Dutch lAnguage Investigation of Summarization technologY
DAISY

STEVIN
Vlaams-Nederlands meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie

1. Project title, Acronym and Abstract

Dutch lAnguage Investigation of Summarization technologY
DAISY

Abstract

Summarization of text is often a necessity when searching and selecting information from document repositories. However, current summarization technology is for a large part restricted to the extraction of sentences. Summarization technology for Dutch is very scarce. The aim of DAISY is to develop and evaluate essential technology for automatic summarization of Dutch informative texts. Innovative algorithms for topic salience detection, topic discrimination, rhetorical classification of content, sentence compression and text generation will be implemented. In addition, a demonstrator will be developed in collaboration with the company Q-Go.

The methods rely on the discourse properties of the texts for the detection of salience and cohesion of content, and on syntactic and functional properties of the sentence constituents in order to identify salient and discriminative content in sentences and clauses. Summarization then takes the form of manipulation (compression) of abstract representations. For the syntactic analysis of Dutch texts which is a prerequisite for the algorithms mentioned above, we build on the Alpino parser. For text generation, no standard tools are available for Dutch. We propose to extend Alpino to include a text generation algorithm to generate Dutch sentences on the basis of an abstract representation. This abstract representation will take the form of an abstraction of dependency structures as developed in CGN and used in Alpino, D-Coi and Lassy.

The developed technology will be made publicly available through a demonstrator. This demonstrator will be a Web-based interface that allows users to summarize sample texts, uploaded documents, or shorts texts which the user enters in a textbox. Three types of output are generated for each input text: 1) A headline type summary; 2) A summary that discusses the main subtopics of a text; 3) A metatag describing the rhetorical role of the text (fragment).

The summarization demonstrator will be tested and evaluated in multiple ways in the QA environment of Q-go on documents in the financial and social security domains. Firstly, the system output will be compared against hand-made abstracts of the documents. Secondly, the effect of adding system-generated headline abstracts on retrieval will be measured. Finally, if suitable training and testing material can be obtained, tests will be done with automated email answering, where the summary of the email is used as input for the Q-go QA system.

The project contributes to the goals of STEVIN by providing essential modules for summarization of Dutch texts, which are incorporable in a wide range of information retrieval applications.

2. Principal Investigator/Co-ordinator

Prof. Dr. Marie-Francine Moens
Interdisciplinary Centre for Law and IT - Department of Computer Science
Katholieke Universiteit Leuven
Tiensestraat 41
B-3000 Leuven
Belgium
e-mail: marie-france.moens@law.kuleuven.be

3. Composition of the Research Team

The research will be carried out by a consortium composed of a Flemish and a Dutch university, and a Dutch company.

Table 1: Consortium.

Short name	Organization	Department	Contact person	URL/Email
K.U.Leuven	Katholieke Universiteit Leuven	Interdisciplinary Centre for Law & IT (ICRI) Department of Computer Science	Prof. dr. M.-F. Moens	http://www.law.kuleuven.ac.be/icri/liir.php marie-france.moens@law.kuleuven.be
RuG	University of Groningen	CLCG/Computational Linguistics	Dr. G.J.M. van Noord	vannoord@let.rug.nl http://www.let.rug.nl/~vannoord
Q-go	Q-go	Research & Development	Ir. B. Bos	http://www.q-go.com bart.bos@q-go.com

The **Katholieke Universiteit Leuven** contributes scientific and development expertise with regard to text summarization. It contributes several summarization modules that will be adapted and integrated in the above summarization technologies.

The **University of Groningen** contributes scientific expertise in the area of robust automatic wide-coverage syntactic analysis of Dutch. It contributes syntactic analysis components and text generation components. In addition expertise with regard to noun phrase coreference resolution and syntactic paraphrasing will be incorporated in the project.

Q-go contributes commercial expertise with respect to application domains and will deliver data, integrate the text summarization technology in a real life environment, as well as test and evaluate the implemented technology.

The members of the consortium have a shared and complementary interest in the development of text summarization systems for Dutch texts. The project also contributes to a knowledge transfer from the academic to the commercial world by having a company, Q-go, defining application scenarios and evaluating the developed technology. The consortium will meet three times a year in order to exchange ideas and project results. A central server will be set up specifically for the project and be accessible by the three partners.

4. Requested Budget

Duration of the project: 3 years.
Start date: January 1, 2008.
Requested budget: 461 900 €.

Table 2: Requested budget.

Partner	Labour	25% overhead	FTE	Bench fee	Total (exclusive taxes)
K.U.Leuven	120 400 €	30 100 €	1 FTE (doctorandus) * 3 years	29 500 €	180 000 €
RuG	128 000 €	32 000 €	1 FTE * 3 years (AiO)	29 500 €	189 500 €
Q-GO	50 000 €	12 500 €	0.2 FTE * 3 years	9 900 €	72.400 €
IP clearance					20.000 €
Total	298 400 €	74 600 €	2.2 FTE * 3	68 900 €	461 900 €

5. STEVIN Priorities

The project addresses the following STEVIN priorities:

Resources:

Annotated corpora.
Summarization and text generation tools for Dutch, which will be made available under GNU Lesser General Public License.
The generation tools will be incorporated in the Alpino system. Alpino is and will remain available under the GNU Lesser General Public License.

