Lineke Oppentocht
Institute for Dutch Lexicology

Towards a Lexical Semantic Model for the Creation of NLP and Human–Friendly Definitions

Abstract

The purpose of this paper is to introduce a tentative formal model for definition generation. The model is based on a semantic subclassification of Dutch verbs. The generation process should result in Dutch verbal definitions which are suitable for Natural Language Processing and at the same time workable for human users. The main question is which semantic information must minimally be represented in such definitions.

1. Introduction

For the sake of a greater efficiency in the making of definitions, the gap between people creating definitions for NLP and those making them for human users, should be bridged. For this purpose, a new kind of definition has to be developed which can satisfy both parties. These definitions should be theoretically sound, easy to handle for human users and at the same time suitable for automatic processing. Definitions of this kind do not yet exist. Until now, some NLP–systems have been working with definitions of the kind we find in traditional dictionaries. This involves many problems (Alshawi 1989). Other NLP–systems work with formalized computational lexicons. However, these lexicons are very small. To gain more insight in the problem of what more convenient definitions should look like, the question was raised which semantic information about Dutch verbs needs to be represented formally in such definitions. To find this out, we needed a better understanding of the global organization of verbs in the lexicon. Therefore, we have started a lexical semantic description of a selection of Dutch verbs based on case–oriented grammars (especially Functional Grammar (FG)).

2. The corpus

The selected verbs (1429 in total) have been extracted automatically on the basis of grammatical codes from the *Van Dale Groot woordenboek hedendaags Nederlands* (Van Dale 1991) which is available on tape. So, although we study the semantics of verbs, we compiled our corpus on the basis of grammatical information. This has been done because, in contrast with LDOCE, there are no codes in the Van Dale dictionary for the semantic properties of entries. Besides, it would be impossible to make a selection of

verbs on the basis of fixed strings occurring in the definitions. These have not been used consistently.

As a first approach, we study verbs which have either intransitive and reflexive meanings (e.g. *amuseren* ('amuse')) or transitive and intransitive meanings (e.g. *afbreken* ('break off')). To provide contextual information, i.e. information about valency, the electronic textual corpus of 55 million words has been used. This corpus is online accessible at the Institute for Dutch Lexicology (INL). The INL–retrieval program delivered the concordances of the selected verbs.

3. Theoretical background

In this study we assume that a minimal set of semantic features can be distinguished by which all verbs from our corpus can be described. The framework of case–oriented grammars was judged most appropriate to elaborate the subclassification system. Fillmore's *case grammar* (Fillmore 1968) serves as a basis to the extent that it is used in Simon Dik's *Functional Grammar* (1978) and in the works of Chafe (1970), Vester (1983), and De Groot (1983).

Some computational models for Natural Language Processing using Functional Grammar have already been developed. Recent work was done by S. Dik and P. Kahrel (1992) and by Martin, Demeersseman and Vliegen (1992). However, the former study is based on a very small corpus (52 words of which 20 verbs). The latter one (the SNIV–project) focusses on grammatical subcategorisation rather than on semantic subcategorisation. Only a small part of its lexicon is also provided with semantic information. Another problem concerning the SNIV–project is caused by the character of its corpus. The macrostructure (the set of lexemes about which information is given) selected for the SNIV–project, is taken from the Van Dale *Basiswoordenboek van de Nederlandse taal* (1987). The basis for the microstructure is taken from a large collection of dictionaries. These sources do not provide explicit information about valency the way a textual corpus does. Therefore it is difficult to make an objective subcategorisation by means of this corpus.

The thesis of Guy Deville (1989), founded on Dik's *Functional Grammar,* has influenced our research to a large extent. His work also forms the basis for the semantic part of the SNIV–project.

For the formal representation of semantic relations, I. Mel'chuk's *Lexical Functions* (Mel'chuk 1988) have been used. The Lexical Functions have been used in a slightly different way than Mel'chuk does. They define the semantic properties of particular States of Affairs and represent the semantic relations between verb meanings and concepts; they do not indicate the lexical co–occurrence of a particular lexeme.

