Introduction

The NWO-programme *Modelling textual organisation: coherence and cohesion* studies the organisation of text into structural units by means of coherence (discourse relations between clausal and larger textual units) and cohesion (lexico-semantic relations between words in textual units). The programme is organised around two related PhD-projects, focussing on coherence and cohesion, respectively. Both projects rely on the availability of an annotated corpus of Dutch texts.

Below, we propose a modification of the first project (on coherence). The original project concentrated on an in depth study of a theoretical issue with practical implications (the structure of discourse relations). In the modified proposal, we emphasize computational issues related to discourse annotation and discourse structure.

In particular, we propose a project which studies coherence in Dutch texts, and the mechanisms that are used to establish coherence. We are interested in the following computational research questions:

- *To what extent does syntactic preprocessing of text contribute to the annotation task?* Syntactic analysis can provide an initial segmentation of the text, and can signal discourse markers. Our hypothesis is that automatic segmentation of text will improve the consistency of segmentation (even if annotators may modify the automatically assigned segmentation), and that signalling of discourse markers will make the annotation process more efficient.

- *To what extent can discourse relations be added to a text automatically?* The use of discourse relations for automatic summarization, coreference resolution, question answering, etc., presupposes that such relations can be added automatically with relatively high accuracy. We will study various systems that have been developed on the basis of corpora for other languages, and develop a system for Dutch. Note that while the statistical analysis and machine learning tasks in such a project are largely language independent, identification of relevant syntactic features and discourse markers is language dependent, and may also depend on the morphological (part of speech tagging) and syntactic (parsing) tools available.

- *Does annotation of discourse relations contribute to automatic summarization of text?* One area where the availability of discourse relations is likely to make a contribution is
automatic summarization. We will study a combination of the system for automatically annotating discourse relations with summarization technology for Dutch. In particular, the Stevin Daisy-project is developing summarization tools for Dutch based on the same parsing technology we plan to employ in the current project. We expect that integration of information about discourse relations in the summarization process, will enable more accurate selection of the parts of the text that need to be included in a summary. We will use the training data of the Stevin Daisy-project (texts and their summaries) to evaluate to what extent discourse relations contribute to accurate summarization.

Part of present proposal is concerned with the production of an annotated discourse corpus, which is a prerequisite for the programme as a whole. The connection with automatic preprocessing and the study of automatic annotation will help in developing the corpus in such a way that it can also be used in other studies (e.g. corpus exploration along the lines of Biber, Connor, and Upton (2007)) and as a starting point for other annotation or natural language processing tasks. Finally, the current theme combines the strength of the two programme supervisors and the research groups within CLCG they are associated with (i.e. discourse and communication and computational linguistics).

In the remainder, we outline the three research questions in more detail, and present a workplan.

**Corpus Annotation**

Large corpora annotated with discourse relations have been developed mainly for English (Carlson, Marcu, and Okurowski, 2003; Wolf and Gibson, 2005; Taboada and Renkema, 2008), but also for languages like German (Stede, 2004). For Dutch, far less material is available. Discourse annotation is hard, and requires explicit guidelines. We can make a head-start by building on the experience gained in previous discourse annotation projects, and other annotation carried out by the project supervisors. Furthermore, important ground work for the current annotation project has already been done. Egg and Redeker (2006) explore the possibilities of automated annotation of discourse structure as foreseen in the current project, and Egg and Redeker (2008) argue for the possibility of underspecification in discourse annotation. In the first year of the programme, a number of texts have already been annotated, using the RST-annotation tool.

1 The role of automatic preprocessing and syntactic analysis may seem limited for the current task. Yet, we believe that it can make an important contribution to the utility of the annotated text. In particular, parsing requires text segmented into sentences and tokenized into words and punctuation. We have ample experience with using the text segmentation and tokenization module of the Alpino parser (van Noord, 2006) for this task. By annotating text that has been segmented systematically, we guarantee that the resulting corpus can also be used as evaluation material for, say, an automatic system for annotating discourse relations which is likely to use the same tools. Furthermore, by presenting segmented text to annotators, we avoid a common source of disagreement, and the annotators can concentrate on the core task, which is the identification of discourse relations. Automatic segmentation is typically highly accurate, but not perfect. Of course, we allow segmentation errors to be corrected (during annotation or before annotation starts).

1 [www.wagsoft.com/RSTTool](http://www.wagsoft.com/RSTTool)
Finally, automatic preprocessing allows discourse markers to be identified and allows sub-sentential discourse units to be identified. Attribution (who said what?, for instance, is a relation that often holds between two parts of a single sentence. Syntactic analysis, as carried out by the Alpino parser, allows the units of this relation to be identified beforehand (see also Skadhauge and Hardt (2005)). The developers of the Potsdam-corpus (Stede, 2004) used a similar methodology, adopting annotation standards that allow multiple annotation tiers to be integrated in the annotation process, and developing annotation tools that allow information from one tier to inform the annotation of a following tier.

