on other than linguistic considerations. A division of the lexicon of a language in which words are brought together in classes and subclasses for substances, artefacts, persons, activities, states, emotions, vehicles etc. is linguistically more relevant than a division in which such words are distributed over various social domains like politics, religion, marriage, sports, weather, etc.

In Konerding and Wiegand (1994) dictionary definitions also play a role in the development of frames. However, in their work they do not form the basis, but are used as proof afterwards to confirm the reality of previously defined frames. Konerding and Wiegand give so called "Minimalframes" for lexicographic practice, after which they determine how segments of definitions of three larger German dictionaries fit in them. Those Minimalframes are reductions of "Matrixframes", which have been composed on the basis of a bunch of queries which are intendend to discover and to record systematically the stereotypical, general "implicit" knowledge of language users.

A content analysis of a definition implies that each definition is divided into segments, the keyword (or sometimes the group of keywords) is marked and the features which express the remaining components are indicated. For a simple definition "adult female bovine" for *cow* such an analysis gives the following result:

KEYWORD: bovine
AGE: adult
GENDER: female

For more elaborate definitions such a segmentation is of course more complex, but in general still well possible although the denomination of the nature of the features can sometimes be hard. Moreover different labels can be used for one and the same feature.

The content analysis was performed on a selection from a list of approximately 20,000 base words (no compounds and no derivations). This selection was more or less random, more or less, because certain conditions were taken into account. On the one hand, the sample could not be too large, but on the other hand it could not be too small either. In order to get as many classes as possible and to be able to find enough representative words per class, one should dispose of a rather large number of words. In addition the three main syntactic categories had to be represented. In the end, the total selection list comprised 3535 words, subdivided as follows: 1331 nouns, 1679 verbs and 525 adjectives.

The decision to start with a selection of base words rests on the insight that in modern lexicography editing should no longer be done alphabetically, but according to "lexicographic types", words that semantically, syntactically and morphologically belong together (Apresjan 1993). The base words form morphologically such a group. They play a "natural", leading role in the formation of new words. A lot of derivations and compounds are formed with them and inherit their conceptual-semantic structure, with some more specifying or additional features. Consequently, with the semagrams of the base words the semagrams of their derivations and compounds are to a large extent given. The semantic description of the latter can be completed faster and more adequate, when one already possesses the semantic descriptions of the former.

The fact that different dictionaries give different definitions for one and the same meaning and that one dictionary gives definitions which are different in size and nature for instances of one and the same semantic class meant that several dictionaries had to be consulted for the analysis. The electronic version of the WNT and the electronic version of the GWHN were chosen. As several dictionaries were used and as for polysemous words more than one primary meaning was taken into account in the analysis, the total number of analysed definitions was more than 7,000.

In a second phase the analyses of the definitions were subjected to various, critical clustering and refining tasks. First the semantic classes were derived from the keywords, which often indicated the adjacent overarching semantic class of the head word. In addition, higher semantic categories were indicated. A few examples:

koe (cow) → keyword: bovine → upper category: animal
oog (eye) → keyword: visual organ → upper category: body part
paniek (panic) → keyword: fear → upper category: sense
ruïne (ruins) → keyword: remains → upper category: building
taxi (taxi) → keyword: hire car → upper category: vehicle.

It was not straightforward to derive classes and subclasses from the analyses. The above examples illustrate this. For *koe* (*cow*) the upper category could also be "mammal", for *oog* (*eye*): "organ", for *paniek* (*panic*): "condition, state", for *ruïne* (*ruins*): "part" (analysis remains problematic), for *taxi* (*taxi*): "car". Briefly, the definition of such conceptual categories is subject to a fair amount of variation, both with regard to the identification of the nature of the categories as with regard to the naming. Consequently further adaptations and uniformations were necessary.

After identification of the classes and subclasses, templates with class features were compiled by adding together all the analysed features for all the words belonging to those classes. Similar to the categories we were confronted with a lot of variation in the qualification of the features. Again clustering, modifications and uniformations were necessary. For the noun this resulted in 266 semantic classes (and as many templates), subdivided into 49 main classes. For the verb we obtained 195 classes in 22 main classes. For the adjective we got 9 main classes.