As in Functional Grammar, we work with the notion *predication.* This is a *predicate* (a verb in our case but nouns, adjectives and adverbs are called

predicates as well) plus *terms* functioning as arguments of the predicate. The arguments are specified by a *case (Agent, Theme, Benefactive...).*

Predications refer to *States of Affairs* (SoA) *(acts, states, processes* and *positions).* With respect to Deville's theory we do not adopt his set of SoA's (SwC's: *Sublanguage world Concepts* in his terminology). However, we do adopt the idea that a predicate is derived from a finite set of *predicate primitives* (in Deville: MVMT–1, LOCATION–2, PR–COGNT–2 etc.). These are defined in terms of binary *primitive features.* Deville distinguishes three types: *typological primitive features (dynamic* and *control), semantic primitive features (attributive, spatial* and *cognitive)* and *valency primitive features (transitive* and *ditransitive).*

However, instead of *dynamic* we use the term *Change*[1] for one of the typological primitive features. In order to avoid terminological confusion, we prefer the terms *mono, di* and *tri* (Martin, Demeersseman, Vliegen 1992) to Deville's *intransitive, transitive* and *ditransitive* for the valency primitive features. The reason for this is that the valency primitive features are not used to indicate whether a verb occurs with a subject only or with a subject and an object (traditionnally referred to by the terms intransitive and transitive). They indicate the number of arguments occurring obligatory with a verb, no matter what their syntactic function is. Finally, our set of semantic primitive features is different from the one in Deville's theory (cf. Section 4).

Every *predicate primitive* is combined with a specific set of *central cases. Central* cases *(arguments* in FG) and *peripheral* cases *(satellites* in FG) *cases* are distinguished according to their necessity in the structure of the *predication.*

Much as our work is influenced by Deville, we did make some alterations in his model. Firstly, in Deville's theory, the semantic primitive features simply combine with the typological primitive features. They are not derived from them. However, in our theory, the semantic primitive features are subclasses from the typological primitive features. By not representing the primitive features at one and the same level, we adopt Chafe's point of view on verb classification (Chafe 1970). Furthermore, we do not agree with the point of view that the semantic primitive features are mutually exclusive. It can, for example, very well happen that a SoA is defined as attributive and spatial at the same time.

Another important difference with Deville is that we work with verbs and *concepts* not with verbs only. More specifically: we assume that the meaning of all verbs refer to a concept. These verb meanings can be defined in the same way as the concept to which they refer. We chose to work with concepts because we aim at a subclassification of verbs in as less groups as possible. The denominator of each of these groups must therefore be as general as possible in order to cover as much verb meanings as possible. It is very unlikely that all of these general denominators can be found in a limited corpus as ours. Another advantage of working with concepts is that our final model for definition generation can also be applied to verbs from *outside* our

corpus. This is possible when the meaning of these verbs refer to a concept also referred to by the meaning of verbs *in* the corpus. This gives the model a broader scope.

4. Semantic subclassification of verbs as a tool for definition making: a formal model

In this paragraph we describe the process of definition generation, using our own model. We want to stress that this is only a tentative model. As our research is still in the early stages, we have not had the occasion to test all the statements made below in detail yet. Besides, we have only analysed about 150 verbs from our corpus at the moment. Consequently, the conclusions drawn below are based on a very small part of our corpus.

The whole process is carried out in eight steps. The tentative model is represented formally in Figure 1. The steps 1 to 8 described in this paragraph correspond to the numbers 1 to 8 in Figure 1. For illustration the verbs *fokken* ('breed'), *indeuken* ('dent'), *mummificeren* ('mummify') and *schilderen* ('paint') from our corpus are defined. These examples only illustrate a part of the complete model represented in Figure 1. That is, they are examples of *actions*. We do not give examples of *positions, processes* and *states* in this paper. However, verbs referring to the latter States of Affairs can be analyzed in a way analogous to action verbs. Just replace *action* by *position, process* or *state* in the formulas given in Figure 1.

The final definitions are only valid for one of the senses of the example verbs. The verb *schilderen* for example is taken from sentences like *een schilderij schilderen* ('paint a picture') and not from e.g. *het huis schilderen* ('paint the house'). In sentences like this one, the verb would be marked with different primitive features and therefore its final definition would be different.

According to their meaning in the concordances, the verbs from our corpus are classified in very general semantic groups. In our first step the typological and semantic primitive features are determined. The *typological primitive features* (ctrl and ch) determine to what kind of SoA the concept refers: an *action* (+ctrl +ch: 'Mary hit the dog'), *process* (–ctrl, +ch: 'Peter falls from the tree'), *state* (–ctrl –ch: 'the chair stands in the corner') or *position* (+ctrl –ch: 'Peter sits in the tree').

A SoA is *controlled* when one of the entities playing a role in the SoA is able to start, stop or continue the SoA. This entity, the *controller,* is not exclusively associated with animacy as is done in FG as well as in Deville (1989). There are three options:

(1) The controller is animate and is then called *agent.*
(2) The controller is not animate but is an entity which has a force of its own, which enables it to perform certain actions (heat, wind...). The controller is then called *force* (Dik 1978). Dik acknowledges that

entities like 'wind' can be presented as an autonomous cause or instigator of a process (Dik 1987: 37), but in his opinion such entities are non–controlled.