Applications:

Automatic summarization and text generation; extraction of information from monolingual texts. The tools will be integrated in a summarization demonstrator. The summarization demonstrator will be evaluated in Q-go's Question Answering system.

6. Description of the Proposed Research Project

6a. Scientific aspects and innovative power

Essential in summarization is the *reduction of content to its most essential* (*salient*) *constituents* and the *generation of a concise summary text or other representation* (e.g., in the form of concepts) that can be easily and efficiently processed by humans or by machines. Research into automated summarization of text goes back several decades, but becomes increasingly important when information has to be selected from or sought in large repositories of texts. For an overview on text summarization we refer to Radev et al. (2002), Hovy (2002), Moens et al. (2003) and the proceedings of the yearly Document Understanding Conference (DUC) (2000-2007). Many current summarization systems just extract sentences that contain content terms occurring frequently in the text, that occur at certain discourse positions, that contain certain cue terms (e.g., "in conclusion"), or learn the importance of these and other sentence scoring features from a training set of example texts and their summaries. Hence, the state of the art in summarization is still far from truly abstractive summarization, fusion of information from different texts, generalizing content, and producing fluent, sensible abstracts. We see a current research interest in moving beyond extraction towards compressing and generating suitable summary sentences (e.g., Turner & Charniak 2005; Barzilay and Lapata 2006; Clarke & Lapata 2006a; 2006b; McDonald 2006; Galley & McKeown 2007). However, research into summarization of Dutch texts is limited (e.g., Moens, Uyttendaele & Dumortier 1997: summarization of court decisions; Vandeghinste & Tjong Kim Sang 2004 and Vandeghinste & Pan 2004: summarization of speech; Moens, Angheluta & Dumortier 2005: summarization of magazine articles). Studies that integrate into the summarization certain pragmatic communication roles of the content are new.

Aims of the project

The general aim of the project is to develop and implement essential methods and supporting algorithms for summarization of informative texts written in Dutch, and apply and evaluate them with texts in the financial and social security domain that are currently posted on the World Wide Web. More specifically, the aim is to develop novel and robust technologies for 1) Segmentation and salience detection of content; 2) Single-sentence compression and sentence generation; 3) Rhetorical classification of informative text. For testing and evaluation purposes a demonstrator will be built that generates complementary types of summary information: 1) A headline type summary of a single text or text segment; 2) A short textual summary composed of compressed sentences; 3) Metadata that describes the rhetorical role (e.g., procedure, definition) of the text or text segment of which the summary is made. This demonstrator will be made available. The combination of the summaries and the metadata should discriminate a text in a document base by the description of topics and the role of the text (segment) in the discourse. The summaries should assist the question answering system developed by Q-go in the search for precise answers to information queries posed to finance and social security information.

Syntactic Analysis

Text analysis for summarization involves an initial syntactic analysis of the texts. For the syntactic analysis of Dutch texts, we build on the various tools developed in the context of the Alpino parser by RuG. This includes wide-coverage robust analysis at the level of part-of-

speech tags, syntactic chunks, as well as full syntactic analysis. RuG is also a participant in the STEVIN-project on anaphora resolution (COREA), which implies that there is easy access to the coreferent resolution technologies that will be developed in this project. Technology developed by RuG for the syntactical normalization of the texts based on equivalence rules will be applied when needed.

Segment recognition and salience detection

A necessary condition for summarization is the detection of the salience of content and the segmentation of content based on the topic discussed and the roles of the segments in the discourse. In the literature we find several approaches for computing topical salience. In a first approach, which is based on lexical cohesion, lexical chains of the text's terms are built with the repeated terms, their coreferents and related terms (Morris & Hirst 1991). The size of the chain gives an indication of the salience of the topic represented by the representative term of a chain (Barzilay & Elhadad 1999; Chali et al. 2003). Related terms are detected with the help of lexical resources that provide the necessary context for word sense disambiguation.

A second approach relies on a linear topic segmentation (e.g., Hearst 1997; Choi 2000) of the texts in order to detect subtopics. Very often, adjacent sentences are grouped based on shared terms (e.g., clustering of the term vectors of the sentences) or segmentation patterns are learned from segmented examples (e.g., Beeferman, Berger & Lafferty 1999). Linguistic and cognitive literature gives us a number of surface features for detecting the topic of a sentence and for detecting patterns of thematic progression in texts. Recently, thematic patterns have been exploited in salience computation of content and in summarization of English and Dutch texts by creating a (possibly hierarchical) table of contents of a text (Moens et al. 2005; Moens 2006; Branavan & Barzilay 2007).

The above approaches mainly apply to expository texts that describe certain topics and subtopics. Segmentation and topic detection in informative texts that contain, for instance, instructions and procedural content are seldom researched. In the DAISY project we will focus on the text type of informative texts in the domain of finance and social security that are posted on the World Wide Web. We will study their rhetoric, thematic and layout features and build a segmentation tool. We expect that the HTML markup of the pages provide valuable cues. We will detect the topics and their salience within the text and the text segments by adapting technology already developed by the main partner that processes expository texts (Moens 2006). The segmentation regards the recognition of a text's constituents that possibly have a rhetorical role in the discourse (see below).

Extracted salient key terms from each segment (or sentences that contain these key terms) form a baseline summary of a text.

