Lineke Oppentocht, Institute for Dutch Lexicology

A Tool for the Lexical Semantic Classification of Dutch Verbs

Abstract

In this paper we will introduce a model for the lexical semantic classification of Dutch verbs. This model aims to contribute to the construction of definitions for monolingual dictionaries to make them more suitable for NLP-purposes. It has the form of a flow chart in which each level contains a further refinement of the distinction in semantic groups at the level preceding. The model can be passed through by answering yes or no to predefined questions. It has been implemented in a computer program which, under strict guidelines, guides the user through the model.

1. Introduction

Because human language is so complex and full of information, computer programs to process it (NLP programs) have to dispose of enormous amounts of varied linguistic data (speech, text, lexicons and grammars). It is virtually impracticable for computational linguists to build such large computational lexicons by hand. Therefore, attempts are made to do this automatically, or semi-automatically by using on-line resources such as machine readable dictionaries (MRDs). In practice however, for reasons exposed in numerous articles (e.g. Levin 1991, Boguraev 1991, Atkins et al. 1986, 1988, Atkins 1991, Klavans 1988 and Alshawi et al. 1989) it often appears to be difficult to take full advantage of the information contained in MRDs. Most problems are related to the fact that a lot of the information in dictionaries is treated neither consistently, nor systematically. Consequently, we do not get a clear picture of the structure of the lexicon. However, to correctly process language, a computer needs explicit information on the structure of the lexicon. Contrary to human users of a dictionary, a computer does not have implicit knowledge of this structure. The model proposed in this paper will hopefully contribute to a better construction of definitions of Dutch verbs so as to be more suitable for NLP-programs, by having them treat semantic information more consistently. The model allows the lexicographer to systematically group semantically similar verbs. It then proposes the semantic information to be treated in the definitions of all verb meanings belonging to a particular semantic group. This consistent

treatment of information should make it easier for NLP-programs to recognize which verb meanings are semantically similar. Our aim is not to build an ideal lexicon for NLP. Such a lexicon is very different from conventional published dictionaries (Klavans 1988). The model merely serves to remind a lexicographer of the basic semantic information about a particular Dutch verb that ought to be treated in a definition that is at the same time human-friendly as well as more suitable for NLP-purposes.

2. Empirical evidence and theoretical framework

The model is based upon the analysis of the meanings of 1485 verbs selected from the *Groot woordenboek hedendaags Nederlands* (Van Dale 1991). All those verbs have either both a transitive and an intransitive meaning or a transitive and a reflexive meaning (Oppentocht 1994:129). The electronic textual corpus of 50 million words which is on-line accessible at the Institute for Dutch Lexicology, has been used to retrieve contextual information on these verbs. The model is entirely based on evidence from the concordances of the 1485 Van Dale verbs in that particular corpus. It is designed to classify verb *meanings*. One verb could eventually be classified in different groups according to its different meanings. As has been described in Oppentocht (1994) the underlying theory is inspired by Simon Dik's *Functional Grammar* (Dik 1989) and by the work of Guy Deville (Deville 1989).

3. The working of the model

The model has the form of a flow chart. We distinguish three levels: the upper level on which verb meanings are classified in States of Affairs, the middle level on which the States of Affairs are semantically more refined, and the bottom level on which the verb meanings are classified according to the number and nature of the obligatory arguments surrounding them. Each of these levels is composed of subpaths which have to be followed by answering fixed questions. This structure has been implemented in a computer program. The program asks the lexicographer those fixed questions and then it classifies a verb meaning in the correct semantic group by combining the given answers. The computer program is accompanied by strict written guidelines and help screens.

3.1 The upper level: classification in States of Affairs

A verb meaning can refer to four kinds of States of Affairs (SoA): an action, a state, a process or a position. The choice among those four possibilities can be made by determining what value the verb meaning has for the features control and dynamism. The first questions asked by the computer program are related to these features. They have to be answered by using a help screen and strict written guidelines. Actions are +control and +dynamic, processes are -control and +dynamic, states are - control and -dynamic and positions are +control and -dynamic. Figure 1 represents this part of the flow chart.