(3) The controller is not animate and does not have a force of its own. It is often used in a metonymical sense. E.g. in 'the ship enters the harbour' it is obvious that there is some non–mentioned animate controller which actually makes the ship enter. The ship is now called a *potent* controller in the sense that it is a controlled (but not controlling) causer of the SoA. The term *potent* is borrowed from Chafe (1970) who uses this term for entities called *force* by Dik.

It will be clear that it can only be determined whether a SoA is controlled by looking at the function of the arguments in the context of the verb referring to the SoA. Therefore information about valency is indispensable for a semantic classification.

A SoA has the feature *Change* if it describes any alteration whatsoever (Vester 1983). We do not limit change to movements only, as is done in De Groot (1983). *Fokken, indeuken, mummificeren* and *schilderen* all are actions (+ctrl +ch) according to their (transitive) meaning in the concordances, so these verbs are symbolised by x in the very first formula of *action* in figure 1 (f_1 = ctrl, f_2 = ch). X_c means that we only talk about verbs from our corpus at this level.

Subsequently, the *semantic primitive features* are determined. Being subclasses of the typological primitives, they are found by asking the questions 'what is controlled?' and 'what changes?' respectively. *Fokken, indeuken, mummificeren* and *schilderen* refer in some of their meanings to a controlled *creation* and a change of *existence.* In another context the semantic primitives assigned to these verbs would be different. *Schilderen* in *het huis schilderen* for example, refers to a controlled *transformation* and a change of *appearance.* To represent formally the subgroups of Control and Change, typological and semantic primitive features are combined in a formula inspired by the form of Mel'chuks Lexical Functions (f(X) = Y), namely: ctrl(creation) ch(existence) = *fokken ... schilderen. Creation* and *existence* replace Sgr^1_i and Sgr^2_j respectively in Figure 1. Other semantic primitive features are: *transfer, movement, place, finality...*

In our second step the *valency primitive features* (mono, di, tri) are determined. That is, by means of the concordances of *fokken, indeuken, mummificeren* and *schilderen* we determine with how many obligatory arguments these verbs are combined (one, two or three). In our example, only those concordances of *fokken, indeuken, mummificeren* and *schilderen* are considered in which these verbs are used in the sense ctrl(creation) ch(existence). Then, semantic functions *(cases)* are assigned to the arguments. In our example there are two arguments; an Agent and a Theme. These cases are combined in a *case frame.* For every State of Affairs there is

a particular case frame, depending on the semantic primitive features filled in for *i* and *j*.

Once all the primitive features determined, the *predicate primitive* can be derived since every predicate primitive has its own combination of typological, semantic and valency features plus case frame. Predicate primitives have not been integrated in the model as yet.

We have the impression that Dutch verbs from outside our corpus can be described in the same way as the verbs from inside this corpus. This is alleged in step 3. Verbs from outside the corpus are in this step categorized according to the method described in step 1 and 2, originally designed for verbs inside the corpus.

In step 4 a *concept* (notated in capitals) is chosen to which the verbs described in step one and two refer. This concept can also be referred to by verbs with the same semantic properties outside the corpus (described in step 3). The concept chosen for *fokken, indeuken, mummificeren* and *schilderen* is CREËREN (CREATE).

Fifthly, a *definition* of the concept is composed. The definition we made for CREËREN, covering all the features distinguished in step one and two, is: *'if X, Y CREËERT, X creates an Y which did not exist before'*.

In our sixth step, Lexical Functions inspired by Mel'chuk relate the concept to verb meanings referring to that concept. *Fokken, indeuken, mummificeren* and *schilderen* are a function of the concept CREËREN. *Fokken* and *schilderen* are verbs indicating (in one of their senses) a way of CREËREN: V_{mod} (CREËREN) = fokken, schilderen (V=verb, mod = way of). *Indeuken* and *mummificeren* refer to the result of CREËREN: V_{res} (CREËREN) = indeuken, mummificeren (a dent *(deuk)* and a mummy are created). Other LF's at this level are V_{loc}, V_{temp}, V_{caus}...

In step 7 is claimed that when a verb from our corpus is the function of the concept to which this verb refers, it gets the same definition as this concept. So, *fokken, indeuken, mummificeren* and *schilderen* are defined as: *'if X, Y fokt ... schildert, X creates an Y which did not exist before'*.

Finally, in step 8 is alleged that this very definition is also valid for verbs outside the corpus whose meaning refer to this concept.

Figure 1: A lexical semantic model for definition making

in which:

X = set of all Dutch verbs. x = one particular verb.
X_C = set of all verbs from our corpus. $X_C \supset X$
f_1 = typological primitive feature Control.
f_2 = typological primitive feature Change.
Sgr^1_i = semantic primitive feature; subgroup of Control. i = 1...n.
Sgr^2_j = semantic primitive feature; subgroup of Change. j = 1...m.
G = function which symbolises the selection of a concept.
Y = set of all concepts. y = one particular concept from Y.
Y_C = set of all concepts used as hypernym for the verbs from our corpus.
y_c = one particular concept from Y_C.
F = set of Lexical Functions: V_{mod}, V_{res} ... n.
f = one particular Lexical Function from F.

