7

Where FrameNet meets the Spoken Dutch Corpus: in the middle

PAOLA MONACHESI AND JANTINE TRAPMAN

Utrecht University, Uil-OTS
Trans 10, 3512 JK Utrecht, The Netherlands
Paola.Monachesi@let.uu.nl, Jantine.Trapman@let.uu.nl

Abstract

In this paper, we investigate to which extent FrameNet could be employed to enrich a syntactically annotated corpus such as the Corpus of Spoken Dutch with semantic role information. To this end, we have taken a language specific phenomenon such as the Dutch adjunct middle construction, as a test case.

7.1 Introduction

The interest for semantic annotation of corpora has grown in the last years. Applications such as information extraction, question-answering, document classification, and automatic abstracting that are based on underlying probabilistic techniques benefit from large corpora for improving their results and this is especially the case if these corpora are enriched with semantic information.

Several initiatives have been launched at the international level showing that it is pos-

Proceedings of the Sixteenth Computational Linguistics in the Netherlands.
Edited by: Khalil Sima'an, Maarten de Rijke, Remko Scha and Rob van Son.

sible to obtain concrete results with respect to the annotation of corpora with semantic information. Projects such as PropBank (Kingsbury, Palmer, and Marcus 2002), which has focussed on annotation of argumentstructure, have demonstrated that creating semantically annotated corpora need not be extremely expensive, and that it is possible to achieve a remarkable degree of consensus on a theory-neutral annotation methodology. On the other hand, projects such as Framenet (Johnson et al. 2002) have shown that it is possible to reach a considerable degree of granularity in the encoding of semantic roles.

However, while most initiatives have focused on English, not much attention has been dedicated to the creation of semantically annotated Dutch corpora, notably the Corpus of Spoken Dutch (CGN) lacks a layer of semantic annotation (Oostdijk et al. 2002). There is the need for appropriate guidelines with respect to the semantic annotation of Dutch corpora which could be adopted both for the annotation of the written Dutch corpus developed within the D-coi project (http://lands.let.ru.nl/projects/d-coi/) and for the already existing CGN.

In this paper, we discuss one type of semantic annotation, that is semantic role assignment. Semantic roles express the relationships identified between items in a text, such as the agents or patients of particular actions. The reason for our choice to focus on role assignment lies in the fact it is a thoroughly attested and feasible type of semantic annotation within corpora such as the already mentioned Framenet and PropBank projects and SALSA, (Erk et al. 2003) which takes the FrameNet dictionary as its basis.

We base our investigation on the already existing CGN in order to establish whether the annotation of semantic roles proposed within the FrameNet project could be adopted for Dutch and to which extent it can be integrated with the syntactic layer already present in CGN. The results, however, should be applicable also to a written corpus such as the one developed within the D-coi project.

Within the FrameNet project, a frame semantic lexicon has been developed which tries to encode all possible semantic and syntactic contexts for each entry. Moreover, the underlying frame ontology makes it possible to relate entries not only through membership of the same frame but also by means of inheritance relations. FrameNet is still under development, however, its methodology has been adopted to develop FrameNets for languages other than English. One important initiative in this respect is the German project SALSA, (Erk et al. 2003) which is not restricted to building a lexicon but it annotates the complete German Tiger corpus, (Brants et al. 2002) using the FrameNet dictionary and adapting it to German.

In order to assess whether the FrameNet lexicon can be employed to annotate a Dutch corpus with semantic role information, we have taken a specific phenomenon into consideration: the adjunct middle construction. This construction is quite similar to the object middle, which occurs both in English and Dutch. However, the adjunct middle does not occur in English (Hoekstra and Roberts 1993) and therefore it seems an appropriate test case to verify whether FrameNet can be adopted and eventually extended to deal with a language specific phenomenon. The adjunct middle construction constitutes a relevant phenomenon also because it is characterized by specific syntactic constraints as well as certain peculiar semantic properties which makes it a relevant case study for the interaction between syntactic and semantic annotation in corpora.

In the next section, we provide a detailed description of the various properties of the adjunct middle construction in Dutch, while in section 7.3 a brief introduction to the FrameNet project is given. Section 7.4 shows how the various adjunct middle verbs can be classified according to FrameNet frames while in section 7.5 the semantic roles which are involved in this construction are presented. Finally, section 7.6 discusses how the FrameNet lexicon can be employed to annotate the Corpus of Spoken Dutch, while 7.7 contains some concluding remarks.