For the superordinate classes templates were developed cumulatively grouping together all features from the subclasses in one all-comprising "type template".

The analyses of the definitions and the following modifications and uniformations resulted in 166 features, divided into 22 main groups. The number of features is much lower than the number of classes because certain features occur in more than one class.

Let me illustrate the above with an example of animal names. There were 92 animal names among the selected words. Analysis of their definitions led to subclasses and type templates for mammal, bird, fish, insect, reptile, amphibian and mollusc and a total type template for the main class "animal" which looks as follows (the numbers in brackets indicate the frequency with which this feature was found in the definitions of the 92 animal names):

VISUAL PERCEPTION
- COLOUR (13)
- APPEARANCE (10)
- SIZE (9)
- SHAPE (7)
- SOUND (1)

COMPOSITION
- PARTS (57)
- BUILD (28)

FUNCTION OR PURPOSE
- FUNCTION (13)

EVALUATION AND VALUE
- VALUE (1)

COMPARISON
- RESEMBLANCE (6)

ORIGIN
- SOURCE OR ORIGIN (3)
- REPRODUCTION (1)

SPACE/PLACE
- HABITAT (21)
- PLACE OF ORIGIN (3)
- PLACE (1)

TIME
- AGE (5)

STATE
- STATE IN GENERAL (6)
- PHASE (3)

LIMITATION
- SORT SPECIFICATION (11)

RELATION AND CONNECTION
- RELATION (1)

PROPERTY/QUALITY
- GENDER (13)
- ABSTRACT PROPERTY (3)

ACT OR ACTIVITY
- BEHAVIOUR (20)
- MOVEMENT (9)
- EFFECT (3)
- PURPOSE AND USE (1)
- TREATMENT (1)

RELATION TO PEOPLE OR OBJECTS
- PRODUCT OR FRUIT (7)

REST GROUP
- EXAMPLE OR SPECIMEN (6)
- CIRCUMSTANCE (1)

Even if we do not take the less frequent features into account, this type template is much more extensive and detailed than the structure schema Wierzbicka (1985) gave for "animal" which consisted of HABITAT, SIZE, APPEARANCE, BEHAVIOUR and RELATION TO PEOPLE. Martsa (1998: 117) simplified this even further by integrating SIZE into APPEARANCE.

Type templates as the one for "animal", whereby, of course, the other classes all have their own set of features, form the basis of the semagram.

4. The semagram

In the introduction I characterised the semagram as a type template populated with concrete word data. The data are derived from our own ANW corpus (104 million tokens), dictionary definitions (in particular WNT, GWNT, GWHN), and encyclopedia information (in particular Wikipedia).

The method of populating the template slots differs significantly from the one used in the example articles from Kammerer (1999) and Smit (2000). They give elaborate frames with long, dense text describing the features in full often quite comprehensive sentences and with a lot of encyclopedic detail. We did not go that route.

For the ANW we adopt the principle that the semagram should be about the description of meaning. Recognising that it is difficult to distinguish semantic and encyclopedic knowledge, does not mean that all sorts of factual knowledge should be included that dictionary users are not interested in. Peculiarities which are too detailed or too encyclopedic and subjective knowledge do definitely not belong in the semagram.

Second, the fact that the ANW is an electronic dictionary has influenced the method of completion in a principled way. An electronic dictionary offers several ways to acces the dictionary information. Not only from word form to content, but also from content to form. It should be possible for a user who cannot think of e.g. the word *apiarist* to find this word through separate content elements (e.g. "bees", "keep") that he does know and could use for a search. However, it should not only be possible to go from content to the appropriate word. It should also be possible to retrieve a set of words on the basis of one or more content features. Thus a user should be able to retrieve all names for female animals in Dutch on the basis of a

query combining the field CATEGORY with the value "animal", and a field GENDER with the value "female". This is also one of the reasons why the features in our structure are and remain more finely split; in the more global meaning schemata from Wierzbicka and Martsa such queries cannot be resolved.

