5

The Contours of a Semantic Annotation Scheme for Dutch

INEKE SCHUURMAN AND PAOLA MONACHESI

Centrum voor Computerlinguïstiek, K.U.Leuven
Maria-Theresiastraat 21, 3000 Leuven, Belgium
ineke.schuurman@ccl.kuleuven.be
and
Utrecht University, Uil-OTS
Trans 10, 3512 JK Utrecht, The Netherlands
Paola.Monachesi@let.uu.nl

Abstract

The creation of semantically annotated corpora has lagged dramatically behind. As a result, the need for such resources has now become urgent. Several initiatives have been launched at the international level in the last years, however, they have focussed almost entirely on English and not much attention has been dedicated to the creation of semantically annotated Dutch corpora. The Flemish-Dutch STEVIN-programme has identified semantic annotation as one of its priorities.Within the project "Dutch Language Corpus Initiative" (D-Coi) we are developing guidelines for the semantic annotation of Dutch and our focus is on two types: semantic role assignment and temporal and spatial semantics.

5.1 Introduction

The realization of an appropriate digital language infrastructure for Dutch is one of the objectives of the Dutch/Flemish STEVIN programme which has been recently launched.[16] In particular, the need for a large corpus of written Dutch, comprising 500-million-words has been identified as one of the top priorities. This corpus should be tailored to the needs of scientific research and commercial applications and should improve the development of other resources and tools. Applications such as information extraction, question-answering, document classification, and automatic abstracting that are based on underlying probabilistic techniques should benefit from it.

All texts in the corpus will conform to standards for character encoding and markup. Furthermore, the corpus will be linguistically annotated. For the various annotation layers, annotation schemes must be decided upon and the aim is to revise and adapt the protocols which have been developed for the Spoken Dutch Corpus (CGN) (Oostdijk, Goedertier, Van Eynde, Boves, Martens, Moortgat and Baayen 2002).

A pilot study is being carried out to this end: the *Dutch language Corpus Initiative* (D-Coi) is a project launched within the STEVIN programme whose aim is a blueprint for the construction of the 500-million-word corpus.[17] The project is concerned with issues related to the design of the corpus and the development (or adaptation) of protocols, procedures and tools that are needed for sampling data, text regularization, converting file formats, marking up, annotating, post-editing, and validating the data. Within the D-Coi project, a 50 million word pilot corpus will be compiled, parts of which will be enriched with (verified) linguistic annotations. The pilot corpus is intended to demonstrate the feasibility of the approach. It will provide the necessary testing ground on the basis of which feedback can be obtained about the adequacy and practicability of various annotation schemes and procedures, and the level of success with which tools can be applied.

One of the innovative aspects of the D-Coi project is that it will focus not only on the revisions of those protocols which have been already developed within the Spoken Dutch Corpus for PoS tagging, lemmatization and syntactic annotation but it will also explore the possibility of integrating an additional annotation layer based on semantic information. This annotation layer was not present in the Spoken Dutch Corpus.

The need for semantically annotated corpora has now become urgent. Several initiatives have been launched at the international level in the last years, showing that the time is ripe for activities in this direction. However, they have focussed almost entirely on English and not much attention has been dedicated to the creation of semantically annotated Dutch corpora. One of the goals of the D-Coi project is the development of a protocol for such an annotation layer. Only a small part of the corpus will be annotated with semantic information (i.e. 3000 words), in order to yield information with respect to its feasibility.[18]

---

[16]http://taalunieversum.org/taal/technologie/stevin/

[17]http://lands.let.ru.nl/projects/d-coi/

[18]The manual for semantic annotation (Monachesi and Schuurman 2006) plus the 3000 word corpus will be

A more substantial annotation effort could be carried out in the framework of the 500 million word corpus. In this follow-up project other types of semantic annotation might also be taken into consideration, as well as their interaction with other levels like PoS tagging and syntactic analysis. We are therefore taking this interaction into consideration when developing the protocols.