Sentence compression

Another important aspect of summarization is single sentence compression. Sentence compression is a first step in generating an understandable shorter sentence (or statement) that conveys the important content of the original sentence by dropping words from it or by syntactically rearranging the words.

> "SNS Bank heeft maatregelen getroffen voor veilig Internet Bankieren" (SNS Bank has taken measures to perform bank transactions in a save way).

In the context of the discourse, the sentence is reduced to

> "Maatregelen voor veilig Internet Bankieren" (Measures to perform bank transactions in a save way).

Current state-of-the-art work on sentence compression focuses on word removal. An advantage of such an approach is that sentence compression can be seen as a machine learning task such as by using a noisy channel model (Knight and Marcu 2002; Turner and Charniak 2005; Galley and McKeown 2007). Clarke and Lapata (2006b) use integer linear programming and McDonald (2006) a large margin online learning approach.

In DAISY sentence compression will be performed by producing grammatically correct reductions, and choosing the best one with regard to the salience and novelty of its content in comparison with the previous discourse and to the required compression rate. We rely on a syntactic parse of the sentence. There are many possible reductions of the sentence that form syntactically correct sentences, clauses or phrases. The task is to rank these possibilities according to their degree of salience and novelty with regard to the discourse context (cf. recent work of Clarke and Lapata 2006a) and according to the degree of compression wanted.

We will study and compare two approaches. Firstly, learning rules or functions that generate all grammatically correct reductions of a sentence and then compute the best reduction, which is considered as a separate classification problem. Or, we can learn classification rules or functions in an integrated way, i.e., combining features for grammatical well-formedness and salience (cf. the research of Knight and Marcu, 2002; Turner & Charniak 2005). The main difference of our work with the one of these authors is that we will take into account features with regard to saliency and novelty in the discourse context (e.g., based on the topic and focus structure of a sentence, e.g., Hajičová 1994 and Moens 2006) and not solely grammatical features and word sequence (*n*-gram) features of example sentences and their compression. We will use state of the art learning techniques (e.g., support vector machines, memory based learning, maximum entropy classifier, and context dependent classification such as conditional random fields).

The results should also give us insight into the necessity to include knowledge of semantic roles of sentence constituents (Mehay, De Busser & Moens 2005) and of the semantics of noun compounds (Moldovan & Girju 2005) in the compression process.

The compression techniques can be used to compress an important sentence of the text and eventually to generate a headline from it.

Sentence generation

The compression techniques yield abstract dependency structures. The task of the sentence generation module is to produce actual grammatical sentences on the basis of such abstract representations, using the declarative grammar of Alpino as its key knowledge source. Although the Alpino grammar can be used to ensure that well-formed sentences are produced, a further fluency module will be developed to ensure that the sentences that are produced are natural and appropriate. Just as parsing needs a (statistical) disambiguation component to select the appropriate parse from potentially large sets of possible parses, we need a fluency component to select the most appropriate sentence from the set of possible sentences given by the grammar.

In Marsi and Krahmer (2005) dependency structures are manipulated in order to obtain fusion of content in the case of sentences which specify partly overlapping information content. In their case, they note that if such fused dependency structures are expressed without reference to an actual grammar:

> "As expected, many of the resulting variants are ungrammatical because constraints on word order, agreement or subcategorisation are violated."

They propose an *n*-gram language model to filter out ungrammatical sentences, but found that such *n*-gram models often produce an inadequate ranking, and they conclude that

> " [..] the realization model clearly requires more linguistic sophistication in particular to deal with word order, agreement and subcategorisation constraints."

Following up on this suggestion, our proposal is to use the freely available Alpino grammar to guide the generation process, in order that syntactic constraints on word order, agreement and subcategorisation are properly taken into account. Alpino is a wide-coverage grammar for Dutch, defined as a unification-based grammar, in which many insights from HPSG have been implemented (examples are the inheritance hierarchy of lexical types and grammatical rules). There has been a lot of work on text generation for unification grammars. Early work in this tradition includes the semantic-head-driven generation algorithms co-authored by one of the project partners (Shieber et al. 1990). More recent work on which we will base our approach includes Carroll and Oepen (2005).

For the fluency component, we propose to develop a machine-learning method similar in approach to the disambiguation component of the Alpino parser. The disambiguation component of Alpino contains a discriminative maximum-entropy model, trained on the Alpino treebank. For statistical ranking of competing surface realizations of the same content, we propose to implement a similar discriminative maximum-entropy model. Velldal and Oepen (2006) show that a discriminative maximum entropy model (with access to structural information) outperforms both a classical *n*-gram model as well as a more sophisticated support vector machine (SVM) classification. Therefore, this choice appears to be attractive since it promises good accuracy, and the required technology is readily available in the Alpino toolset.

Rhetorical classification

A text may fulfill various pragmatic communication roles. For instance, it may describe a procedure, inform about a fact, or give a definition. Such roles are signaled by certain rhetorical linguistic cues. It is important to type a text (segment) according to its rhetorical function, as such typing has been proven a valuable part in summarizing textual content (Moens & Teufel 1999, Hachey & Glover 2005). In this project, we use rhetorical typing in order to answer certain types of questions with text to which a suitable role is attached in a question answering system. Rhetorical structures of texts have been studied by Mann and Thompson (1988) and used for summarization of expository texts by Marcu (2000).