7.2 The adjunct middle

The middle construction is characterized by an active voice, in the form of an intransitive verb, or a transitive verb that is used intransitively. Furthermore, a non-Agent is promoted to the subject position. An example is given by the active sentence in (12a) which can be transformed into the middle sentence in (12b). While sentence (12a) contains an Agent in the subject position and a Theme in the position of the direct object, in (12b) the Agent is no longer syntactically present and the direct object is now in the position of the subject:

(12) a. De padvinder schilt de aardappelen met een mesje.
The boy scout peels the potatoes with a knife
'The boy scout peels the potatoes with a knife'

b. Deze aardappelen schillen makkelijk.
These potatoes peel easy
'These potatoes peel easily.'

This type of construction, the object middle, is attested both in English and in Dutch, but in Dutch, another type of middle construction can be employed: the adjunct middle. It is characterized by the presence of an adjunct in the subject position, as exemplified by example (13a) below. No object is present in the middle construction which is consistent with its purpose: to focus on the (former) adjunct. In addition to adjuncts, demonstratives and the particle *het* ('it') can also occur as subjects, as shown in (13b), eventually in combination with *zijn* ('be') and an infinitive verb, as exemplified in (13c):[27]

(13) a. Dit mesje schilt handig.
This knife peels neat
'This knife is neat for pealing.'

b. Dat/het fietst prettig hier.
That/it cycles nice here

[27] The examples in this section are taken from (Ackema and Schoorlemmer 1993), (Ackema and Schoorlemmer 1995), (Haeseryn et al. 1997), (Peeters 1999) and (Hoekstra and Roberts 1993).

'It is nice to cycle here.'

c. Het is hier lekker zitten.
It is here nice sitting
'It is nice to sit here.'

The middle owes its name to the fact that it shares some of its properties with passives on the one hand, while on the other hand it shows some similarities with ergatives. In the rest of this section, the most important properties of the Dutch adjunct middle construction are summarized. Special attention is dedicated to those characteristics which directly affect the syntactic structure or the interpretation of the relationship between the verb and its arguments. These properties will eventually enable us to:

- identify the adjunct middle construction within the syntactically annotated data of the CGN;
- to assess whether we can represent it correctly within the theory of Frame Semantics as exemplified in FrameNet.

In particular, we will discuss the type of verbs which can be attested in this construction, the constraints on the subject, the presence of an implicit Agent as well as that of the compulsory modifiers, for more details we refer to (Peeters 1999).

The adjunct middle verb Not all verbs allow middle formation. The ones which allow adjunct middle formation are mostly intransitives although there is a number of verbs which allow both object and middle formation. However, If a verb of the latter group appears in a middle construction its object cannot be present. (Peeters 1999) divides the intransitives that trigger middle formation into three classes:

1. verbs of position;
2. verbs of physical activity, implying no locomotion;
3. (agentive) verbs of manner of motion (expressing no directional endpoint).

The subject The grammatical subject in an adjunct middle construction has to meet certain syntactic and semantic requirements. Three types of adjuncts are allowed in the subject position, that is an instrument (14a), a location (14b) or an external circumstance (14c), as shown by the examples below:

(14) a. Deze stoel zit lekker.
This chair sits comfortable
'This chair is comfortable to sit on.'

b. Deze sportzaal turnt prettig.
This gym does gymnastics nicely
'In this gym it is nice to do gymnastics.'

c. Regenweer wandelt niet gezellig.
Rainy weather walks not pleasant

’It is not pleasant to walk in rainy weather.’

In a regular matrix clause, these adjuncts are preceded by a preposition, but in the middle construction these prepositions have disappeared, as a comparison between (15) and (14a) reveals:

(15) Men zit lekker op deze stoel.
One sits comfortable on this chair
’One sits comfortably on this chair.’

(14a) Deze stoel zit lekker.
This chair sits comfortable
’This chair is comfortable to sit on.’

It is the following hierarchy which regulates the degree of acceptability of adjuncts:

$$\text{Instrument} \ll \text{Location} \ll \text{External Circumstance}$$

The leftmost element is the most eligible for middle formation while elements more to the right are less eligible. Thus, a middle verb which allows an adjunct of external circumstance in the subject position, automatically allows a Location or an Instrument in that position. As we have mentioned before the focus of the adjunct middle is on its subject which makes the presence of another element (e.g. an object, a purpose clause) not desirable. An additional constraint is that the subject should not represent a human entity.