For the moment, we are only dealing with two types of semantic annotation and their interaction, that is semantic role assignment and temporal and spatial semantics. The reason for this choice lies in the fact that semantic role assignment (i.e. the semantic relationships identified between items in the text such as the agents or patients of particular actions), is one of the most attested and feasible types of semantic annotation within corpora. On the other hand, temporal and spatial annotation was chosen because there is a clear need for such a layer of annotation in applications like information retrieval or question answering.

5.2   Semantic role assignment in D-Coi

During the last few years, corpora enriched with semantic role information have received much attention, since they offer rich data both for empirical investigations in lexical semantics and large-scale lexical acquisition for NLP and Semantic Web applications. Several initiatives are emerging at the international level to develop annotation systems of argument structure, within the D-Coi project we intend to exploit existing results as much as possible and to set the basis for a common standard. We want to profit from earlier experiences and contribute to existing work by making it more complete with our own (language specific) contribution given that most resources have been developed for English.

The following projects have been evaluated in order to assess whether the approach and the methodology they have developed for the annotation of semantic roles could be adopted for our purposes:

- PropBank (Kingsbury, Palmer and Marcus 2002);
- FrameNet (Johnson, Fillmore, Petruck, Baker, Ellsworth, Ruppenhofer and Wood 2002);

Given the results they have achieved, we have taken their insights and experiences as our starting point. In the rest of this section, we will consider them more in detail in order to evaluate their strengths and weaknesses and to assess which features of the existing systems we want to include in the scheme for the semantic annotation of the D-Coi corpus.

available through the TST-centrale early 2007 (http://www.tst.inl.nl/).

5.2.1 PropBank

PropBank aims at adding a layer of semantic annotation to the Penn English TreeBank (Marcus, Santorini and Marcinkiewicz 1993). It provides a semantic representation of argument structures that are labeled consistently in such a way that the data are usable for automatic extraction. PropBank uses a very restricted set of argument labels.

The PropBank lexicon, which was added first to facilitate annotation and later evolved into a resource on its own, is constructed following a 'bottom-up' strategy: starting from the various senses of a word, a framefile is created for every verb. Such a framefile contains thus all possible senses of the verb plus a set of example sentences that illustrate the context in which the verb can occur. For each sense of the verb, a roleset and example sentences are available. Therefore, when a verb has two senses its framefile contains two different rolesets as is the case with *leave* from (Babko-Malaya 2005):

(3)   a.   Frameset leave.01 *move away from*
           Arg0: entity leaving
           Arg1: place left
           Mary left the room

      b.   Frameset leave.02 *give*
           Arg0: giver
           Arg1: thing given
           Arg2: beneficiary
           Mary left her daughter-in-law her pearls in her will

To create a framefile, relevant sentences are extracted from the corpus. Based on those sentences, the most frequent and/or necessary roles are selected and one or more rolesets are formed. In this way, the most common senses of the verb are stored in the framefile. An interesting feature of the PropBank project is that the corpus has been annotated automatically with 83% accuracy and then corrected by hand on the basis of the developed lexicon. Furthermore, the goal of the project is to provide training data for supervised automatic role labelers. This is a desirable objective since it will be possible to annotate corpora of the size of D-Coi with semantic role information only if the the process is semi-automatic.

5.2.2 FrameNet

Contrary to PropBank, FrameNet does not annotate a complete corpus, but one that contains example sentences that illustrate all possible syntactic and semantic contexts of the lexical items taken into consideration. Besides the corpus, two other components can be distinguished in FrameNet, that is a set of lexical entries and a frame ontology.The development of the ontology is based on the frames. A frame represents a certain prototypical situation which is described by the frame definition. Every frame contains also a list of frame elements and a set of lexical units that can evoke the frame. The term *lexical unit* is used for a word in combination with one of its senses (Johnson et al. 2002). The frame elements fulfill a certain semantic role within the situation that is evoked by one of the

lexical units. For every lexical unit a set of sentences is selected that illustrate all possible occurences of the lexical unit; all possible semantic roles are annotated in these sentences.

For example, the verb *leave* would evoke the frame *Departing* which is (partly) shown below:

- Departing
  An object (the Theme) moves away from a Source. The Source may be expressed or it may be understood from context, but its existence is always implied by the departing word itself.
- Frame Elements: Source, Theme, Area, Depictive, Distance, Manner, Goal etc.