This research extends previous work on text segmentation. We will study the text corpus further manually. Based on the literature of discourse theories (see above, possibly complemented by more specific studies: e.g., Kosseim & Lapalme 2000), we will define a limited, but important set of rhetorical types that are characteristic of the informative texts (e.g., definition, procedure,

example, goal, …) and that also correspond to the types of questions with which people interrogate the finance and social security texts (see corpora below). The studies will also provide more refined features that can be considered in role recognition (e.g., syntactical features, lexical items, morphological features such as aspects of verbs). The detected rhetorical roles can be attached as meta-data to texts and their summaries.

> Example of a procedure "Verzenden met EasyStamp" (Send with EasyStamp)
>
> "selecteer het adres of typ postcode en huisnummer in
> kies het gewicht van het poststuk
> selecteer een envelop of etiket (veel soorten en maten zijn al gedefinieerd)
> kies eventueel voor een logo of afbeelding die u mee wilt printen
> druk op de printknop"
>
> (select the address or type postcode and house number
> choose the weight of the mail piece
> select an envelope or label (many types and sizes are defined)
> choose optionally a logo or image that you want to print
> push the print button)

Our research will build further on Moens and Teufel (1999) and Hachey and Glover (2005), who respectively detect the rhetoric function of textual passages in respectively scientific articles and legal cases.

We will train a classifier automatically, based on annotated examples. We will use classifiers that adhere to the maximum entropy principle as they prove to be successful in case of incomplete data. We assume that our training data will be incomplete because we might lack all language patterns that signal rhetorical relations and some rhetorical relations are only implicitly present in the texts. We will also have to deal with ambiguity of rhetorical markers. Context-dependent classification techniques might be useful as the recognition of some rhetorical relations might be dependent on the presence of other relations in the previous discourse.

The project here will process informative texts in the finance and social security domain. We will incorporate and extend the findings of Auoladomar (2005) who studied the properties of procedural texts written in French.

This research will also contribute to a refined segmentation of a text and a more advanced selection of important sentences. The demonstrator will return for each input text (fragment) its rhetorical role.

Detection of the differences between texts

In information processing tasks it is interesting to generate a discriminative summary. This is a summary that makes explicit the differences in content between the target text and the other documents of a set. This is an ambitious goal and existing research on this topic is very limited (e.g., McKeown & Radev 1999; Mani & Bloedorn 1999).

In this project we focus on two types of differences in content: 1) topical differences as reflected by the headlines of a text or text segment; and 2) role differences of the texts.

Given two texts concerning the same topic from the same document set, their respective summaries should be expected to give an indication of the difference between the two. See for example http://www.postbank.nl/ing/pp/page/product/detail/0,2819,1859_309498,00.html and http://www.postbank.nl/ing/pp/page/product/detail/0,2819,1859_642111694,00.html?linktype=int. Both texts provide information concerning "internet banking", but the first text provides general information, while the second focuses on how to print statements. In order to distinguish such documents at first site, it is of imminent importance that the summarization of the latter document mentions the concept of statement printing, in addition to mentioning "internet banking" or "Mijn Postbank.nl".

A second way to distinguish two texts when treating the same topic is by the rhetorical role they play in the discourse. For instance, the topic of two texts might be "internationaal rekeningnummer" (international account number), but the role of one text is providing a definition and of the other text is explaining the procedure of how to acquire an international account number.

Innovative aspects

The novelty of our approach lies in 1) The advancement of the state of the art in segmentation and salient content detection in Dutch informative texts; 2) Improvements of current sentence compression technologies for Dutch texts by considering the discourse context; 3) Development of standard text generation technology for Dutch - integrated with the standard Dutch text analysis tools; 4) Classification of the rhetorical role of a text segment or sentence.

These tasks regard essential tasks in summarization of informative content and are important when more precise answers to information queries have to be found in informative texts. The summarization demonstrator can already be considered as an application. The summarization techniques will be integrated in the question-answering system of Q-Go, but can be integrated in a variety of information search and filtering tools. The issue of the *Communications of the ACM*, 49 (4), 2006 discusses the need for innovative exploratory search, for which summarization of content is of primordial importance.

Demonstrator

The technology developed in the project will be integrated in a demonstrator. The demonstrator will extract essential content in the form of a headline from a single document and expand this very short summary to a small text composed of (possible) condensed sentences. In addition, it will return the rhetorical role of the text (fragment) in the form of a metatag. Figure 1 shows what the interface may look like. The demonstrator should also provide batch mode and verbose output functionality.

The demonstrator will be evaluated in a real setting. The summaries produced by the demonstrator will be integrated in the question answering system of Q-go. The goal is twofold: improve retrieval and shorten the manuel implementation process. In addition, the demonstrator and the Q-go QA system may be combined to form a new application, namely automated email answering.

Currently, Q-go processes user questions based upon a lexical, syntactic and semantic analysis, which results in a formal representation. The application matches such representations against similar representations in a database. These database entries are the result of the linguistic

analysis of manually created "template questions". The template questions are created manually, and each question is associated with an answer, which may be a piece of content on the customer website, or a brief textual answer and a link to the relevant webpage.