The Agent The prototypical adjunct middle construction contains an Agent which does not surface in syntax, but is only implicitly present at the semantic level. The Agent can be characterized by the features [+animacy] and [+volitionality] (i.e. conscious and deliberate), but it is often interpreted as [+human]. In the agentive counterpart of the middle construction, the Agent is indicated by the arbitrary (pro)noun *men* (’one’, ’people’), as illustrated by example (16a) compared to (16b), which represents the adjunct middle version of (16a):

(16) a. Men loopt lekker op deze schoenen.
One walks nice on these shoes
’One walks nicely on this shoes.’

b. Deze schoenen lopen lekker.
These shoes walk nice
’On these shoes one walks nicely.’

Only under certain conditions, it is possible for an Agent to appear explicitly in the middle construction. In this case, it is represented by a PP introduced by the the preposition *voor* (’for’), this is possible in the case the Agent is generic or non-specific, as shown in (17a):

(17) a. Een krukje zit vervelend voor oude mannen /een oude man /?Hans.
A stool sits tedious for elderly men /an old man /?Hans
’A stool is tedious to sit on for elderly men / an old man /?Hans.’

b. Dit ijs schaatst goed genoeg voor Hans.
This ice skates good enough for Hans
’For Hans this ice is good enough to skate on.’

A sentence like (17b), where the Agent is a referential expression, is only allowed if the modifier has a restrictive, hence a comparative meaning.

The modifier The modifier encodes information on how the action of the predicate can be carried out with respect to the entity specified by the subject (Fagan 1992). The modifying element can be an adjective, as shown in the previous examples, negation (18b) or a stressed element (18a) and it has a dyadic character; On the one hand, the modifier refers to the subject, on the other hand, it is needed to identify the Agent:

(18) a. Dit ijs SCHAATST.
This ice skates
’This ice DOES skate.’

b. Dit ijs schaatst niet / lekker / *glad.
This ice skates not / nice / smooth
’This ice does not skate / skates nicely / *skates smoothly.’

Modifiers which are exclusively related to the subject or the Agent are excluded from middle formation, as is the case for the adverb *glad* ’smoothly’, in example (18b).

Due to the presence of the modifier, an implicit division automatically arises among the set of elements to which a certain property does (not) apply. This division can be quite explicit, as in (18b), where the distinction is made between ice that does skate (nicely) and ice that does not skate (nicely).

The semantics of the adjunct middle The adjunct middle construction focuses on (the properties of) the instrument, location or external circumstance, instead of the Agent. The passive sentence shows a similar character: the direct object occupies the position of the subject. Although the middle has some properties of passives, it is not sufficient to assign it a passive meaning as (Fagan 1992) does: "being able to be V-ed." (Peeters 1999) gives a somewhat different meaning description for the middle with structure ’NP V X’: "Adjunct NP enables whomever, to (un)succesfully V." The role of the modifier is left out of both descriptions, but could be filled in by adding "in an X manner".

Furthermore, the adjunct middle has the following semantic characteristics:

- non-eventiveness;
- it does not express or imply a completed change of location or state;

- it does imply an Agent;
- the Agent does not control the quality of the process (the Agent is more like an Experiencer);
- the modifier provides the middle with a comparative character.

After this general introduction of the properties of the adjunct middle, we will discuss in the next sections whether FrameNet can be assumed to classify the adjunct middle verbs according to its frames and whether the various elements of this construction can be labelled with appropriate semantic roles labels.

7.3 FrameNet

The Berkeley FrameNet project is based on the theory of Frame Semantics. Each frame represents a system of concepts related to each other (Petruck 1996). Words derive their meaning from the frame they belong to and their meaning is related to other words.

An example to illustrate the way FrameNet is structured can be given on the basis of the concept *buy*. The concept *buy* is included within a more abstract frame containing related concepts, e.g. *rent, spend, pay, cost* in this case. In FrameNet, this frame is called Commerce_buy (Johnson and Fillmore 2000). Concepts within the same frame may differ from each other due to the way in which the action is carried out, for example: pay with a bank/chip card or pay cash or because of the person involved in the transaction as in the case of *buy* vs. *sell*.

Besides the concepts which can be evoked in a frame, that is the so-called Frame Evoking Elements (FEEs) there are also Frame Elements (FEs) present in a frame. The elements *Buyer, Goods, Seller, Money* belong to the core of the concept associated with the verb *buy*. These frame elements represent the situational roles of the predicate. In case of *buy* the *Buyer* and *Goods* are obligatory, the other roles are optional. This information is encoded in the typical scenario which is described by a definition that covers all the possible contexts: each concept has such a prototypical scenario as basis.