A sentence annotated with semantic roles on the basis of the FrameNet information, would receive the following representation:

(4) $[_{Theme}$ We all $]$ left $[_{Source}$ the school$]$ $[_{Time}$ at four o'clock $]$.

Although FrameNet is still under development, its approach has been adopted for the annotation of semantic roles for languages other than English. An example is provided by the German project *Saarbrücken Lexical Semantics Annotation and analysis* (SALSA) (Erk, Kowalski, Pado and Pinkal 2003), but there are also projects based on FrameNet for languages such as Spanish, French and Japanese. However, SALSA distinguishes itself from the others by the fact that it is not restricted to building a lexicon but it annotates the complete German Tiger corpus using the FrameNet dictionary and adapting it to German. Unlike FrameNet, SALSA is not committed to always assigning a single sense (frame) to a target expression, or a single semantic role to a constituent but either more than one or an *Underspecified* sense tag can be assigned in case of vagueness or ambiguity.

5.2.3 Comparing approaches

The main differences we have noticed between FrameNet and Propbank are related to the methodology employed in the construction of the lexicon and the way the lexicon is structured. More generally, the classification attested in PropBank is based on *word senses* which are grouped in the 'shallow' framefiles while the FrameNet classification is driven by the *concepts* which are structured in the ontology of frames and thus based on hierarchically structured semantic classes.

Furthermore, the two projects differ with respect to the granularity of the role labels employed. FrameNet uses labels which immediately reflect the semantic role of the constituent and its annotation is rich in information. PropBank labels require more careful investigation about the meaning of the constituent in question.

The FrameNet labels are rather rich in information, however, they might not always be transparent for users and annotators. On the other hand, the advantage of the PropBank appoach is that by employing neutral labels, less effort is required from annotators to assign them. Furthermore, it creates the basis for the development of semi-automatic

annotation of role labels, which is a necessary requirement if we want to annotate large corpora.

5.3 Merging approaches

In developing a scheme for the semantic annotation of the D-Coi corpus, we are faced with several options.

We could assume the FrameNet approach and develop a Dutch lexicon based on the English (and German) one and employ it for the annotation of the Dutch corpus. We would thus follow the strategy employed within the SALSA project and we could even exploit their results given the similarity between Dutch and German. A disadvantage of this choice is related to the fact that in order to annotate the corpus further we are bound to construct new frames (with their definitions and their frame entities) manually and this is a rather expensive process. Furthermore, we believe that the labels used to identify the frame entities are not very transparent and difficult for annotators to use.

The other possibility would be to employ the PropBank approach which has the advantage of providing clear role labels and thus a transparent annotation for both annotators and users. Furthermore, the annotation process could be at least semi-automatic. However, a disadvantage of this approach is that we would have to give up the classification of frames in an ontology which could be very useful for certain applications, especially those related to the Semantic Web.

Within the D-Coi project, we have chosen for a third option which wants to reconcile the rather pragmatic PropBank approach to role assignment which is essentially corpus based and syntax driven with the more semantic driven FrameNet approach which is based on a network of relations between frames. More generally, we would like to adopt the conceptual structure of FrameNet, but not necessarily the granularity of its role assignment approach. With respect to role assignment, we would like to adopt the annotation approach of PropBank. A risk we take is that we will end up with a semantic annotation layer which is too similar to the syntactic representation which is assumed in D-Coi. This will be an extension of that developed for the Spoken Dutch Corpus, that is a dependency structure which carries information about heads, complements and modifiers. However, a preliminary study carried out in (Stevens 2006), has shown that this is not the case and that the PropBank role labels provide additional useful information, especially with respect to modifiers. Stevens has suggested a heuristic strategy to map nodes in a D-coi dependency tree to PropBank argument labels. This strategy is implemented in a rule-based semantic role tagger (XARA), which has assigned 65% of the roles of the selected corpus correctly.

