

Basics of a comprehensive semantic categorization of Dutch verbs

Frans Heyvaert

Keywords: *semantic categorisation, verb meaning, definition structure.*

Abstract

Based on a definition analysis project carried out at the INL, this paper puts forward a proposal for a semantic categorisation of Dutch verbs in which existing category systems like the ones in Framenet and in Levin (1993) are integrated with some findings that emerged from the project. A multilayered category structure is proposed in which data about Aktionsart, conceptual field and systematic semantic analysis are all made explicit. This type of categorization is to be used in a Dutch dictionary project, the *Algemeen Nederlands Woordenboek* (ANW). It is meant as a tool to make common verb definitions in the dictionary more uniform and systematic, but also as a service to scholarly dictionary users whom it will enable to extract easily and systematically bodies of semantically related data from the dictionary.

1. Origins

In this paper we present some ideas about how verbs can be categorised semantically, based on a research project into the structure of verb definitions in Dutch dictionaries carried out at the Instituut voor Nederlandse Lexicologie (INL) between 2003 and 2006. For this project some 4000 meaning definitions of monomorphematic verbs, taken from the most important dictionaries of Dutch, were analysed into headwords and modifications. The headwords were clustered in groups of synonyms and near-synonyms in order to form categories; the modifications were grouped according to the type of information they contained (such information categories as Agens, Patiens, Manner, Place, Time etc.). These manipulations resulted into a provisional set of Dutch verb categories each of which was defined by a set of properties distinguishing it from all other categories, and by a set of modification types typical for that category. The aim of the whole operation was to establish a vehicle for a uniform treatment for all Dutch verbs to be used in a new lexicographical project, the *Algemeen Nederlands Woordenboek* (hence ANW), a monolingual dictionary of present-day common Dutch. A provisional overview of Dutch verb categories based on this project can be found in Heyvaert (2006). At present the original set of categories is being refined in order to be incorporated in the ANW.

The next two sections of the paper will be devoted to some earlier proposals that have significantly inspired the INL project. In the last section some conclusions based on our own findings will be presented and an attempt will be made at developing an integrated categorisation system in which insights from different perspectives are combined into one consistent whole.

2. Basic categories

Proposals as to how verbs should be categorised semantically differ with respect to how deep and how detailed one wants this categorisation to be. To our knowledge the most basic type of categorisation has been proposed in Dik (1978). Dik distinguishes only four categories, on the basis of two oppositions: [STATIC/DYNAMIC] and [+/- CONTROLLED]. DYNAMIC are those verbs that express that some change in the world takes place, whereas verbs are STATIC when this is not the case. A verb is +CONTROLLED when what it expresses is

performed willingly (normally by humans or sometimes animals) and –CONTROLLED when there is no idea of intentionality. So on this view there are basically four types of verbs: (1) STATIC/–CONTROLLED (*staan* ‘stand’¹, *slapen* ‘sleep’, *regenen* ‘rain’, *lijden* ‘suffer’), (2) STATIC/+CONTROLLED (*werken* ‘work’, *spelen* ‘play’), DYNAMIC/–CONTROLLED (*vallen* ‘fall’, *barsten* ‘burst’) and DYNAMIC/+CONTROLLED (*bouwen* ‘build’, *verven* ‘paint’). More or less the same distinction is to be found in Miller and Johnson-Laird (1976), but within a more complex context in which there are more distinctive features involved on the same level.

Basically this is a plausible way of organising the verbal universe: these are four fundamental things that can be the case in the world and that can be expressed by verbs. But there are some problems with it. The first can best be termed an aesthetic problem. The category of static verbs unites such words as *behoren* ‘belong’, *betreffen* ‘concern’, *regenen*, *draaien* ‘rotate’, on the –CONTROLLED side and *werken*, *spelen* as well as *luieren* ‘be idle’, *rusten* ‘rest’ in the +CONTROLLED part. Intuitively one feels that some refinement is possible here: *behoren*, *betreffen*, *luieren*, *rusten* show a kind of staticness different from that of *regenen*, *draaien*, *werken*, *spelen*. The first group denotes real, complete stasis, whether intended or not. The second group denotes some kind of static ‘operativeness’. This distinction may play no role at all in deep theoretical semantics, but it certainly does on a more superficial level, as in dictionary definitions: it is hard to explain to a dictionary user that, if someone works, nothing happens. A second obstacle is that some verbs are neutral with respect to one or both of the categorisation criteria. So e.g. *roll* can be used in one and the same sense as +CONTROLLED and –CONTROLLED: *Jan rolde van de heuvel* ‘John rolled off the hill’ (because he stumbled ↔ because that was the quickest way to get down). Lots of verbs can also be DYNAMIC or STATIC while used in the same sense, depending on the context in which they are used. This is the case e.g. with telic and atelic use (cf. Verkuyl 1993). *Jan at een koekje* ‘John ate a cookie’ is dynamic, *Jan eet elke middag koekjes* ‘John eats cookies every afternoon’ is static, while the sense of *eat* in the two sentences remains the same.