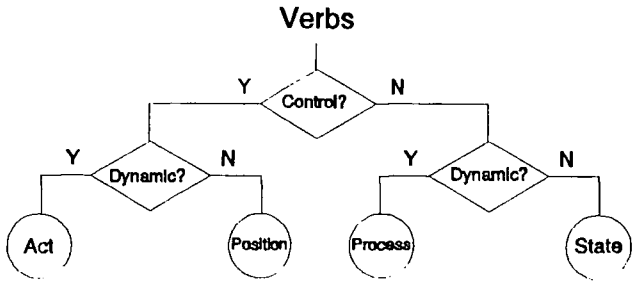

Figure 1: Classification in States of Affairs

A verb meaning has the feature +control when its subject argument (in active sentences) refers to an animate or likely animate entity which has the power to determine whether or not the event or situation referred to by the verb will take place. The question asked by the computer program is: "Is the entity being the subject of the sentence in control of the event?" Likely animate entities are usually metonyms that stand for a collective of animate entities. For example, *this institute* in *This institute takes part in the conversation* is a controlling entity. However, a (likely) animate entity functioning as a verb's subject argument, is not necessarily a controller. Therefore, we use two additional tests. For a verb to be labelled as [+control], all tests have to be affirmative. Firstly, when a verb's subject entity is a controller, the sentence can be put in the imperative.[1] The sentence resulting from this needs to be both grammatically and semantically acceptable. Secondly, words like *opzettelijk* ('deliberately') and *vrijwillig* ('voluntarily'), indicating a conscient attitude, can be added to such sentences. For example, *de kinderen* ('the children') in *De kinderen balen omdat het regent* ('The

children are fed up with the rain') are not controlling entities. When we put this sentence in the imperative the result is semantically strange (**Kinderen, baal omdat het regent!*: *'Children, be fed up with the rain'). Similarly, the sentence becomes semantically strange when we add a word like *opzettelijk* or *vrijwillig* to it (**De kinderen balen opzettelijk/ vrijwillig omdat het regent*: *'The children are deliberately/voluntarily fed up with the rain').

A verb meaning has the value [+dynamic] if one of its arguments refers to an entity which changes because of the event referred to by the verb. This can be a spatial change, a change of one or all of its properties, or a change in the (degree of) activity of this entity. Non-dynamic States of Affairs (states and positions) can be divided into two groups. Either the verb refers to a static State of Affairs ('John sits in his chair'), or it refers to a State of Affairs which is still in progress ('John talks/laughs'). During this progress however, nothing changes.

According to these definitions of the features control and dynamism, we classify *opknappen* ('restore') in *Mijn broer knapt de schuur op* ('My brother restores the barn': X restores Y) as an action (+control, +dynamic). My brother is the controller and the properties of the barn change. *Remmen* ('slow down') in *De zuren remmen de ontwikkeling van de cellen* ('The acids slow down the development of the cells': X slows down Y) will be classified as a process (-control, +dynamic). The acids do not control the event but the development of the cells does change (its activity decreases). *Betogen* ('argue') in *De auteurs betogen dat er sprake is van onrecht* ('The authors argue that it is a matter of injustice': X argues Y) will be classified as a position (+control, -dynamic). The authors control the arguing but neither the authors nor the injustice change. *Hangen* ('hang') in *De was hangt aan de lijn* ('The laundry hangs on the line': X hangs on Y) will be classified as a state (-control, -dynamic). The laundry does not control the situation and neither the laundry nor the line change.

3.2 The middle level: refinement of the States of Affairs

The next step in the development of our model consists of the lexical semantic refinement of the classes action, state, process and position. These classes are too coarse-grained to adequately describe the basic semantic differences between our 1485 verbs. Therefore we have further divided those classes into subgroups. These subgroups correspond to semantic domains (see Fellbaum 1990 and Levin 1991, 1993 for this term). The classification of both action and process verb meanings has

been refined by analyzing the kind of change occurring to the entities referred to by a verb's arguments (change of properties, spatial change or change in (degree of) activity).[2] The classification of states and positions has been refined by analyzing the exact nature of the situation referred to in the concordances (does it refer to a state of mind, a location, a possession etc.). In what follows we will elaborate on one part of the flow chart, namely the class of actions. Every now and then, however, we will refer to similarities and differences we encounter with the other classes (processes, positions and states). Figure 2 represents the middle level of the flow chart of action verbs. This flow chart has to be considered as the continuation of the act circle in figure 1. Depending on whether the entities referred to by the verb's arguments change as to properties, place, or activity, a different path (tree) will be followed. These paths consist of fixed questions (represented in brief in the diamonds of figure 2) which, while using the computer program, have to be answered by using a help screen and very strict written guidelines. The middle level of the flow chart of processes looks exactly the same. That of states and of processes is however very different. They have a much flatter structure than that of actions and processes. In general they do not go any deeper than the level at which in the flow chart of actions and processes it is decided whether there is a change as to properties, place or activities.