In order to assess the feasibility of our approach we have carried out a pilot study involving the integration of PropBank and FrameNet. The goals behind this study are:

- to assess whether it is possible to merge FrameNet frames with PropBank role labels and whether this merging has to be manual or whether it is possible to make it at least semi-automatic;

- to investigate to which extent we can use resources already developed for other languages;
- to assess whether we can extend existing resources on the basis of our language specific annotation and whether we should include the language specific features in the original resource;
- to investigate whether it is possible to extend the merged resources by exploiting the best features of both and in this way facilitate the process.

In our study we have considered a language independent phenomenon such as the classification of verbs of communication and a more language specific phenomenon, such as the classification of (adjunct) middle verbs. We refer to (Monachesi and Trapman 2006) for more details about the pilot study, in the rest of this section we will exemplify the merging approach on the basis of the *Communication* frame.

5.3.1 Merging approaches: The Communication frame

As previously discussed, FrameNet provides a rich semantic representation of language because its lexicon not only encodes word senses but also relations among words. Words can be related to each other on the basis of the frame they share, but also because a relationship among frames is established. The ontological relations add extra information to word senses. In FrameNet, every frame has its own definition which distinguishes it from other frames while the frame elements can be the same across frames due to (partial) inheritance. However, there is a great variety of elements that are frame specific creating thus a quite complex structure which is not always very transparent for annotators and users. Furthermore, by taking into account the *Communication Frame*, which comprises a mother frame with six daughters, we have noticed that the inheritance relation is not as strict as we had assumed. Our aim is thus to reduce the FrameNet frames to a simpler form in which the set of frame elements is restricted to a number of elements that is comparable with the PropBank arguments. Since the interpretation of a PropBank argument label depends on the word senses of the individual word, we wanted to make their interpretation more uniform as well. This can be achieved by assuming Levin's classes and diathesis alternation and the revisions implemented within VerbNet . Verbs within the same Levin class, sharing the same diathesis alternations, should have the same roleset. Thus, the second step is to group together those verbs that share the same FrameNet frame, the same Levin class and diathesis alternation and assign this group one roleset; this roleset is derived from PropBank by selecting the most common arguments from one group of verbs. Regrouping the verbs this way decreases the number of rolesets in comparison with the number of PropBank rolesets. The advantage is that we can determine rolesets using a simple algorithm that results in an intersection of the FrameNet and the Levin classification. Assigning rolesets to these newly created classes takes less effort than manual role assigment for every individual verb. We then compared this roleset with the frame elements that are normally used.

As a test case we took the frame *Communication* which has six daughters:

Communication-manner, Communication-noise, Communication-response, Gesture, Reassuring and Statement. For example, the verbs comprised in *Communication-noise* belong to four different Levin classes with the majority of the verbs belonging to the class *Verbs of Manner of Speaking*. These are verbs like *babble, bellow, croon, hiss, wail, whine* etc. A general roleset for this group could be:

PropBank	FrameNet
Arg0: speaker, communicator	Speaker
Arg1: utterance	Message
Arg2: hearer	Addressee

TABLE 3 Roleset for *Communication-Noise*

The alignment in the case of the communication verbs is rather straightforward and it seems to suggest that indeed this methodology might be appropriate.

5.4   Temporal semantics

5.4.1   Background

The layer of temporal and spatial annotation is meant to be useful for both scientific research as well as applications (information retrieval, question answering, multidocument summarization, etc.). Within the STEVIN-programme this layer of annotation is part of a whole series of annotations, from part-of-speech (or morphosyntax) over syntax to several semantic ones. It goes without saying that it is to reflect the state of the art. In that respect TimeML (Saurí, Littman, Knippen, Gaizauskas, Setzer and Pustejovsky 2006) comes to mind as far as temporal annotation is concerned.[19]
TimeML is a temporal markup language, a joint effort reflecting many ideas from other, earlier approaches (it is strongly based on an earlier version of TIDES ((Ferro, Gerber, Mani, Sundheim and Wilson 2002)) as well as on (Setzer 2001) whose authors were among the developers of TimeML, the main difference being that more types of phenomena are annotated, especially those related to events and to tense and aspect). But note that also for TIDES there is an adapted version (TIDES 2005) of their manual, still concentrating on the core time expressions, so-called timexes (such as calendar dates). Within D-Coi we wanted to annotate states and events as well, cf. TimeML.