The frame comprises a frame definition, a list of frame elements and a list of lexical units – the frame evoking elements. A lexical unit (LU) spells out all the various meanings of a word. The lexical entry encodes the valence description showing, by means of illustrative sentences, the various semantic and syntactic structures in which a LU can appear together with its frame elements. If a word has four different meanings, it has four lexical entries in FrameNet.

The complete description of a verb thus contains its frame definition, the elements of that frame, the grammatical properties of the verb and the various syntactic patterns in which it can appear (Petruck 1996). One problem concerning frame labels is that several parts of a sentence can evoke several frames simultaneously.

Not only lexical units are related to one another, frames themselves are mutually connected as well by means of subframes and *inheritance* or *using* relations. Inheritance is a

De stoel	zit	lekker	(Frame: Posture)
Location		Depictive	CNI: Protagonist
Ext		Mod	
NP		AdvP	

FIGURE 10 An adjunct middle sentence in FrameNet

"IS-A"-relation between the mother frame and a daughter. The daughter inherits the semantic (sub)type and the subframe structure from the mother. In addition, a daughter can include extra frame elements. The difference between *inheritance* and *using* is that the former implies complete inheritance whereas the latter involves incomplete inheritance. Notice that a daughter can have several mothers.

In summary, FrameNet is built out of three components (Fillmore, Baker, and Sato 2004):

1. the frame ontology (the set of frames)
2. the set of annotated sentences (examples of evoking the frames)
3. the set of lexical entries

The example in figure 10 illustrates how an adjunct middle sentence can be represented using FrameNet. The annotation of FrameNet encodes not only information about FEEs and FEs but also information about the syntactic function of the elements involved and information about their part of speech. The verb from our example evokes the frame **Posture** which is associated with the following definition: " The words in this frame describe the stable body posture of an Agent". *Protagonist, Depictive, Direction, Distance, Goal (e.g. lean against the wall), Location* and *Manner* are some FEs related to this frame. The adjective *lekker* constitutes also an FEE; it evokes the frame **Aesthetics**. But since this paper is only concerned with argument structure, the adjectival FEE is not discussed further. The Agent, which is called here the Protagonist, is syntactically absent, but it is present at the semantic level. In FrameNet, it is expressed at the end of the clause it belongs to, and the tag CNI: Constructionally licensed Null Instantiation is used to express this information.

7.4 Classifying adjunct middle verbs according to FrameNet frames

After this brief overview of the FrameNet system, we can now assess whether it can be employed to annotate the Dutch adjunct middle construction. The first step in this process is to establish to which frame a given verb belongs: the existing frame classification of FrameNet is used for this purpose.[28] The classification is based on English, but our assumption is that it should also be applicable to Dutch. In the rest of this section, we discuss under which frames the Dutch middle verbs can be grouped and which similarities and relationships these frames share.

[28] A complete overview can be found on the FrameNet website: http://framenet.icsi.berkeley.edu/

We have investigated sixty verbs which are extracted from example sentences in the literature and classified according to the three categories proposed by (Peeters 1999). They are listed in figure 11.

Each verb evokes one or several frames and different verbs can of course evoke the same frame. Since FrameNet is still under developement, it is incomplete; it does not contain every middle verb from our list. In those cases where the verb was not found in FrameNet, we have tried to assign it to an existing frame or to introduce a new frame if there was no appropriate one available.

A list was made of the frames that contain one or more middle verbs and if there were also non-middle verbs in the frame, we have verified whether they were eligible for middle formation. Finally, we have investigated how the frames that contain middle verbs are related to each other. This could be a direct relationship in which one frame inherits from or uses another frame. However, the relation could also be more indirect in the case two or more frames have the same mother.

For example, all the verbs belonging to the first colum, in figure 11, that is verbs of position evoke the frame **Posture**. All the additional verbs belonging to this frame can undergo middle formation in Dutch.