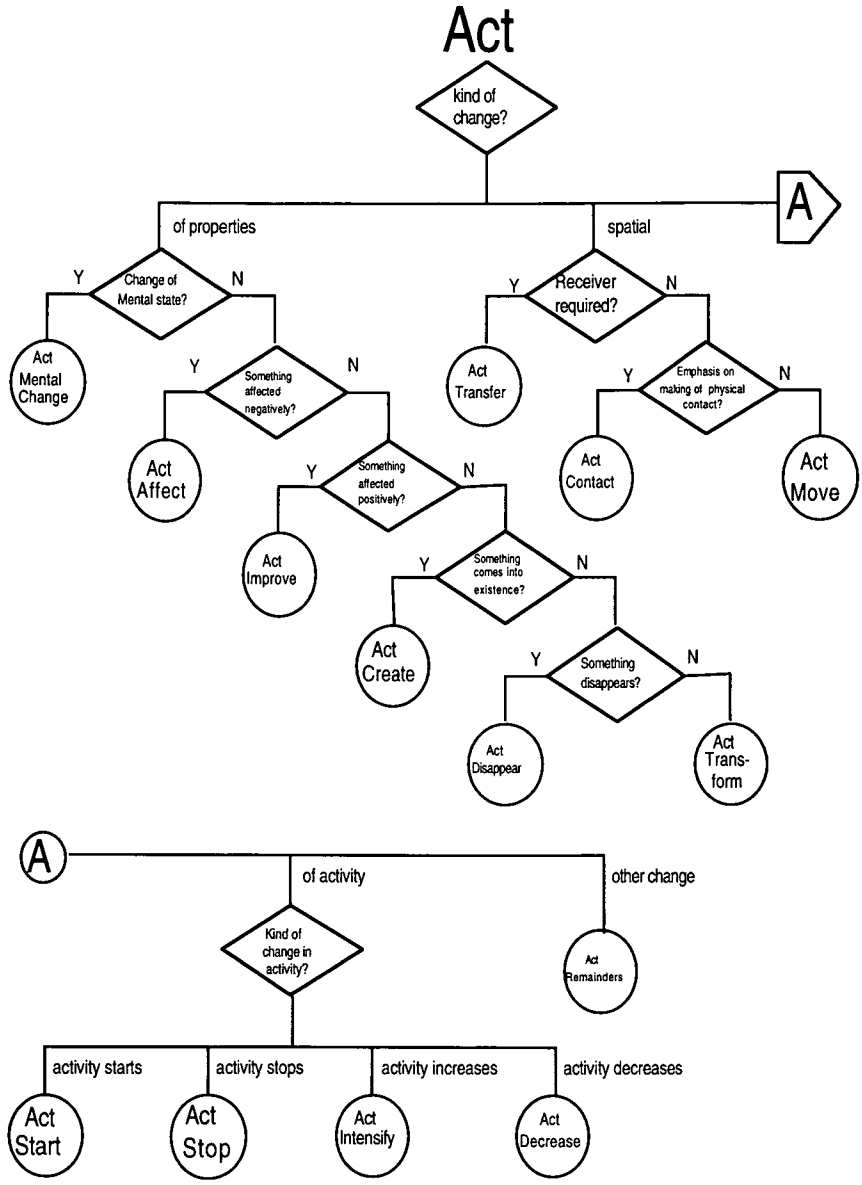

Figure 2: Middle level. Refinement of the States of Affairs

We will give one example of each of the three subtrees in the flow chart of actions (change as to properties, place and (degree of) activity):

• **ACT-improve**: change of properties. Action verbs (in a particular concordance) belong to this group when in the State of Affairs described

by the verb meaning, one or more entities referred to by the arguments of the verb are affected positively. This means that one or more properties of this/these entity/ies become(s) nicer, more beautiful or qualitatively better. An example is *opknappen* ('restore') in *Mijn broer knapt de schuur op* ('My brother restores the barn'). Here the properties of the barn are affected positively by my brother.