We give the characterisation of those features in terms of short statements about the headword, i.e. "introductory sentences" which follow more or less naturally from the nature of the feature (e.g. CATEGORY → "is a"/ "is", MATERIAL → "is made of", FUNCTION → "is used to", "is used for", PLACE/LOCATION → "is located on/in/between/under" etc., COMPOSITION → "consists of", "contains", etc). Such sentences are particularly well suited to get an impression of the meaning starting from the word form, i.e. for "semasiological" queries. To facilitate the retrieval for queries from content or parts of the content to the matching words, the "onomasiological queries", we complement those sentences after a "#" character (a hash) with one or more keywords with or without adding relevant synonyms or other relevant words. The data after the hash will not be visible to the dictionary user on the screen though.

Below the semagram for *cow* (translation of the original Dutch *koe*):

A COW

UPPER CATEGORY:	is an animal # animal; mammal; ruminant
CATEGORY:	is a bovine (animal) # bovine; ruminant
COLOUR:	is often black and white spotted, but also brown and white spotted, black, brown or white # black and white; brown and white; red and white; spotted; black; blackspotted; white; brown; rusty brown
SIZE:	is big # big
SOUND:	moows/lows, makes a sound that we imitate with a low, long-drawn " boo' # moo; moows; low; lows; boo
PARTS:	has an udder, horns and four stomaches: paunch, reticulum, third stomach, proper stomach # udder; horns: paunch; rumen; honeycomb bag; reticulum; third stomach; omasum; proper stomach; abomasum
BUILD:	is big-boned, bony, large-limbed in build # big-boned, bony, large-limbed
FUNCTION:	produces milk and (being slaughtered) meat # milk; flesh; meat; beef; milk production; meat production
PLACE:	is kept on a farm; is in the field and in the winter in the byre # farm; farmhouse; field; pasture; meadow; byre; cow-house; shippon; stable
AGE:	is adult, has calved # adult; calved
GENDER:	is female # female
PROPERTY:	is useful and tame; is considered as a friendly, lazy, slow, dumb, curious, social animal # tame; domesticated; friendly; lazy; slow; dumb; curious; social
BEHAVIOUR:	grazes and ruminates # grazes; graze; ruminates; ruminate; chew the cud
TREATMENT:	is milked every day; is slaughtered # milk; slaughter
PRODUCT:	produces milk and meat # milk; meat
VALUE:	is useful # useful

Although the lexicographer does not need to worry too much about the split between semantics and encyclopedia, he still has to decide what is relevant and appropriate for inclusion and what is not relevant and can therefore be ignored. Another problem is that in some places the division of the features is too fine-grained and that certain ones are so close to each other that a particular value can be included in more than one place. For the semagram for *cow* this seems to be the case for instance for FUNCTION and PRODUCT, FUNCTION and USE, PROPERTY and BEHAVIOUR, PROPERTY and VALUE, and for PLACE one could wonder whether HABITAT would not have been

possible too here. It is therefore reasonable to assume that based on such experiences the editorial process will require further modifications and additions to the type templates.

I will now discuss in some detail the functioning of the semagram in the dictionary practice. In the ANW we distinguish the following five main search options:

- search for information about a word (combination, expression, idiom, proverb);
- search for a word;
- search for words with one or more common features;
- search for example sentences;
- search for other dictionary information.

The semagram plays a role in the first three of them. In the first one, the option that can be identified with the traditional search for information about a word in the article of a printed dictionary, the semagram is presented together with the definition. Its function here is: to give more semantic and encyclopedic information than can be given in the definition in a systematised, explicit and consistent way. In this search option, the semagram will be shown to the user in the form illustrated in the *cow*-example, without, as mentioned, the hash and the words behind it.