The other verbs listed in figure 11 belong to the frames summarized in figure 12, in which the various relations among frames are illustrated. Our aim was to generalize over types of verbs and frames which can be evoked in the adjunct middle construction. Figure 12 shows that middle verbs cannot be grouped under one frame but they belong to several ones. The most important mother frames are **Posture** (previously discussed) as well as **Intentionally_act** and **Motion**, represented in figure 12, which, however, are not connected with each other. The second frame itself does not contain middle verbs but it is included in the diagram because it has several daughters that do. It should be noticed that frames containing only one of the sixty investigated verbs include other verbs that can undergo middle formation, but also many verbs that cannot. Hence, we cannot simply state that if one frame includes some middle verbs, all the other verbs belonging to this frame can undergo middle formation. Furthermore, we should point out the presence of the **Sport** frame in figure 12. This frame does not exist in FrameNet, however, we have introduced it to group middle verbs that express sporting activities. The relations between the **Sport** frame and the other frames are only generally sketched. It should be left to the developers of FrameNet to assess the validity of this introduction further.

7.5 Assigning Frame Elements to adjunct middle verbs

In order to provide a complete representation of adjunct middle sentences, it is necessary to assign a label to the adjunct which is in the subject position, to the implicit Agent and to the modifier. Therefore, for each verb we checked which frame elements from the frame they evoked provided the suitable label. The list of frame elements (i.e. semantic roles) is ordered according to coreness and alphabetical order. So it seems that once we have

Verbs of position	**Verbs of physical activity**	**Verbs of manner of motion**
hangen	breien	draven
leunen	dansen	fietsen
liggen	eten	galopperen
rusten	golfen	glijden
staan	gooien	klimmen
steunen	kaarten	lopen
zitten	koken	rennen
	laden	rijden
	lezen	reizen
	praten	schaatsen
	roken	skiën
	schaken	springen
	schermen	stappen
	schillen	varen
	schoonmaken	vallen
	schrijven	vliegen
	schudden	wandelen
	slapen	zeilen
	spelen (toneel)	zwemmen
	spelen (sport, spel)	
	tekenen	
	tennissen	
	turnen	
	typen	
	vechten	
	vegen	
	voetballen	
	vrijen	
	werken	
	winkelen	
	zingen	

FIGURE 11 List of Dutch adjunct middle verbs analyzed

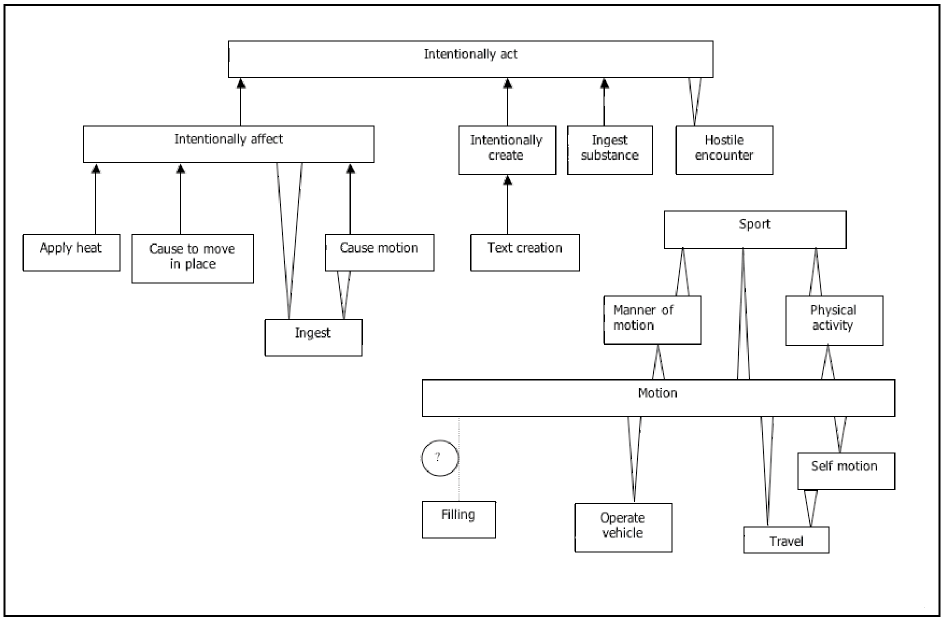

FIGURE 12 Frames including adjunct middle verbs and their relations

manually established to which frame a given adjunct middle verb belongs to, we have to detect the relevant frame elements whose label can be assigned manually to the various situational roles of the predicate.

In particular, for each verb, we have established which frame element would represent the (implicit) Agent. FrameNet uses various labels for what is traditionally called the Agent: e.g. *Ingestor, Self_mover, Cook, Interlocutor_1/Interlocutors, Author, Sleeper, Driver* allowing for a high degree of granularity but making it rather difficult to eventually automatize the annotation process.