• **ACT-transfer**: spatial change. Action verbs (in a particular concordance) belong to this group when in the State of Affairs described by the verb meaning (a transfer of something) there is a receiver. According to our definition this receiver can only be animate. He or she is not necessarily mentioned in the sentence but is necessarily there in reality. An example is *toesteken* ('hand') in *De douanebeambte steekt ons de paspoorten toe* ('the customs officer hands us the passports'). In this sentence the passports change place. They are given by someone to an animate receiver (us).

• **ACT-decrease**: change of activity. Verbs (in a particular concordance) belonging to this group indicate that the intensity of the activity of one or more entities or events referred to by its arguments, decreases. We are not talking about the quality of what the entity(ies) is/are doing here (ACT-improve). The only thing that counts is a reduction of the speed, frequency or pace of the activity carried out by the entity/ies, or of an event referred to by an argument. An example is *vertragen* ('slow down') in *Hij vertraagt het onderzoek* ('He slows down the the investigation'). Here someone causes the pace of the investigation to slow down.

3.3 The bottom level: number and nature of obligatory arguments

A verb is obligatorily combined with one, two, or three arguments.[3] For NLP-programs (but for human users of a dictionary as well) it is important to know the number of obligatory arguments as well as their nature. Unfortunately, in most conventional dictionaries this information is not systematically given. We use two tests to determine whether an argument is obligatory. For an argument to be obligatory, one of these tests should be affirmative. Firstly, an argument is obligatory when the sentence in which it occurs with the verb becomes ungrammatical when the argument is left out. For example, in Dutch, in *Drie mannen rennen de heuvel af* ('Three men run down the hill') both *mannen* ('men') and *heuvel* ('hill') are obligatory. Neither **Rennen de heuvel af*, nor *Drie mannen rennen af* are grammatically correct sentences. This test usually works well, although there are instances in which people judge differently as to grammaticality. This especially happens when a verb

meaning does not belong to our every day vocabulary. The second test says that when the meaning of a sentence changes when an argument is left out, this argument is obligatory. For example, in *De monteur beweegt de ruitewissers heen en weer* ('The mechanic moves the windshield wipers up and down'), both *monteur* ('mechanic') and *ruitewissers* ('windshield wipers') are obligatory. When we leave out the argument *ruitewissers* ('windshield wipers'), the sentence becomes *De monteur beweegt heen en weer* ('The mechanic moves up and down'). This is a grammatical sentence but its meaning differs from that of the first sentence. We will give one example of a flow chart at this level, namely the one that is the continuation of the ACT-improve circle in figure 2:

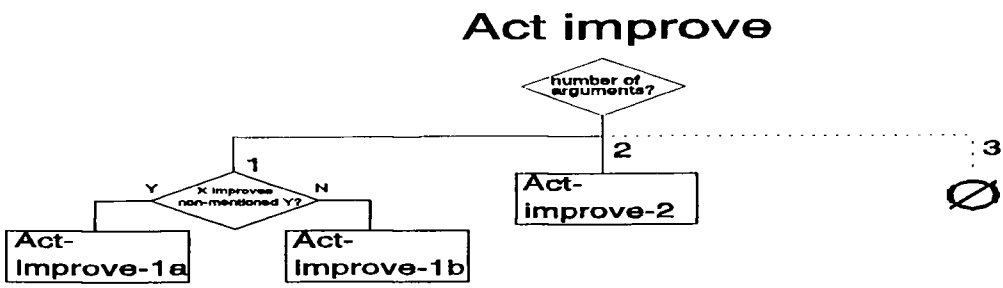

Figure 3: The bottom level.