Another observation is that Dik’s STATIC – DYNAMIC opposition closely resembles the STATE – EVENT distinction made in Jackendoff (1983). What is added by the [+/-CONTROLLED] opposition is only some information about the subject type and about the intentionality of what is happening. However, Jackendoff’s distinction concerns not verbs but sentences. The same holds for information on the semantic role of the subject, which concerns the categorisation of sentences as well as verbs. The observation that the same categorisation can be used for both sentences and verbs may raise some doubts about Dik’s proposal. The problem can be illustrated by an example. Both *the paperclip keeps the papers together* and *the papers lie on the table* are sentences that denote states, or static, uncontrolled states of affairs. But the verbs in them attribute a quite different role to the subject. In the first sentence a subject is needed that causes the state, in the second sentence a subject that just is in that state. Similar examples can be found for controlled states and for uncontrolled events, not at first sight for controlled events. It appears that the opposition between causative and non-causative verbs is something quite fundamental that pervades the whole of verb categorisations. The elaboration of this observation will, however, be postponed to a later stage of the project. But the observation itself makes the proposal on deep semantic verbal categories made below only provisional.

In the categorisation system proposed here Dik’s system is adopted with some modifications. In the first place Dik’s definition of his categories, which is basically presented as an adjectival modification², has been translated into something more verbal in nature: some formula that can be used as a – remote – hypernymic expression over the verbs in the category. Second, the four basic categories that Dik puts forward are extended to five.

[+CONTROLLED, DYNAMIC] is translated into *een handeling verrichten* ‘perform an action’. [-CONTROLLED, DYNAMIC] becomes *betrokken zijn bij een gebeurtenis* ‘be involved in an event’. [+CONTROLLED, -DYNAMIC] is termed *een bezigheid hebben* ‘have an occupation’³. [-CONTROLLED, -DYNAMIC] is split into two basic categories, *in een toestand verkeren* ‘be in a situation’ and *betrokken zijn in een proces* ‘be involved in a process’, in order to make the distinction between pure stasis and continuous activity. The counterexamples to Dik’s division above also make it necessary to discern, next to the five categories proposed, disjunctive basic categories such as *either perform an action or have an occupation* and *either have an occupation or be involved in a process*. So in principle the number of basic categories can be multiplied to 25. In principle even disjunctions of three or more basic categories are possible, so that the number of possible categories increases significantly, but quite a lot of them will be practically excluded for there being no verbs to fill them, like *either perform an action or be in a situation*.

3. Conceptual domains and inheritance

Next to this rather elementary form of categorising verbs according to just two basic oppositions, there is a series of inspiring major studies of a more detailed and comprehensive nature that cannot be overlooked in any attempt to find some structure in the verb stock of a language. Especially prominent are the *Framenet* account and the account of Levin (1993). Not only do they offer a very rich and detailed stock of categories, but they also reveal a lot about what should be the basic philosophy behind semantic verb categorisation.