1
$$X_{Cij}^{action} = \{x \in X_C \mid x = f_1(Sgr^1_i) \wedge x = f_2(Sgr^2_j)\}$$
$$X_{Cij}^{position} = \{x \in X_C \mid x = f_1(Sgr^1_i) \wedge x \neq f_2(Sgr^2_j)\}$$
$$X_{Cij}^{process} = \{x \in X_C \mid x \neq f_1(Sgr^1_i) \wedge x = f_2(Sgr^2_j)\}$$
$$X_{Cij}^{state} = \{x \in X_C \mid x \neq f_1(Sgr^1_i) \wedge x \neq f_2(Sgr^2_j)\}$$

2
X_{Cij}^{action} is combined with [case frame $\alpha_{1...\alpha n}$]
$X_{Cij}^{position}$ is combined with [case frame $\beta_{1...\beta n}$]
$X_{Cij}^{process}$ is combined with [case frame $\gamma_{1...\gamma n}$]
X_{Cij}^{state} is combined with [case frame $\delta_{1...\delta n}$]

3 Analogous to step (1) and (2), but read X_{ij} instead of X_{Cij}.

4
$$\left.\begin{array}{l} G(X_{Cij}^{action}) = y_{cij}^{action},\ G(X_{ij}^{action}) = y_{ij}^{action} \\ y_{cij}^{action} = y_{ij}^{action} \end{array}\right\}$$
Analogous for position, process and state.

5 y_{cij}^{action} is described as $X_{Cij}^{action} \Rightarrow$ definition $\alpha_{1...\alpha n}$
Analogous for position, process and state.

6 $f(y_{cij}^{action}) = x_{1...n} \in X_{ij}^{action}$
Analogous for position, process and state.

7 Definition of x being the function of y_{cij}^{action} = definition of y_{cij}^{action}
Analogous for position, process and state.

8 Definition of x being the function of y_{ij}^{action} = definition of y_{ij}^{action} = definition of y_{cij}^{action}
Analogous for position, process and state.

5. Conclusion

This paper presents a method for the creation of a semantic subclassification of verbs in terms of primitive features and cases. This method has been inspired by Functional Grammar. We have tried to show that a formal model for definition generation can result from this semantic subclassification. The method has been developed in order to classify 1429 Dutch verbs. However, by integrating concepts in our model, it can also be applied to other verbs.

This model is a step towards definitions which are more suitable for NLP and still workable for human users for two reasons:

1. Similar verbs are treated in the same systematic way.
2. There is nothing too much and nothing too little in the definitions. The issue at stake is not the exhaustiveness of the definitions. Only those meaning components are of interest to the system which are common to a certain class of lexemes. Individual semantic peculiarities of a particular lexeme are not taken into consideration.

Certainly, a lot more is needed to make the definitions perfectly suitable for either NLP or the human user. This model is just meant to contribute to the bridging of the gap between people making definitions for one of these purposes.

Notes

1 Our feature *Change* is not identical to Vester's *change.*

References

Alshawi, H. 1989. "Analysing the dictionary definitions" in B. Boguraev, T. Briscoe (eds.), *Computational Lexicography for Natural Language Processing.* London / New York: Longman.

Chafe, W. 1970. *Meaning and the Structure of Language.* Chicago: University of Chicago Press.

Deville, G. 1989. *Modelization of task–oriented utterances in a man–machine dialogue system.* Thesis, Antwerpen.

Dik, S.C. 1978. *Functional grammar,* North–Holland linguistic series 37. Amsterdam.

Dik, S.C., Kahrel, P. 1992. "ProfGlot: a multilingual natural language processor". *Working papers in functional grammar* 45. Amsterdam.

Fillmore, C. 1968. "The Case for Case" in E. Bach, R.T. Harms (eds.), *Universals in Linguistic Theory.* New York: Holt, Rinehart and Winston.

Groot, De, C. 1983. "Typology of States of Affairs". *Linguistics in the Netherlands* 1983:73–81.

Martin, W., Demeersseman, H., Vliegen, M. 1992. *SNIV–Project, beschrijving, opzet en verantwoording.* Amsterdam.

Mel'chuk, I. 1981. "DEC Dictionnaire explicatif et combinatoire du français contemporain". *Recherches lexico–sémantiques* I, 3–49. Montréal: les Presses de l'Université de Montréal.

Vester, E. 1983. *Instrument and manner expressions in Latin.* Amsterdam, thesis.