We think of the manually crafted templates questions and the short textual answers as one or more summarizations reflecting the gist of the target document, which is why we think that an applied summarization system can replace or at least substantially help a large part of the editorial procedure needed in the current setup. Furthermore, we hope to improve the retrieval by associating automatically created summaries to templates as an alternative for matching. Finally, the demonstrator may serve as a preprocessor for automated email answering: the one sentence summaries that form the headline type output can be considered user questions, and treated similarly.

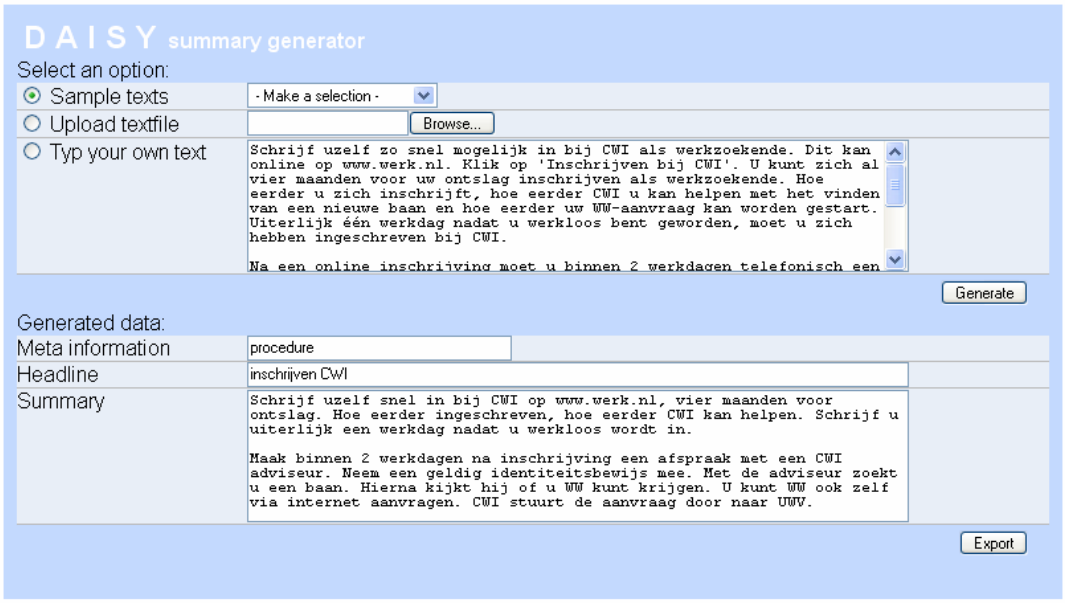

Figure 1. The demonstrator with sample output

Corpora building

During the project several annotated corpora will be built (e.g., for segmentation, sentence compression and rhetorical classification) on which the pattern recognizers can be trained. Whenever possible the corpora will be built automatically (e.g., from existing Dutch texts and their headlines or from texts with example summary sentences). When needed the university partners will engage a job student paid by the bench fee. We will receive a test corpus from the Q-go partner with about 5.000 web pages and short texts, and their hand crafted digestions into short statements or questions.

> Example: http://www.rabobank.nl/info/execute/node?node_id=81936, digested into the main topic “Wat is IBAN?” (What is IBAN?), and for instance the following variants: “Hoe kan ik geld terugboeken in het buitenland?” (How can I credit money in a foreign country”, “Hoe betaal ik met IBAN” (How do I pay with IBAN?)

This corpus will be available for training and testing purposes. A portion of it will be made available for demonstration purposes in the summarization demonstrator. We hope to get the

permission of at least some of the actual clients of Q-go for their data to be transferred to the TST-centrale. A sum has been reserved for IP clearance.

Reuse of existing resources

- Alpino parser and complementary tools developed by RuG.
- Treebanks developed in the contexts of the Alpino parser developed by RuG.
- Feature extraction tool developed by the K.U.Leuven.
- The topic detection and segmentation tool developed by the K.U.Leuven.
- A collocation detection tool developed by the K.U.Leuven.
- Open source pattern classification tools (e.g., Yale, WEKA, OpenNLP, Mallet Package)
- Sentence compression tools developed at the K.U.Leuven: the implementation of the Knight and Marcu algorithm (2002) and the Hedge Trimmer algorithm (Dorr et al. 2003) will be adapted for input of Dutch parsed sentences.

6b. Economic aspects

Because of the information overload in many professional domains, there is a real need for tools that generate summarized overviews of the information (Moens 2000). The different modules that will be designed and developed will satisfy a number of pressing needs in automated summarization, and most importantly will be integrated and tested in a real retrieval and questioning answering (QA) system. Together with the company Q-go, we will build a summarization tool and we will apply it to automatic answer generation, and possibly automatic email answering.

Nevertheless the developed modules can be used in any technology where the presentation of a synopsis of textual content is needed. Compressing and generating content at different levels of detail is useful in many situations where screen sizes demand an unconventional presentation of the information (e.g., very small and very large screens). Summaries are especially valuable for mobile users of information in the selection of and navigation across information.

The number of mobile users of information is constantly increasing. In addition, text sources are increasingly used to automatically annotate images in the training of image content recognizers. Besides information extraction technology, techniques that pinpoint the most salient content in the discourse are important in the alignment of image and text (Deschacht & Moens 2007).