According to the concordances of the verbs belonging to the ACT-improve group, those verbs can be combined with one or two obligatory arguments, depending on their meaning. We have not found ACT-improve verbs occurring with an obligatory third argument. When an ACT-improve verb meaning is combined with only one obligatory argument, an additional question must be answered, namely whether the subject entity in active sentences (X) in the reality referred to, improves the properties of another entity or not. This additional question serves to determine the perspective from which the State of Affairs is seen: the subject entity (X) can either be the causer of a change of properties of another entity (Y) (not mentioned in the sentence but derivable from the surrounding context), or be the causer of a change of properties of him/herself. This distinction does not have to be made when there are two obligatory arguments with an ACT-improve verb meaning. In the latter case the subject entity (X) always is the causer of a change of properties of another entity (Y). When this bottomlevel flow chart has been passed through, the computer program places the verb meaning in a final verb

class (represented by the squares in figure 3) by combining all the answers given at the top level, middle level and bottomlevel. We will now give an example of each of the three final ACT-improve classes:

• **ACT-improve 1a**: *mesten* ('fertilize') in *We moeten nog mesten* ('We still have to fertilize'). Here the quality of a non-mentioned entity (probably land, soil) is improved by a group of controllers (we).
• **ACT-improve 1b**: *afslanken* ('slim down') in *Zij is aan het afslanken* ('She is trying to slim down'). A controlling entity (she) causes her own appearance to improve.
• **ACT-improve 2**: *opknappen* ('restore') in *Hij knapt de schuur op* ('He restores the barn'). A controlling entity (he) causes the improvement of the barn. This second entity is mentioned in the sentence, as opposed to what is the case for ACT-improve 1a verb meanings.

4. Argument roles and template definitions

To each of the final classes (represented by the squares in the flow charts at the bottom level), a case frame corresponds which shows the semantic role of the obligatory arguments that should be represented in the definition, and a template definition which very succinctly describes the basic meaning of all verbs belonging to a particular final group. Once a verb has been classified in a final class at the bottomlevel of the flow chart, this information is shown on the lexicographer's screen. For example:

• **ACT-improve 1a**. *Case frame*: [Agent] (a controller which causes a change). *Template definition*: Een of meer levende of daarop gelijkende wezens verbeteren een of meer eigenschappen van een niet genoemde entiteit ('One or more (likely) animate entities improve one or more properties of a non-mentioned entity').
• **ACT-improve 1b**. *Case frame*: [Agent experiencer] (a controller which causes a change he experiences himself). *Template definition*: Een of meer levende of daarop gelijkende wezens verbeteren een of meer eigenschappen van zichzelf ('One or more (likely) animate entities improve one or more properties of themselves').
• **ACT-improve 2**. *Case frame*: [Agent, Patient] (a Patient undergoes a change caused by a controller). *Template definition*: Een of meer levende of daarop gelijkende wezens verbeteren een of meer eigenschappen van een of meer andere genoemde entiteiten. ('One or more (likely) animate

entities improve one or more properties of (an) other mentioned entity ies').

The case frame helps to identify the lexical semantic differences between the obligatory arguments of verbs belonging to different final groups in a particular bottom level flow chart (in our example between those of ACT-improve 1a verbs, ACT-improve 1b verbs, and ACT-improve 2 verbs). The template definitions are not meant to be put in a dictionary in their exact wording. They are to be considered as a guideline for the lexicographer who wants to treat semantically similar words in a consistent and systematic way. By using the classification model it is made explicit which verb meanings are semantically similar (according to theory of the organization of the lexicon which is at the basis of the model). The template definition then tells what is the basic meaning all those verbs have in common. The lexicographer can now make sure that all verbs with the same basic meaning according to the model, get a definition in which this basic meaning is represented by the same wording. This basic definition can be complemented afterwards with specific information indicating how verbs with a similar basic meaning differ from each other. The basic definition then establishes an explicit link between semantically similar verbs. In this way the computer can get a better insight into the structure of the lexicon than it would using definitions from most conventional dictionaries. For example, both *opknappen* ('restore') in *Hij knapt de schuur op* ('He restores the barn') and *versoepelen* ('relax') in *Hij versoepelt de contacten met Afrika* ('He causes the contacts with Africa to be more relaxed') belong to the ACT-improve 2 class according to the model. However, in the wording of the definitions of those two verbs in the *Groot woordenboek hedendaags Nederlands* (van Dale 1991) there is nothing telling the computer that there is a connection between those verbs. *Opknappen* in this sense is defined as 'netter, beter maken' ('tidy up, make better'). *Versoepelen* is defined as 'soepeler maken' ('render more flexible'). When taking into account the template definition of ACT-improve 2 verbs, the lexicographer could begin describing both verbs as, for example 'Iemand verbetert het genoemde' ('Someone improves the mentioned entity'). The basic definition of *opknappen* could then be completed with specific information like 'door het netter te maken' ('by tidying it up'). The basic definition of *versoepelen* could be completed with 'door het soepeler te maken' ('by rendering it more flexible').[4]