TimeML is designed as a common meta-standard for temporal annotation covering the recognition of all temporal elements (i.e. expressions and events (for the latter notion, see pt 2 below)), anchoring of these elements and relating them to each other.
There are four meta data structures to be annotated, cf. (Day, Ferro, Gaizauskas, Hanks, Lazo, Pustejovsky, Saurí, See, Setzer and Sundheim 2003):

---

[19]In this paper we will concentrate on temporal annotation. Spatial annotation is done in a similar way, concentrating on spacexes instead of timexes.

- Events: these describe all situations that occur or happen. States are considered to be events, only a subset of these will be annotated.
- Times (Timex3): points, intervals or durations. These may be referred to by fully specified temporal expressions (like *May 4th, 2005*), by underspecified expressions (*Monday*), contextually dependent expressions (*last week*), . . .
- Signals: elements (like prepositions, conjunctions) indicating how temporal objects are to be related.
- Links: these describe the relations between events, and between events and times or signals. There are three kinds of links: temporal links (like `before`, `immediately after`, `included in`), subordination links (like `modal`, `negative`, `factive`), and aspectual links (like `initiation`, `continuation`).

TimeML is by far the most elaborated annotation scheme around these days, there is also a still rather small corpus available (TimeBank) that is annotated according to the TimeML guidelines.

There are, however, a few problems when adopting (and adapting) a scheme like TimeML for the Flemish/Dutch STEVIN programme:

1. we want to make use of information available through other layers like Syntactic Analysis (SA) and Part of Speech tagging (PoS) when analysing the sentences
2. the semantic foundation should be a sound one
3. the annotation should be useful for the scientific community
4. annotation should be feasible in a semi-automatic way although we want to annotate all sentences in a text, i.e. also those without so-called timexes.

With respect to point 1, like most annotation schemes around TimeML does start from scratch, not really taking into account other annotation layers (at least not in a way a script can be aware of it). Within D-Coi we wanted to make use of all information available (such as Part of Speech, Syntactic Analysis)

The issue under 2 is of a more serious nature. In TimeML, states are considered particular types of events, which is not correct: they are at the same level, and they both belong to the 'eventualities' (or 'situations').[20] The other types of 'events' used in TimeML are also not standard ones, cf. above. We therefore will not make use of this part of TimeML, although we do see the merits of a characterization of verbs in order to rate the relevance of a temporal expression. We are using a separate feature to accommodate this.

Point 3 is related to point 2: an elaborated tense and aspect component is often not considered necessary for applications, especially when the corpus to be annotated consists of news items, cf. (Setzer 2001). We wanted to make use of a more elaborate theory of tense and aspect than the one used in TimeMl as this is of importance for temporal semantics (as opposed to temporal annotation), the more as we have to annotate all kinds of texts,

---

[20]Cf. also (Mani, Pustejovsky and Gaizauskas 2005), p. 491.

that is not only news items, but also fiction and the like. We therefore want to merge the ideas behind TimeML (and TIDES) with those of theories like Discourse Representation Theory (Kamp and Reyle 1993). This is in fact our most serious content-related objection to TimeML.[21]

The last point seems to be contradictory: we want to annotate more phenomena, and at the same time we want to do it in a semi-automatic way, whereas other annotation schemes seem to rely heavily on a firm amount of manual annotation. We do so by making use of compositionality, exploiting all the regularities in the language. We also need to annotate full texts, not isolated sentences.[22]

Our approach

As remarked in the previous section, in our approach we cover more or less the same phenomena as in TimeML: we annotate all temporal expressions (nouns, prepositions, adverbs, adjectives, conjunctions, as well as the relations between them); eventualities (events, states and processes); characterization of eventualities as reporting, perception etc in order to be able the reliability of an statement (compare: *Bill Gates is CEO of Microsoft* vs *Bill Gates claims/is said to be CEO of Microsoft*); tense and aspect.