The above situations point to a large demand for summarization technology that can be integrated in search engines, automated information services, information and navigation systems and ubiquitous access to information. The project will develop and evaluate essential modules for summarization of Dutch informative texts. The presence of the company Q-go will ensure that realistic and efficient applications are built and enforce that the standards for a commercial application of the technologies are met.

The project focuses on a number of innovative tracks in summarization technology. As such it will also simulate industrial activity with regard to language technology in the Netherlands and Flanders. When this language technology is integrated in search engines and other information processing technology, it will certainly contribute to the strengthened position of the Dutch language in the world.

6c. Contribution to the STEVIN-programme

DAISY contributes to the STEVIN priorities by:

- Implementing and evaluating innovative and essential technology for the summarization and generation of Dutch informative texts that can be integrated in a wide variety of summarization and text processing applications, such as information retrieval, question answering and navigation of information. The technology will be easily accessible through a web-based demonstrator.
- Development of infrastructure for summarization of Dutch texts (annotation tool according to international standards, annotated texts, test corpora, standard text generation tools).
- The integration of the technology in a real life QA system.

The project mainly focuses on the STEVIN priority *Applications* (*Automatic summarization and text generation applications*), and to a lesser extent contributes to *strategic research* (several innovative approaches are considered) and to *the construction of resources.*

6d. IPR and standards

All results of the project, notably algorithms and software implementations and annotated corpora will be made available as Open Source. In particular, this includes the summarization demonstrator and the text generation modules bundled with Alpino. When we use existing open-source software, which is not developed in this project, we will make sure that it can easily be cleared for commercial licensing.

The software components will be integrated by means of object code and API. The software results of the DAISY project will also be of a modular design, accessible by other software components through a clear and open API.

Q-go will evaluate the summarization demonstrator within its question answering system, so that the technology can extrinsically be tested. The evaluation will be covered by the project, but is irrelevant to the functionality of the module. The code to perform this integration of the demonstrator in Q-go's question answering system will not feature in a deliverable. This is not problematic, since the functionality of the summarization modules including the demonstrator will be fully independent of the Q-go environment. Q-go will demonstrate, for example in the form of a webcast, in which way DAISY and Q-go QA integrate, and how the resulting application works.

The corpora that are built and annotated during the project and which do not have copyright or other restrictions attached will be made available as Open Source after the project. Some of the corpora provided by Q-Go are the propery of clients of Q-Go. We will make an effort to obtain permission to make these corpora available as well.

Annotations will be made available in XML, and will follow standards for encoding text corpora such as TEI and EAGLES and to standards that are already available for Dutch, e.g., CGN, D-Coi and Parole, where applicable.

Table 3. Summary table of IPR issues.

Type of tool	Agreements
Know-how developed by project partners in DAISY including demonstrator	Will be made available under GNU Lesser General Public License.
Interface between demonstrator and QA system of Q-go	Remains property of Q-go.
Background knowledge of partner, which is not open source	Remains property of the partner.
All open source pre-existing know how	Remains open source.
Corpora	Upon IPR clearance, the corpora will be made available to the TST-centrale. An IP clearance fee is foreseen in the budget.

6e. Co-ordination and project management

The management of DAISY will be the responsibility of the project co-ordinator, who is responsible for monitoring the overall progress on the basis of regular reports from each work package, identifying any deviations from the work plan and ensuring that suitable corrective measures are implemented. The project partners will meet three times a year to take important design decisions, to synchronize the efforts, to discuss the project's progress and to collaborate on the evaluation. Standard best practices (web-site, groupware, cvs, bugzilla, etc.) will be used for joint development and communication between the partners, and for dissemination of the results. The co-ordinator will stimulate the research groups to participate in international summarization competitions such as the Document Understanding Conference (DUC) organized by the National Institute of Standards and Technology (NIST), USA, and to present the project results at leading conferences (ACL, HLT, COLING, SIGIR) or in prestigious journals (e.g., Computational Linguistics, Journal of Natural Language Engineering and Information Processing and Management).

6f. Evaluation, validation and success criteria

Progressive evaluation is important, being both intrinsic and extrinsic. With intrinsic evaluation, the system's output is compared with humans' output and their congruence is computed. Extrinsic evaluation on the other hand, measures the quality as needed for other information tasks (e.g., filtering and retrieval). We will test the technology in the domains of finance and social security, where summaries and condense representations of textual content are already available.

We will use as evaluation metrics the ones commonly used at the Document Understanding Conference (such as "Content responsiveness", "Pyramid" and "Rouge"). Defining convenient evaluation measures for the innovative information service here described will certainly be part of the research. Because of the problem of subjectivity of human summarization, wherever possible three or more summaries of the same text will be collected. It is expected that good system-made summaries have a sufficient amount of congruence with at least one of the human-made summaries. The model summaries are obtained from the company Q-go. Very often variant summaries made by different persons are available. In each step, both a baseline approach and the research approach will be compared with the model summary(ies). Also a number of qualitative questions on syntactical correctness and coherence will be answered by manual inspection of the summaries. The qualitative questions will be inspired by questions drafted by the DUC conference program committee.

Evaluations of intermediate steps will also be performed on limited hand-coded test sets (e.g., of the segmentation, salience detection, compression, generation and rhetorical role recognition).