The second search option can be used when one has forgotten a word, or when one wants to know whether the language has a word for a certain concept or not. Here the semagram as such plays a different, implicit role. It does not become visible to the dictionary user, but functions behind the scene. The user is offered two search methods to arrive at an answer. First, he can give a definition, description, paraphrase or sum up synonyms or other words that he can associate with the word he is looking for. If this does not lead to a satisfying result, he can make use of a guided search method. In this guided search the user can choose the category (the semantic class or subclass) of the word from a menu (is it a thing? a person? an animal? a vehicle? a tree? an emotion? an activity?, etc.). Once a category has been chosen, the features of the type template for that category appear on the screen and the user is asked to fill in the value(s) that spring to mind. We illustrate this with a screen capture for the animal-example:

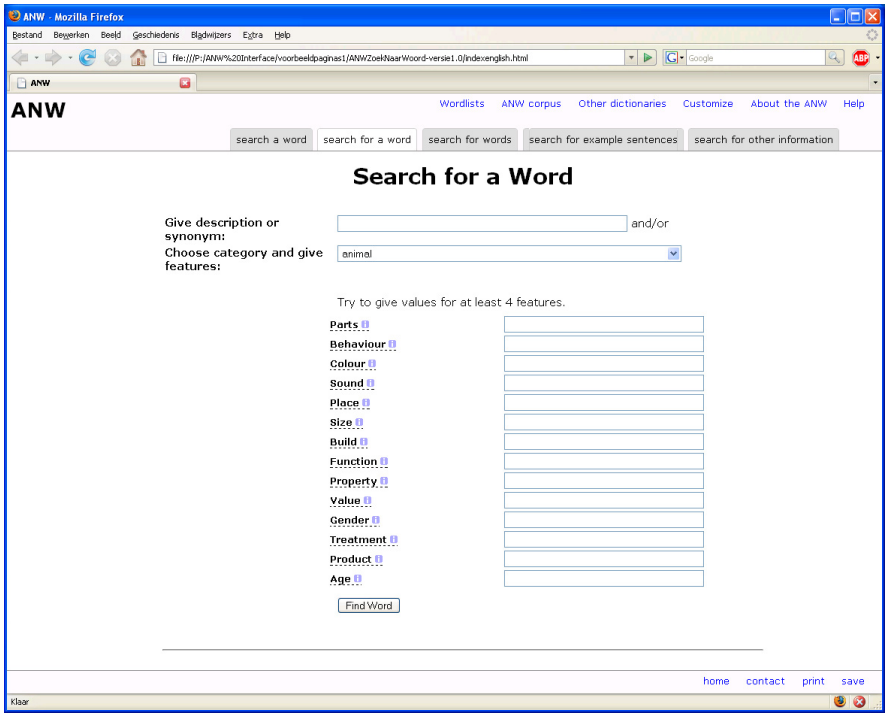

Figure 1. Screen for “Search for a Word”

With a press on the button “Find word” the data inserted by the user will be compared with the data of the semagrams in the dictionary database. Now the words behind the hashes are also involved in the retrieval process and the matching cases (in the best scenario just one!) are shown.

The third option is especially relevant for linguists and other language professionals who need examples and other materials for their work. It enables them to gather words that share one or more identical features within the main dimensions orthography, pronunciation, morphology, pragmatics, meaning, combinatorics, idioms, etymology. The semagram is, of course, active in searches in the semantic domain. Its role is here to some extent comparable with its role in the search for a word, going from content to form. In the third option, however, users can search in the dimension “Meaning and Use” for all the words that belong to a certain semantic class, for all the words that share one or more particular features, or for all the words sharing both class and certain features, instead of searching for just one word. To realise such a semantic search the user needs to assign values to the boxes for the category and the features of the type templates, that will be presented on the screen. The data of the completed form will then be compared with the data in the database and the words where there is a matching semagram will be shown on the screen. For instance, someone who had filled in “animal” for the CATEGORY and “farm” for the feature PLACE will get a list of all animals kept on a farm, including our *cow*.

5. The importance of the semagram

Insertion of semagrams into the semantic dimension of an electronic dictionary leads to a much richer semantic description, in which the implicit knowledge of the definitions has been made explicit and more (also encyclopedic) knowledge data are recorded than can be represented in the traditional definition formats. More lexical semantic relations than the well known traditional ones can also be discovered.