Similarly, when it comes to possible adjuncts which can be on the subject position, three types are identified by (Peeters 1999), that is Instrument, Location and External Circumstance. FrameNet however, exhibits a higher degree of granularity. Therefore, for each relevant frame, we have established which frame element is allowed in the subject position of an adjunct middle verb. They can be both core and non-core elements. Potential subjects are *Goal, Theme, Instrument, Place, Area, Supporting_Bodypart, Vehicle, Circumstance*. From our investigation, it appears that frame elements with the same name are attested in various frames, however, not always with the same definition in each one. Therefore, since we cannot be sure that the description of a frame element is consistent through the whole lexicon we are obliged to examine for each element whether its definition varies across different frames. This uncertainty, in addition to the high degree of granulairty, makes it also in this case difficult to make the annotation process automatic.

In addition to the more fine grained labelling of adjuncts, there is another difference in terminology between FrameNet and the information found in the literature. In the sentence *De stoel zit lekker* ('The chair sits comfortably'), Peeters classifies the subject as an *Instrument*, while according to the description of **Posture**, *de stoel* ('the chair') is labelled as *Place*.

It is not standard that the modifier is present in a frame, however, if attested, it is usually represented as *Depictive*. For further details with respect to the classification of adjunct middle verbs according to FrameNet frames and for the labelling of the various semantic roles involved, we refer to (Trapman 2005).

7.6 Annotating the Spoken Dutch Corpus with FrameNet

In the previous sections, we have shown that it is possible to classify Dutch adjunct middle verbs according to FrameNet frames and to establish the semantic roles (Frame Elements) related to the various elements present in this construction. In this section, we illustrate how this information can be employed to enrich an existing corpus such as the Spoken Dutch Corpus with a semantic annotation layer. In particular, we discuss how a sentence in which the adjunct middle construction is attested can be annotated on the basis of the FrameNet information.

The Spoken Dutch Corpus includes about 8.900.000 words from both Flemish and Dutch sources including spontaneous conversations, telephone dialogues, news bulletins, read aloud texts etc. All together roughly 800 hours of spoken material in modern Dutch have been collected. The transcribed material has been enriched with part-of-speech tagging while a smaller part of the corpus has been annotated with phonetic, prosodic and syntactic information.

In order to indentify the adjunct middle construction in the corpus, we have employed the syntactically annotated part as well as the lexicon of the CGN. The middle construction can be identified as a predicate-argumentstructure which lacks an object, and some kind of AP has to be present within the dependency structure. Unfortunately, in the CGN, information about dependency structures and subcategorization is available but in two separate modules of the query tool. Therefore, the subcategorization information is not available while one is searching in the syntactic annotated part.

Despite these shortcomings, we were able to identify adjunct middle sentences correctly. In the rest of this section, we will provide some examples of annotation taken from the CGN, however, we will assume that the required information about subcategorization is available within the syntactic annotated corpus.

The sentences in figures 13, 14 and 15 are examples taken from the corpus. The annotation starts from the verb, which is the frame evoking element. A verb can evoke several frames at a time; other sentence elements determine the exact frame. In addition to their part-of-speech and their syntactic labels, lexical verbs, adjectival and nominal phrases get a semantic label, as well. In the case of the verb, the label represents the frame

(19) 'nou een luchtbed slaapt op zich wel heel erg fijn.'
well an air-bed sleeps in itself indeed very much comfortably
'well, an airbed in itself does sleep very comfortably indeed.'

<fn000682.326>

word	pos	syn	sem
nou	BW()		

een
luchtbed
	LID(onb,stan,agr)
	N(soort,ev,basis,zijd,stan)
		SU:NP	FE:Location

slaapt	WW(pv,tgw,met-t)	HD:V	(Sleep)
op	VZ(init)		
zich	VNW(refl,pron,obl,red,3,getal)		
wel	BW()		
heel erg fijn	ADJ(vrij,basis,zonder) ADJ(vrij,basis,zonder) ADJ(vrij,basis,zonder)	MOD:AdvP	FE:Depictive
.	LET		
		CNI: Sleeper	

FIGURE 13 A CGN sentence enriched with semantic information derived from FrameNet

(20) 'Nou en die bank zit niet zo lekker (marnix als de bank waar wij
well and that couch sits not as nice (marnix as the couch where we
nou op zitten.)'
at the moment on sit
'Well, sitting on that couch is not as nice (marnix as on the couch we are sitting on at the moment.)'