In the *Framenet* system (<https://framenet.icsi.berkeley.edu/>) the basic principle for the formation of semantic categories is the idea of syntactic-semantic frames: sameness of the environment of arguments and modifications. Verbs, themselves being the organising constituents of sentences, are of course excellently suited to such an approach. Frames are defined in terms of the semantic roles of the arguments and the modifications. The classic example is the “commercial event”, introduced in Fillmore (1976), a frame defining verbs like *buy, sell, let, rent, purchase, hawk, auction* and others. Necessary argument roles in this frame are a buyer, a seller and goods; an optional argument can be the means of paying; an optional modification could be a time period (in the cases of *let* and *rent*). Frames as described here are of course not mere verb categorisations. They are organisations of the vocabulary of a conceptual domain and as such the commercial event frame also contains nouns such as *buyer, seller, client, salesman, goods, price* and adjectives such as *cheap* and *expensive*. Alongside this organisation in terms of conceptual frames, yet another structuring principle is used: inheritance. So *sell* inherits parts of the frame it is used in from *give*, which means that *give* is categorised in a more general frame of which that of *sell* is a specification. The same kind of inheritance holds for *buy* and *get*. In fact one could say that these inheritance relations, on condition that they inherit the whole frame and not just part of it, are another way of presenting hyponym-hypernym relations. Notice, by the way, that also within the commercial event domain one can find some hyponym-hypernym relations. *Auction* and *hawk* name special ways of selling. Auctioning goods is selling them by presenting them at an auction and hawking is selling goods by presenting them door-to-door. Their semantic content consists of the meaning element ‘sell’ and a specific value for the modification type ‘means’. *Auction* and *hawk* inherit from *sell*. Taken independently, the two principles do not lead to the same categories. By the conceptual field principle, *buy* and *sell* are taken together in one category. The inheritance principle defines a semantic hierarchy *give* – *sell* – *auction*, but *buy* is excluded from that. There is no systematic or predictable interaction between the two principles.

A comparable double principle of semantic organisation can be found implicitly in Levin (1993). Levin bases her categorisation on a correspondence between verb meaning and some regularities in the syntactic behaviour of verbs, notably the regular patterns in argument alternation. Consistent application of this criterion in most cases yields categories that are also coherent from the semantic point of view. But in some cases it results in distinctions that can hardly be called semantic. So e.g. the alternation criterion leads Levin to discern a category ‘verbs with predicative complements’, which groups together two types of concepts that have nothing in common from a semantic point of view: ‘find’ as in *I find her stupid* and ‘make’, as in *They made him king*. The former in fact joins with her category of verbs of judgement, the latter with her verbs of change of state. Apart from that, the criterion sometimes also prevents her from keeping related verbs together. To indicate only two cases: *kill* and *destroy* are placed in two different categories (in fact they are category names) without any mutual reference; there is a (sub)category of sound emission but sounds made by animals are placed apart. It looks as though the alternation criterion most of the time correctly predicts semantic differences but is insufficient to relate categories that are semantically next of kin. Another apparent weakness is that the syntactic distinction criterion cannot be mapped onto one single distinctive principle in terms of semantics. It leads to a wealth of categories with different semantic motivations. So there is e.g. a category of ‘verbs involving the body’ that is semantically motivated by the belonging of its members to the same conceptual domain, in this case everything related to the body. A category like ‘verbs of emission’ on the other hand, would form, when thought of in terms of a conceptual domain, a very diffuse category, but the verbs belonging to it happen to share the same hypernym: *emit*, and thus, as is common for hyponyms, inherit the semantic properties of that verb. A consequence of this lack of a single organisation principle is that cases of double categorisation will be inevitable (and numerous). To confine ourselves to the examples just given: verbs like *sweat*, *bleed* and *breathe* are at the same time verbs involving the body and verbs of emission. All this also has its advantages. Levin offers a pretty complete catalogue of verb types, albeit in the form of a flat and unstructured list, and at the same time shows us implicitly, and maybe unintendedly, the different possible basic approaches to semantic categorisation of verbs, which happen to be pretty much the same as those ran across in *Framenet*.

Both the conceptual domain approach and the purely semantic categorisation in terms of hypernyms and hyponyms have their limitations. As for the latter, it is a well-known fact that this kind of paradigmatic relation with verbs is much less richly developed as for nouns. A notorious problem with conceptual domains is that there is a certain degree of arbitrariness in them. What conceptual domain does *shop* belong to? Commerce or leisure? Domains can serve as a good guide to direct the word searcher in a certain direction but there is no sound system in them: there are no objective criteria for any judgement about whether some specific slice of our experience should be a conceptual domain or not.