We exploit lexica with temporal items (lemmata, sometimes with a very specific PoS label in order to be able to refer to a 'token'; expressions). In order to give an idea how entries would be combined, have a look at the following table with the temporal semantic information attached to the leaf nodes of the temporal expression *23 maart 1967 om twintig na drie* (the 23rd of March 1967 at 20 minutes past three).

23	t-ent="yes" t-value="D23"
maart	t-ent="yes" t-value="M03"
1967	t-ent="yes" t-value="Y1967"
om	t-ent="yes" mod="at"
twintig	t-ent="yes" t-value="T20M"
na	t-ent="yes" mod="after"
drie	t-ent="yes" t-value="T03|15H"

In combination this will become **1967-03-23T03|15:20**.[23] In case we would have known that it would be *drie uur 's middags* (three o'clock in the afternoon) the full expression would get the value **1967-03-23T15:20**. In the first expression the value of the hour is left underspecified.

There will also be a lexicon containing those expressions that at first sight seem to be temporal as they contain an item that usually is used in a temporal way: *Zwarte September*

---

[21] Within the framework of this paper it is not possible to give a detailed overview of our approach as far as tense and aspect is concerned. We refer the interested reader to the manual (Monachesi and Schuurman 2006).

[22] Note that for phenomena like coreference we rely on a project like COREA, another STEVIN-project, to solve these computationally.

[23] We will use | as a symbol meaning 'or'.

(Black September), *De Morgen* (a Flemish newspaper), *een dagje ouder worden* (be getting on a bit), *ouden van dagen* (elderly people). Such expressions will be excluded from temporal annotation. Note that some expressions are ambiguous in this respect: *Het is vijf voor twaalf* (a) It is five to twelve; b) We are on the verge of disaster). In such a case the broader context will be dicisive.
On the other hand, one also needs a lexicon containing expressions that do get a temporal interpretation although they don't contain temporal items, like *Tweede Wereldoorlog* (Second World War).

As said above, up till now the only annotation scheme in which events (or rather eventualities) get an inclusive treatment , cf. (Mani et al. 2005) (but see above) is TimeML.

But also when dealing with timexes (thus neglecting eventualities) some problems arise. In (Mani et al. 2005) three types are mentioned as problematic:

1. indexicals: contextual dependent expressions, like *Wednesday* (which Wednesday?) or *next week*
2. relational expressions: times are specified in relation to other times *two weeks after Christmas*
3. vagueness: times with inherently vague boundaries *spring, evening*

In the next section we will say more about the third category (vague expressions) as they occur in general language, and the way we deal with such expressions.

Vague expressions

When reasoning with time, for example to answer a *when*-question, one is confronted with an annoying human characteristic: people tend to use their language in a very sloppy way.
It is therefore sometimes rather 'dangerous' to deduce temporal information and to reason with it (but see (Pan, Mulkar and Hobbs 2006).) There are several ways of being sloppy when temporal information is concerned. We will try to accommodate these in various ways.

Case A:

Suppose today is Friday, March 10, 2006. When refering to Saturday the 18th in the Netherlands one will use the expression in (5), in Flanders the one in (6):

(5) morgen   over een week
     tomorrow over a   week

     a week from tomorrow

(6) morgen   over acht dagen
     tomorrow over eight days

     a week from tomorrow

The Flemish expression *over acht dagen* will be in the lexicon with the meaning *over* **zeven** *dagen*.[24]

The following case is more serious:

Case B:

Suppose it is April 18th 2006, the day after Easter Monday. That Tuesday you can mention that in 14 days time there is already another public holiday, Labour day (Monday 1st). Nobody will correct you, saying that it should be '13 days'. One should have protested in case you did say *in exactly 14 days*.
14 days or *2 weeks* are a kind of rather global containers, meaning *more or less 14 days/2 weeks*, whereas for example *12 days* or *16 days* only have a strict meaning.
In order not to jump to false conclusions (like: Labour Day is May 2nd) we use a boolean feature *noise*. That way we would still conclude that Labour Day is May 2nd, but with the warning that this can be wrong. Note that in a case like this a human corrector will correct the mistake as everybody knows that the reference is to May 1st. The point however is that such sloppy (temporal) containers are used time and again. In case of expressions like *14 dagen*, we add the feature noise="yes" which expressions like *16 dagen* will not have.