<fn00729.11>

word	pos	syn	sem
Nou	BW()		
en	VG(neven)		

die
bank
	VNW(aanw,det,stan,prenom,zonder,rest)
	N(soort,ev,basis,zijd,stan)
		SU:NP	FE:Location

zit	WW(pv,tgw,met-t)	HD:V	(Posture)

niet
zo
lekker
	BW()
	BW()
	ADJ(vrij,basis,zonder)
		MOD:AdvP	FE:Depictive

		CNI: Protagonist	

FIGURE 14 A CGN sentence enriched with semantic information derived from FrameNet

that is being evoked, resp. *Sleep*, *Posture* and *Operate*$_V$*ehicle*. Furthermore, adjectival and nominal phrases are labelled according to the frame element they represent. In the

(21) 'de auto rijdt makkelijk'
the car drives easy
'Driving the car is easy'

<fn008066.260>

word	pos	syn	sem
de	LID(bep,stan,rest)		
auto	N(soort,ev,basis,zijd,stan)	SU:NP	FE:Vehicle
rijdt	WW(pv,tgw,met-t)	HD:V	(Operate_vehicle)
makkelijk	ADJ(vrij,basis,zonder)	MOD:AdvP	FE:Depictive
		CNI:Driver	
.	LET()		

FIGURE 15 A CGN sentence enriched with semantic information derived from FrameNet

first example sentence, the subject 'een luchtbed' gets the role *Location* assigned, while the modifier gets the label *Depictive*. It should be noticed that not only verbs are FEEs, other elements of the sentence can also be a FEE. FrameNet has a strategy to deal with this phenomenon, but we will ignore it in this paper. At first sight, there is no difference in the annotation of middles and other verbs. The difference lies in the presence of the Agent: if there is an FE, other than the Agent, in the subject position, then the Agent is automatically represented as CNI (in special cases it surfaces as a voor-PP). In our first example sentence, the Agent is a *Sleeper*.

7.7 Conclusion

The goal of this paper was to verify to which extent FrameNet could be employed to enrich a syntactically annotated corpus such as the CGN with semantic role information. To this end, we have taken a language specific phenomenon such as the Dutch adjunct middle construction, as a test case.

From our investigation, we can conclude that there is only a partial correspondence between the classification of the Dutch adjunct middle construction as attested in the literature (Peeters 1999) if it is compared with the FrameNet classification. This is due to the wide distribution of the adjunct middle verbs over the frames which goes beyond the division in three classes proposed by Peeters. However, we can distinguish a restricted set of frames that contain middle verbs, i.e. Intentionally_act, Motion and Posture indicating that FrameNet is suitable for making linguistic generalizations.

On the other hand, when it comes to frame elements this is not the case, since the traditional Agent role gets many different labels across various frames. Other frame elements are more constant across frames although their definitions are not always the same. As for the labelling of the adjuncts which surface in subject position, we also see a more fine grained division in FrameNet than that postulated in the literature. More generally, FrameNet reaches a level of granularity in the specification of the semantic roles which might be desirable for certain applications (i.e. Question Answering). However, it makes automatic annotation of semantic roles rather impossible and might even raise problems with respect to uniformity of role labelling even if human annotators are involved.

Furthermore, incompleteness constitutes a serious problem, i.e. several frames and relations among frames are missing mainly because FrameNet is still under development. Adopting the FrameNet lexicon for semantic annotation means contributing to its development with the addition of (language specific) and missing frames. Incompleteness is also a problem within the CGN since at its present stage the corpus lacks information about subcategorization, which, however, can be inferred on the basis of the dependency structure.

In our study, we have assumed that the FrameNet classification even though it is based on English could be applicable to Dutch as well. Although Dutch and English are quite similar, there are differences on both sides. For example, in the case of the Spanish FrameNet it turned out that frames may differ in their number of elements across languages (cf. (Subirats and Petruck 2003) and (Subirats and Sato 2004)).

On the basis of our preliminary investigation, we can conclude that FrameNet offers a way to correctly classify the Dutch adjunct middle verbs. Even though some problems have emerged, our test case indicates that the FrameNet lexicon can be employed to semantically annotate the Spoken Dutch Corpus. However, we need to verify in more details to which extent the English frames translate into Dutch frames. In this respect, we can benefit from results from projects like SALSA ((Erk et al. 2003)) where FrameNet is used to annotate the German Tiger Corpus ((Brants et al. 2002)).