More importantly, the technology will be deployed and evaluated in real life systems that are implemented for Q-go customers from the financial and governmental sector. This extrinsic evaluation is very important. Q-go monitors the recall and precision of its question answering system. These metrics can be expanded with mean reciprocal rank, a classical metric for evaluating QA systems. The testdata can be reused in order to test whether recall, precision and reciprocal rank of the retrieval can be improved by adding automatically generated summaries to the system, or by replacing the hand-made abstracts with system summaries. Q-go will report on the results, and demonstrate the combined system in the form of, for example, a webcast.

A third type of evaluation may consist of a test with automatic email answering. Currently, Q-go analyses user questions and matches them with the linguistic analyses of database questions. If the DAISY system produces high-quality summaries, then Q-go can extend this system to automatic email answering: emails are reduced to single sentence summaries, after which they are processed in the same way as user questions on a web interface. The realization of this experiment depends on the availability of a representative set of emails and (standard) answers. Q-go will attempt to collect such a set from an existing customer. Due to privacy legislation, it is unlikely that the collection will be made publicly available. The deliverable will consist of a detailed test report.

7. Work Programme

WP 1 Management

Tasks:

Setting up the infrastructure for cooperation (server, groupware, website), communication, organization of trimestrial consortium meetings and co-ordinating the dissemination of the results.

Responsible:

K.U.Leuven

Time span:

M1-M36.

WP 2 Corpus building and preprocessing

Tasks:

Building of the necessary corpora; POS-tagging; sentence parsing and coreference resolution.

Partners:

K.U.Leuven
RuG
Q-go

Time span:

M1-M6

Expected risks and alternatives:

The performance of the noun phrase coreference resolution borrowed from the COREA project might be insufficient at early stages of the project. In this case we might only use it in late stages of the project or only have a restricted use of it (e.g., pronoun resolution).

The clients of Q-go do not give permission for making the data publicly available. In this case we will only use the documents for training and evaluation and not include them in the demonstrator.

No suitable collection of emails can be obtained for developing an automated email answering system based on the DAISY summarization tools. In this case we will fall back on intrinsic and extrinsic evaluation on the standard Q-go QA system.

IPR and standards:

Corpora on which there are no intellectual property rights attached will be made available.

WP 3 Segment recognition and salience detection

Tasks:

Informal study of the characteristics of the informative texts; selection of features; design and implementation of the feature extraction; adaptation of existing annotation tool; annotation; training and testing of tool for segmentation; detection of topics and their salience.

Partners:

K.U.Leuven

Time span:

M2-M10

Expected risks and alternatives:

No large risks are expected because we use state of the art technology adapted to the needs of informative texts and to the Dutch language.

WP 4 Sentence compression

Tasks:

Generation of candidate compressions: design and implementation; computation of the most suited compression: Feature selection, training and testing of tool for sentence compression; adaptation of existing implementation of the Knight and Marcu, and Hedgetrimmer

algorithms for Dutch; implementation of the Clarke and Lapata algorithm; implementation of our own algorithm.

Partners:

K.U.Leuven
RuG

Time span:

M6-M24

Expected risks and alternatives:

We expect our approach to perform better than the existing state of the art algorithms for sentence compression. If this would turn out not to be true, we continue with state of the art algorithms.

WP 5 Text Generation

Tasks:

Definition of abstract dependency structures; implementation of text generation algorithm to construct well-formed Dutch sentences on the basis of abstract dependency structures (using Alpino grammar); implementation of fluency component, to select appropriate Dutch sentence from the set of grammatical sentences produced by previous step.

Partners:

RuG

Time Span:

M0-M30

Expected risks and alternatives:

Alpino grammar/dictionary might be too complex as a target for text generation. Alternatives: Use of a subset and/or a variant of the Alpino grammar/dictionary.

WP 6 Rhetorical classification

Tasks:

Study of linguistic and cognitive discourse theories on rhetorical structure applicable on the informative texts; selection of features; design and implementation of the feature extraction; adaptation of existing annotation tool; annotation; training and testing of tool for rhetorical classification.

Partners:

K.U.Leuven

Time span:

M23-M34

Expected risks and alternatives:

Rhetorical signaling cues might be ambiguous or not explicitly present. Only rhetorical roles that are assigned with sufficient certainty will be stored. The late start date of this task is justified because we can work on this task any time when the segmentation is finished (see Figure 2). The results of this task can be integrated in the final demonstrator. It is more important that K.U.Leuven works first on sentence compression, the results of which can be integrated in WP 5.

WP 7 Demonstrator

Tasks:

Design and implementation of the demonstrator: building of a simple Web-based interface and integration of the modules of WP 3, WP 4, WP 5 and WP 6: M33; implementation of Webcast.

Partners:

Q-go

Time span:

M30-M33

Expected risks and alternatives:

There are no expected risks.

WP 8 Evaluation

Tasks:

Intrinsic and extrinsic evaluation of the tasks and subtasks.

Partners:

K.U.Leuven
RuG
Q-go

Time span:

M6-M36

Expected risks and alternatives:

The difficulties of the evaluation steps are described in the concerned work packages.

Figure 2. Dependencies of the WPs.