Some cases even seem to suggest that conceptual domain and hypernym yield complementary types of information that together constitute a category: there happen to be hypernyms that can serve for more than one conceptual domain and there are conceptual domains within which more than one hypernym can govern over the category members. This can be observed in the commercial event mentioned above. Both the verbs *buy* and *sell* belong to the conceptual domain, but they have different hypernyms: *get* in the first and *give* in the second case. But, to take just one of them, *give* has also hyponyms outside the realm of the commercial event: *pass* and *hand*, to name two. One can only find a homogeneous category in the intersection of commercial event words and hyponyms of *give* and *get*: a category *give with respect to the commercial event* and one *get with respect to the commercial event*. The former would contain verbs like *auction*, *let*, *retail* and *sell*, but not *hand* and *pass*, the latter *buy*, *purchase* and *rent*, but not *receive*.

4. A proposal based on definition analysis

One feature that, in the course of the INL definition analysis project, immediately drew attention as a basic difference between noun definitions and verb definitions was the presence of auxiliaries and modal operators in the latter. This is part of an explanation of why hyponym-hypernym chains occur to a much less pervasive extent with verbs than with nouns. Normal analytic noun definitions are of the form

Category name + modifications

and every noun defined in that way can in turn become a category name and take new modifications, thus constituting another level in the hyponym-hypernym chain. A verb definition, compared to this, takes as its most extended form

(Modal operator) + (auxiliary) + category name + modifications

and in every instance where one category is formed out of another by means of an auxiliary the chain of hyponyms is interrupted. *Doze* ('sleep lightly', a verbal category with a degree modification) is a hyponym of *sleep*, but *fall asleep* ('begin to sleep', a verbal category subjected to an inchoative auxiliary) isn't.

The most common auxiliaries in Dutch verb definitions are the aspectual ones (*beginnen* 'begin', *ophouden* 'stop', *doorgaan* 'continue' and the causal auxiliary *doen* 'cause, make'. Also *trachten* 'try' and *kunnen* 'be able' have been encountered a number of times in the definitions under analysis. Modal operators that have been found are the negative element *niet* 'not' and a set of modal adverbs, all indicating some lower or higher degree of possibility.

It is our guess that the definition structure formula above will enable us to organize the verbs in the same kind of hierarchic categorial structure as can be found with nouns. There is even clearly one basic predicate that we can start from: *zijn* 'be' in its use as a two-place verb, as '*be [in a position, situation, place, with a property etc.]*'. Semantically this verb is practically empty and it functions only as a bearer of syntactic-semantic information about tense, person, active/passive distinction etc. From this basic category one can start building other, derived categories, using operators, auxiliaries and modifications. Some brief illustrations to show how the system works: direct derivations from the basic category and a modification type that we will call, for the sake of convenience, *status*, are categories such as *op een plaats zijn* 'be in a place', *in een toestand zijn* 'be in a situation', *in het bezit zijn* 'be in the possession'. Verbs belonging to the category *op een plaats zijn* are e.g. Dutch *vertoeven*, *verblijven* (both 'stay'), *overwinteren* ('winter, hibernate'), *wonen* ('live, inhabit'); from this we can derive *beginnen op een plaats te zijn* (*komen* 'come, arrive'), *ophouden op een plaats te zijn* (*vertrekken* 'leave', *verdwijnen* 'disappear'), *doorgaan met op een plaats te zijn* (*blijven* 'stay' as opposed to 'leave'), *niet op een plaats zijn* (*ontbreken* 'be missing') and further on *doen beginnen op een plaats te zijn* (*plaatsen* 'place, put'), *doen ophouden op een plaats te zijn* (*wegnemen* 'take away'), *doen doorgaan met op een plaats te zijn* (*houden* 'keep') and so on. One should notice that operators and auxiliaries on the one hand and modifications on the other are objects of different kinds. Operators and auxiliaries form a limited set of abstract structuring elements; modifications are ways to give a predicate some place in the world: they are an in principle unlimited set of restrictions serving to link the abstract predication to some larger or narrower conceptual domain. One should also notice that this type of categorial structure offers a systematic means to discern different senses of a word.