Case C:

There is yet another type of global temporal expressions, those in which the uncertainty is made explicit in the wording:

(7) Het was ongeveer   middernacht toen …
    It   was more or less midnight   when …

    It was about midnight when …

(8) Het was rond   middernacht toen …
    It   was around midnight   when

    It was around midnight when …

In these cases we will use mod="approx", in order to modify the value T24:00, cf. (Ferro, Gerber, Mani, Sundheim and Wilson 2005) and (Saurí et al. 2006). Note that in case of

(9) Het liep   tegen   middernacht
    It   got on towards midnight

    It was getting on towards midnight

(10) Het was net na middernacht
      Het was net na middernacht

---

[24]Note that in French the expression is *quinze jours* (fifteen days) instead of *two weeks*.

It was just after midnight

the value of the modifier will not be just "approx", but the more specific "just-before" resp. "just-after".

Case D:

The seasons of the year are also often used in a sloppy way, refering globally to those particular months. In most annotation schemes they are annotated as SP, SU, FA or WI respectively (for 'spring', 'summer', 'fall' and 'winter'). Within D-Coi we refer to the seasons with months as we want to be able to order eventualities as in 11:

(11) De ring   rond   Antwerpen wordt deze zomer   vernieuwd. Eind Mei
     The ring road around Antwerp   will be this summer renewed.   End May
     wordt   de Singel aangepakt.
     will be the Singel dealt with.
     The ring road around Antwerp will be renewed this summer. The end of May the Singel will be dealt with.

A human annotator will know that the Singel is likely to be renewed before the ring way, as May is in the spring, and therefore before the summer. A machine will not know that unless it is specified. As May has the value M05, and summer M07/09[25] the machine does know that May is ordered before summer. Of course end of May next year could have been meant, but in that case this would have been said so explicitly.

Notions like *meteorologische zomer* of *weerkundige zomer* (meteorological summer) and *astronomische zomer* (astronomical summer) are as such part of the lexicon with complex entries.

The noise feature is added because people tend to use the names of the seasons in a sloppy way, for example influenced by the weather, as they are not aware of the exact dates at which the seasons change.

Case E:

This is in fact the case we mentioned in section 5.4.1 in expressions like *23 maart 1967 om twintig na drie* (the 23rd of March 1967 at 20 minutes past three). With respect to *drie uur* there are two options: T03 or T15. In case the context doesn't make it clear which one is meant, we will take "T03|15" as its value.

---

[25]The / is used as a symbol meaning 'up to and including'.

	type	solution
A	*morgen over acht dagen*	lexicon
B	*over twee weken*	noise="yes"
C	*ongeveer middernacht*	mod="about"
D	*de herfst*	t-value="M09/12" noise="yes"
	in de ochtend	t-value="T05/13" noise="yes"
E	*om drie uur*	t-value="T03|15" noise="yes"

5.5   Integration of annotation schemas in D-Coi

As already mentioned, within the D-Coi project a choice has been made, with respect to which types of semantic annotation should be developed: annotation of semantic roles as well as of temporal and spatial semantics. At the moment, we keep the two annotation levels separate, to make it easy to produce alternative annotations of a specific type of semantic information without need to modify the annotation at the other level. By keeping the different types of annotation separately, it will be possible to enable progress on techniques for one type of semantic processing without need to wait for the development of high-performing systems for other aspects of semantic interpretation. However, we are aiming at a comprehensive annotation scheme which should ensure compatibility among the various types of semantic information.

Since all linguistic levels interact closely in order to determine the meaning of a whole sentence, the meaning of an expression will be characterized not only by its word meanings, but also by the manner in which they are put together: syntactic structure plays thus a relevant role. In the D-Coi project, the two different types of semantic annotations will be carefully integrated with the other layers of annotation, that is syntactic and morphosyntactic. Allowing semantic annotation to proceed in parallel with the other levels of annotation is a great advantage. There are several examples of treebanks which were extended with semantic information at a later stage such as PropBank or the Prague Dependency Treebank. While these additions are possible, they are not trivial since they often require modifications in the previous annotations, such as changing the labels or some design principles. With D-Coi, we are in the privileged position of developing these annotations in parallel, taking thus into account possible interactions and being able to exploit the available information. The guidelines developed in this respect can constitute the basis for further research as well as a reference for similar initiatives.