5. Concluding remarks

Our model for verb classification has been designed as a tool for lexicographers to help them treat semantic information on verbs in a consistent and systematic way, necessary to enable the computer to get insight into the lexical structure of verbs when using the dictionary. Writing this paper we are in the middle of testing the model. Different testees are asked to classify the verbs in a number of sentences by using the verb classification program and the strict written guidelines. The question is whether a classification model designed by one person is objective enough to be used by other persons. We hope to present the results of this test at the '96 Euralex congress.

Notes

1. This test only applies to positive sentences. For example, the sentence **Children be fed up with the rain* is strange, but the sentence *Children don't be fed up with the rain (because I know something we can do...)* is more acceptable.
2. These are the kinds of changes that were most frequently found in the INL corpus. There are act and process verbs which cannot be classified in one of those groups. We have decided to focus our attention first on the large and frequent groups. Verbs that could not be classified in those large groups have preliminarily been put aside (cf. the remainders circle in figure 2). The same method has been used for state and position verbs.
3. The concordances in the INL corpus show that the state and position verbs (verb X in a particular meaning) among the 1485 verbs from our corpus only very rarely occur with an obligatory third argument whilst this is not at all exceptional for action and process verbs.
4. Another possibility for dictionaries compiled by computer is to give the template definitions and data on particular verb classes automatically in a separate field in the database, next to the more classic definitions composed by the lexicographer.

References

Atkins, B.T.S., Kegl, J. & Levin, B. 1986. "Explicit and implicit information in dictionaries". *Advances in Lexicology.* Proceedings of the second annual conference of the UW centre for the new Oxford English Dictionary. Waterloo, Canada. 185–203.

Atkins, B.T.S., Kegl, J. & Levin, B. 1988. "Anatomy of a verb entry". *International journal of lexicography.* Vol. 1 no 2. 84–126.

Atkins, B.T.S. 1991. "Building a lexicon, the contribution of lexicography". *International journal of lexicography.* Vol. 4 no 3. 167–

303.

Alshawi, H., Boguraev, B., Carter, D. 1989. "Placing the dictionary on-line". In: B. Boguraev & T. Briscoe (eds.): *Computational lexicography for Natural Language Processing*. London/New York: Longman. 41–63.

Boguraev, B. 1991. "Building a lexicon, the contribution of computers". *International journal of lexicography*. Vol. 4 no 3. 227–259.

Deville, G. 1989. *Modelization of task-oriented utterances in a man-machine dialogue system*. PhD. thesis. University of Antwerpen.

Dik, S. 1989. *Functional Grammar. Part I: the structure of the clause*. Dordrecht: Reidel.

Fellbaum, C. 1990. "English verbs as a semantic net". *International journal of lexicography*. Vol. 3 no 4. 278–299.

Klavans, J.L. 1988. "Building a computational lexicon using Machine Readable Dictionaries". In: T. Magay & J. Zigany (eds.): *BudaLEX '88 Proceedings*. Akademia Kiado, Budapest. 265–279.

Levin, B. 1991. "Building a lexicon, the contribution of linguistics". *International journal of lexicography*. Vol. 4 no 3. 205–226.

Levin, B. 1993. *English verb classes and alternations. A preliminary investigation*. The university of Chicago press.

Oppentocht, A.L. 1994. "Towards a lexical semantic model for the creation of NLP and human-friendly definitions". In: W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg & P. Vossen (eds.): *EURALEX 1994. Proceedings*. Amsterdam.

Van Dale 1991. *Groot woordenboek hedendaags Nederlands*. 2nd edition. Van Dale Lexicografie, Utrecht/Antwerpen.