Semagrams which are in first instance derived from definitions themselves, should in turn result in better and more uniform definitions. Certainly when editing is done modularly per semantic class decisions can be made about the choice of keywords and about the number and nature of the features to be included in the definition and the semagram.

Furthermore, an electronic dictionary with semagrams opens a lot of new perspectives for onomasiological queries, going from content to form. Introducing the semagram, the ANW will bring us nearer the realisation of an old lexicographer’s dream: the combination of the semasiological and the onomasiological dictionary.

Bibliography

Apresjan, J. (1993). "Systemic Lexicography as a Basis of Dictionary-making". *Dictionaries* 14. 79-87.

Bublitz, W.; Bednarek, M. (2004). "Nur im begrenzten Rahmen: *Frames* im Wörterbuch". In

Herbst, T. et al. (eds.). *Lexikografie, ihre Basis- und Nachbarwissenschaften. (Englische) Wörterbücher zwischen "common sense" und angewandter Theorie*. Tübingen: Max Niemeyer Verlag. 35-51.

[GWHN]. *Van Dale Groot woordenboek hedendaags Nederlands*. CD-ROM 2.0. Utrecht: Van Dale Lexicografie b.v., 2002.

[GWNT]. *Van Dale Groot woordenboek van de Nederlandse taal*. 14th edition. CD-ROM. Utrecht: Van Dale Lexicografie b.v., 2005.

Kammerer, M. (1999). "Zur framebasierten lexikographischen Bedeutungsbeschreibung von substantivischen Lemmazeichen". *Lexicographica* 15. 229-263.

Koenen Wolters' Woordenboek Nederlands. Groningen: Wolters-Noordhoff, 1992.

Konerding, K. P. (1993). *Frames und lexikalisches Bedeutungswissen. Untersuchungen zur linguistischen Grundlegung einer Frametheorie und zu ihrer Anwendung in der Lexikographie*. Tübingen: Max Niemeyer Verlag.

Konerding, K. P.; Wiegand, H. (1994). "Framebasierte Wörterbuchartikel. Zur Systematisierung der lexikographischen Präsentation des Bedeutungswissens zu Substantiven". *Lexicographica* 10. 100-170.

Louw, P. (2000). "Presentation of Semantic Information in General Monolingual Dictionaries". *Lexikos* 10. 119-137.

Moerdijk, F. (2002). *Het woord als doelwit*. Amsterdam: Amsterdam University Press.

Moerdijk, F. (2004). "Het Algemeen Nederlands Woordenboek (ANW)". *Nederlandse Taalkunde* 9. 175-182.

Martsa, S. (2000). "On Animal Frames in English and Hungarian". In Mogensen, J.; Pedersen, V.; Zettersten, A. (eds). *Proceedings of the Ninth International Symposium on Lexicography April 23-25, 1998 at the University of Copenhagen*. Tübingen: Max Niemeyer Verlag. 113-124.

Smit, M. (2000). "Using Frames in Special-Field Lexicography: An Ethnomusicological Case Study". *Lexikos* 10. 173-188.

Verschueren Groot encyclopedisch woordenboek. 13th edition. Antwerpen: Standaard Uitgeverij, 1996.

Wegner, I. (1985). *Frame-Theorie in der Lexikographie*. Tübingen: Max Niemeyer Verlag.

Wiegand, H. (1989). "Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch". In Hausmann, F. et al. (eds.). *Wörterbücher. Ein internationales Handbuch zur Lexikographie/Dictionaries. An International Encyclopedia of Lexicography/Dictionnaire. Encyclopédie internationale de lexicographie*. Berlin: Walter de Gruyter. 530-588.

Wiegand, H. (1992). "Elements of a Theory towards a So-called Lexicographic Definition". *Lexicographica* 8. 175-289.

Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma Publishers.

[WNT]. *Woordenboek der Nederlandsche Taal* (1864-1998). CD-ROM. The Hague: Sdu Uitgevers, 2003.