In particular, our input sentences are syntactically analysed in another layer using the Alpino parser,[26] in this way the meaning of an entity (expression, sentence) will not only be characterized by the meaning of the constituting words, but also by the manner in which these are put together. Note that for example in temporal semantics the interaction of a verb and other constituents (especially their prepositional and/or nominal heads) is crucial in order to decide whether the verb is refering to a bounded or an unbounded event. PoS information is also still available at the syntactic level, for example with respect to

---

[26]http://odur.let.rug.nl/~vannoord/alp/

temporal information associated with the verbal forms. Similarly, in the case of semantic role labelling, the syntactic structure encoded will guide the assignment of the role labels, this is the case in the distinction between arguments and modifiers.

5.5.1 Conclusion

Our work on both types of semantic annotation discussed in this paper shows that it is feasible to annotate a rather substantial corpus with this kind of information as well, for example a subset of the 500-million-word corpus mentioned in section 1. Our annotation is designed in a way that other semantic layers, like coreference, negation, lexical sematics, can be added as well.

Such a project fits very well in the international state of affairs, cf. recent efforts in the States by Hovy (cf. his keynote lecture at CLIN 2005, Amsterdam) and Pustejovsky (Pustejovsky, Saurí and Littman 2006) and their groups.

References

Babko-Malaya, O.(2005), *Guidelines for Propbank framers*.

Day, D., Ferro, L., Gaizauskas, R., Hanks, P., Lazo, M., Pustejovsky, J., Saurí, R., See, A., Setzer, A. and Sundheim, B.(2003), The TimeBank Corpus, *Corpus Linguistics 2003*, Lancaster.

Erk, K., Kowalski, A., Pado, S. and Pinkal, M.(2003), Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation, *Proceedings of ACL 2003*.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson, G.(2002), *Instruction Manual for the Annotation of Temporal Expressions*, MITRE Washington C3 Center, McLean, Virginia.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson, G.(2005), *TIDES 2005 Standard for the Annotation of Temporal Expressions*.

Johnson, C., Fillmore, C., Petruck, M., Baker, C., Ellsworth, M., Ruppenhofer, J. and Wood, E.(2002), FrameNet: Theory and Practice.

Kamp, H. and Reyle, U.(1993), *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, Vol. 42 of *Studies in Linguistics and Philosophy*, Kluwer Academic Publishers, Dordrecht, Boston, London.

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

Mani, I., Pustejovsky, J. and Gaizauskas, R. (eds)(2005), *The Language of Time. A Reader*, Oxford University Press.

Marcus, M., Santorini, B. and Marcinkiewicz, M.(1993), Building a large annotated corpus of English: The Penn treebank, *Journal of Linguistics*.

Monachesi, P. and Schuurman, I.(2006), *Semantic Annotation for D-Coi. A manual*, Universiteit Utrecht and Katholieke Universiteit Leuven. version 0.1.

Monachesi, P. and Trapman, J.(2006), Merging FrameNet and PropBank in a corpus of written Dutch, *Proceedings of LREC 2006*.

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J., Moortgat, M. and Baayen, H.(2002), Experiences from the Spoken Dutch Corpus Project, *Proceedings of LREC 2002*.

Pan, F., Mulkar, R. and Hobbs, J.(2006), An Annotated Corpus of Typical Durations of Events, *Proceedings of LREC 2006*, Genoa.

Pustejovsky, J., Saurí, R. and Littman, J.(2006), Argument structure in TimeML, *Workshop on Merging and Layering Syntactic Information*, LREC, Genoa, Italy.

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A. and Pustejovsky, J.(2006), *TimeML Annotation Guidelines, version 1.2.1*.

Setzer, A.(2001), *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*, PhD thesis, University of Sheffield.

Stevens, G.(2006), *Automatic semantic role labeling in a Dutch corpus*, Master's thesis, University of Utrecht.