There are also some types of categorial information that this structure is unable to express. For one thing, it is not able to make exact predictions about the selection restrictions applying to the members of a category. Verbs with different selection restrictions can belong to the same category. To some degree the selection restriction for a verb correlates with its basic semantic function: whether it denotes an action (subject: persons), an occupation (subject: persons), an event (subject: things or persons behaving as things), a process (subject: things or persons behaving as things) or a situation (subject: things or persons behaving as things). For everything beyond that, the selection restriction is verb-specific and should be incorporated in the definition of the single word, not of the category. But the degree to which selection restrictions can be predicted on a regular base should be incorporated in the category definition, just in order to respect the principle that regular things should be described on a regular level and particular things on a particular level. The conclusion from is that also Dik's categorisation or what we have made out of it should play a role in our own organising system in order to increase its defining power.

The system presented here also lacks explicit information about the conceptual domain to which a verb category is linked. So e.g. *change* (with respect to clothing) and *shave* clearly belong to different categories in the sense that we have just sketched, but they both belong to the conceptual domain of body care. This information may in principle be implicitly or explicitly present in the modifications that form part of the category definition, but there is nothing systematic in that. So also information on conceptual domains could better be overtly expressed in the categorising system.

Therefore we opt for a triple-layered categorisation structure. The deepest level shows the most abstract and general division into performing an action, having an occupation, being in a state or being involved in a process or an event (or leaving the choice between two or more of these)⁵. On the intermediate level the conceptual domain is specified. For this level at the present stage of the project the division by Levin has been adopted – with some adaptations – as a working hypothesis. The surface level contains the categories resulting from definition analysis as sketched in the present section. This triple-layered category structure runs parallel to a proposal concerning adjective categorisation made in Heyvaert (2010).

To make this a little more concrete with an (English) example: a verb like *shave* is categorised as

To shave
is to perform an action
is related to body care
is to cause to disappear (= stop to be in a place)
 Features: place: on one's face
 object: hairs
 instrument: with a razor

To avoid possible misunderstanding: what is shown here is not intended as a model for verb definitions to be presented to the common dictionary user. It is meant as an underlying structure for these definitions, invisible to the user but available for various retrieval operations like synonymy search and search of words starting from meaning data. In the ANW, the dictionary this proposal is meant for, it has even a visible place in the entry structure, in the so-called semagram, an extended and systematically organised meaning description that is presented side by side with the definition but is in the first place meant for linguists and specialists in other disciplines who want to make use of the dictionary for specialised purposes.

Notes

¹ With a thing as subject; with a human subject it would normally be +CONTROLLED.

² One may wonder what in fact is modified by these adjectives: the verb itself or the event or state that it denotes? The second answer seems to be the most plausible.

³ Let use of the word *bezigheid/occupation* here be some rhetorical trick by which the category is enabled to contain at the same time verbs of the *rest* type and *work* verbs.

⁴ Things are only goods in contexts of possession or trading.

⁵ There are a few verbs that fall beyond the reach of this categorization: some of the *verba sentiendi et declarandi* in one particular use as in *John thinks that circles are square*. These verbs denote no action, situation, activity or process. *Thinks* does not describe what John is doing: the idea that John is actually performing some thinking is completely irrelevant to the sentence meaning. *John thinks* here defines the condition under which the subordinate clause *circles are square* is valid. There seems to be one categorization level that is deeper than Dik's proposal. Dik categorises descriptive verbs, but apart from these there also appears to exist a small category of non-descriptive verbs that also must find its place in the verbal category structure. Our proposal is to discern a category of reflective verbs as opposed to descriptive verbs.

References

- Dik, S. C. 1978.** *Functional Grammar*. Amsterdam: North Holland Linguistic Series
- Fillmore, C. 1976.** 'Frame semantics and the nature of language.' In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*. Volume 280: 20–32.
- Heyvaert, F. 2006.** *Werkwoordelijke betekenisklassen en semantische kenmerken*. Internal project report, INL, Leiden.
- Heyvaert, F. 2010.** 'An outline for a semantic categorisation of adjectives.' In A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy / Afuk.
- Jackendoff, R. 1983.** *Semantics and Cognition*. Cambridge (Mass.): MIT Press.
- Levin, B. 1993.** *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Miller, G. A. and P. Johnson-Laird. 1976.** *Language and Perception*. Cambridge, Mass: The Belknap Press of Harvard University Press.
- Verkuyl, H. 1993.** *A theory of aspectuality: the interaction between temporal and atemporal structure*. Cambridge: Cambridge University Press.