In our study, we have assumed the Spoken Dutch Corpus as our basis. We still have to assess whether the FrameNet lexicon is also suitable for the semantic annotation of the written Dutch corpus which is being developed within the D-Coi project which employs the Alpino parser to add the syntactic layer of annotation to the corpus. Furthermore, we did not yet discuss the possibility of applying the PropBank approach to role assigment (Kingsbury, Palmer, and Marcus 2002). This approach is essentially corpus based and syntax driven and while the more semantic driven FrameNet approach which is based on a network of relations between frames. Another difference is that in PropBank verbs are not categorized under a specific concept but for each verb its *sense(s)* are classified under a framefile and the set of possible semantic roles is more restricted. In this respect it is worth noticing that the PropBank framefiles are quite different from the FrameNet frames. In our follow-up study (Monachesi and Trapman 2006) we examine in more detail the differences and similarities of the two approaches and the possibilities they provide for semantic annotation. We also consider in this paper the reconciliation of the two since this might result in a scheme which includes ontological information, without having a too fine grained list of possible roles.

References

Ackema, P. and Schoorlemmer, M. (1993). The middle construction and the syntax-semantics interface *Lingua 93, pp. 59-90*.

Ackema, P. and Schoorlemmer, M. (1995). Middles and Nonmovement, *Linguistic Inquiry 26, pp. 173-197*.

Brants, S., Dipper, S., Hansen, S., Lezius W. and Smith G. (2002). The TIGER Treebank, *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol.

Erk, K., Kowalski, A., Pado S. and Pinkal, M. (2003). Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of ACL 2003*. Sapporo.

Fagan, S. (1992). The syntax and semantics of middle constructions. Cambridge University Press.

Fellbaum, C. (1986). On the middle construction in English. Bloomington, Indiana: Indiana Univ. Linguistics Club.

Fillmore, C.J., Baker, C.F. and Sato, H. (2004). FrameNet as a net, *Proceedings of LREC*, Lisbon, Elra. Volume 4, pp. 1091–1094.

Haeseryn, W., Romijn, K., Geerts, G., De Rooij, J. and Van den Toorn, M.C. (1997). Algemene Nederlandse Spraakkunst. Tweede, geheel herziene druk, 1997. Groningen/Deurne, Martinus Nijhoff uitgevers/Wolters Plantyn, pp. 50–55.

Hoekstra, T. and I. Roberts (1993). Middle constructions in Dutch and English, *Knowledge and Language*. Kluwer Academic Publishers, Dordrecht, pp. 183–220.

Johnson, C.R. and Fillmore C.J. (2000). The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure, *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, Seattle WA, pp. 56–62.

Johnson, C.R., Fillmore, C.J., Petruck, M.R.L., Baker, C.F., Ellsworth, M.J., Ruppenhofer, J., and Wood, E.J. (2002). FrameNet: Theory and Practice (e-book), `http://framenet.icsi.berkeley.edu/book/book.pdf`

Kingsbury, P., Palmer, M. and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank, *Proceedings of the Human Language Technology Conference. HLT-2002*. San Diego, California.

Monachesi, P. and Trapman, J.R. (2006). Merging FrameNet and PropBank in a corpus of written Dutch, *Proceedings of the workshop Merging and Layering Linguistic Information, LREC-2006*. Genoa, Italy.

Oostdijk, N., Goedertier, W., Van Eynde, F., Bovens, L., Martens, J.P., Moortgat, M. and Baayen, H. (2002). Experiences from the Spoken Dutch Corpus Project, *Proceedings of LREC-2002*, pp. 340–347.

Peeters, R.J. (1999). The adjunct middle construction in Dutch, *Leuvense Bijdragen*, jaargang 88, pp. 355–401.

Petruck, M.R.L. (1996). Frame Semantics, *in* Verschueren, J., Östman, J., Blommaert, J. and Bulcaen, C. (eds.), *Handbook of Pragmatics 1996*. Philadelphia: John Benjamins.

Subirats, C. and Petruck, M.R.L. (2003). Surprise: Spanish FrameNet!, *in* Hajicova, E., Kotesovcova, A. and Mirovsky, J. (eds.), *Proceedings of CIL 17*. Prague: Matfyzpress.

Subirats, C. and Sato H. (2004). Spanish FrameNet and FrameSQL, *Proceedings of the workshop Building Lexical Resources from Semantically Annotated Corpora, LREC-2004*. Lisbon, Portugal.

Trapman, J.R. (2005). Where FrameNet meets the Dutch Spoken Corpus: in the middle. Bachelor thesis. Utrecht University.