Corpus annotation needs to be stored in a neutral and machine readable way, if it is to be used as a source for statistical analysis, as an evaluation corpus, or as a starting point for other annotation tasks. XML is currently almost uniformly used as standard for annotating corpora. We will investigate which annotation tools provide the functionality we need. The RST-annotation tool stores annotation results in XML, but using an ‘in-line’ annotation format, where text is segmented using XML-elements, and attributes are used to encode the discourse relations. The annotation tool MMAX (Müller and Strube, 2006) has also been used in discourse annotation projects. It produces XML with so-called ‘off-line’ annotation, where the text and the annotation are stored in separate files. The advantage of the latter method is that it provides more flexibility, especially for adding several layers of annotation to a single document. We have used MMAX, for instance, for adding coreference annotation to texts that had already been annotated with syntactic structure. We will also investigate which of these tools is most suitable for the current project, or whether a combination is possible.

Automatic identification of discourse relations

By developing explicit guidelines for the annotation of discourse relations in Dutch text, and by assembling a corpus of text annotated consistently with such relations, we provide the ideal conditions for studying the possibilities of automatic annotation of such relations.

Automatic systems typically rely on the presence of discourse markers, cue phrases, etc. Part of speech tagging and syntactic analysis can help in identifying discourse markers more accurately (i.e. the adverb dan is a discourse marker, but not if it is part of the phrase meer dan drie, conjunctions can be used to connect discourse units, but not if the coordinates are NPs, etc), as is illustrated for English by Soricut and Marcu (2003), who develop a discourse parser based on the output of a statistical syntactic parser. To increase labelling accuracy (i.e. the choice of the appropriate relation), statistical models can be added which disambiguate specific discourse markers. Miltsakaki et al. (2005), for instance, use maximum entropy to disambiguate adverbs such as since and while into temporal or causal readings.

An important first step toward automatic annotation for Dutch is found in Timmerman (2007), who outlines a system for annotation of Dutch medical text with discourse relations based on RST. We hope to show that a more accurate system can be developed by using a larger annotated corpus, containing texts from various genres, as starting point. In particular, as the corpus includes texts from different genres (i.e. fundraising letters, expository texts, and reviews), we will be able to investigate to what extent discourse relations are influenced by genre. Furthermore, Timmerman makes only limited use of the Alpino-parser (mainly as a means for finding part of speech tags). We hope to use the dependency analysis of Alpino as a means for recognizing intra-sentential discourse relations (following the suggestions in Skadhauge and Hardt (2005)).
Access to discourse relations seems particularly valuable for automatic summarization systems (Teufel and Moens, 1998; Mani, 2001; Thione et al., 2004), and the connection between the two has been explored in, among others, Marcu (2000) and more recently, Bosma (2008) (in the context of a question answering system). Summarization has also applied notions of lexical cohesion and lexical chains (Morris and Hirst, 1991), which is the topic of the other PhD-project within the programme. Our goal is to investigate to what extent information about coherence relations can be integrated in current summarization technology for Dutch.

We hope to benefit from cooperation with the Stevin-project Daisy, in which the computational linguistics group of the University of Groningen participates. Within the project, a corpus of texts and summaries will be collected, and summarization techniques (salience detection, sentence compression and generation, detection of structural relations) will be adopted to Dutch. The syntactic modules of the system will be based on Alpino, and thus, we foresee that our automatic method for annotation of discourse relations can be integrated or combined with technology developed within Daisy. At the very least, the corpus material collected in that project will provide us with useful evaluation data.

1 Workplan

1. Corpus Preparation and Annotation [Month 1-6]
   - Corpus selection, annotation guidelines, selection and configuration of annotation tool, linguistic preprocessing

2. Corpus Exploration [7-12]
   - Evaluation of the annotation process (annotator agreement, role of preprocessing, granularity of the relations, ...), corpus analysis of the annotated corpus (frequency, distance between arguments of a relation, order of arguments, presence of discourse markers, intra vs inter-sentential relations, etc).

   - using the results of 2. for building a automatic (rule-based or statistical) annotation system, evaluation tools, integration of syntactic analysis, case studies on specific relations (attribution, ...), study unsupervised methods (Marcu and Echihabi, 2002).

4. Discourse relations and summarization [25-36]
   - Using the results of 3. to study the role of discourse relations in automatic summarization of Dutch texts from various genres. For evaluation, we will texts and summaries that are collected within the Daisy-project. To ensure sufficient coverage in all genres we are interested in, summaries may have to be produced for a number of additional texts within the current project.

5. Completion of Thesis [37-48]
References


Stede, Manfred. 2004. The Potsdam commentary corpus. In ACL workshop on Discourse Annotation, Barcelona. ACL.