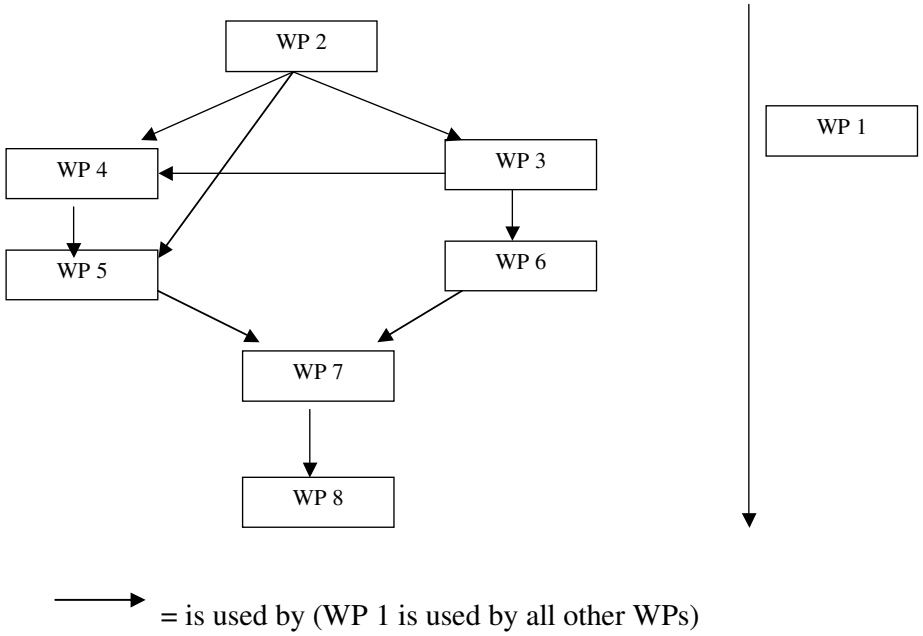

Table3. Table with deliverables, responsible partners, PMs and timing.

Deliverable	**Responsible partner**	**Foreseen number of person months**	**Timing**
WP 2: Corpus building and preprocessing			
Corpora of texts and summaries	Q-go	1	M1
Syntactically parsed corpora	RuG	2	M2
Corpora with coreferences resolved	RuG	2	M6
WP 3: Segment recognition and salience detection			
Report on the feature selection	K.U.Leuven	2	M4
Design and first version of software tool	K.U.Leuven	5.5	M8
Refined version of software tool	K.U.Leuven	2	M10
WP 4: Sentence compression			
Report on the feature selection	RuG	1	M13
Adapted software of the Knight and Marcu, and Hedgetrimmer algorithms for Dutch and implementation of other state of the art algorithms	K.U.Leuven	5	M15
Design and software of our compression algorithm	K.U.Leuven	4.5	M20
WP 5: Generation			
Specification of abstract dependency structures	RuG	2	M6
Initial text generation algorithm	RuG	8	M12
Final text generation algorithm	RuG	4	M18
Specification of fluency component	RuG	2	M21
Implementation of fluency component	RuG	8	M24
Annotated corpus for fluency	RuG	4	M24
WP 6: Rhetorical classification			
Design and first version of software tool	K.U.Leuven	10	M30
Refined version of software tool	K.U.Leuven	3.5	M34
WP 7: Demonstrator			
Demonstrator	Q-go	4	M33
WP 8: Evaluation			
Short evaluation report on the corpus building and preprocessing	RuG	1	M6
Short evaluation report on segment recognition and topic salience detection	K.U.Leuven	0.5	M14
Short evaluation report on sentence compression	K.U.Leuven	0.5	M26
Short evaluation report on rhetorical classification	K.U.Leuven	0.5	M34
Final evaluation report: intrinsic and extrinsic evaluation of the demonstrator results	Q-go	1	M36
Webcast demonstrating the integration of DAISY in Q-go QA	Q-go	1	M36
International publications (throughout the project)	All partners	4.2	M6-M36

We foresee 3 PMs for WP1 (management) financed by the university appointment of the co-ordinator.

8. International Perspective

Summarization is a current research topic at the yearly Document Understanding Conferences (DUC) organized by the National Institute of Standards and Technology, USA. The current

project contributes to this line of research and addresses essential summarization technology that is not yet developed for summarizing Dutch text.

The **Katholieke Universiteit Leuven** has more than 10 years expertise in text summarization research and is author of many international publications on this topic. The summarization tool SUMMA developed at this university has successfully participated in the Document Understanding Conferences in 2002, 2003 and 2004, where for two tasks it obtained a second position. It integrates technologies for table of content generation, a module for the detection of redundant sentences and a sentence compression function. SUMMA processes English and Dutch texts. The sentence compression function is only developed for English. The research group has participated/participates in 18 research projects sponsored by the European Commission, and the Belgian and Flemish governments (Belgian Science Policy, FWO, IWT, IBBT) in the domains of text summarization, information extraction and information retrieval. The research group has very good contacts with the following authorities in the domain of text summarization: Eduard Hovy and Daniel Marcu (University of Southern California, USA), Kathy McKeown (University of Columbia, USA), Simone Teufel (University of Cambridge, UK) and Dragomir Radev (University of Michigan, USA).

The **RuG** has extensive experience and expertise in the automatic syntactic analysis of Dutch. The group recently developed the wide-coverage Alpino parser for Dutch. On news-paper texts, the parser achieves an accuracy of more than 89% per sentence concept accuracy (proportion of correct named dependency triples) (Malouf & van Noord 2004). Using error-mining techniques (van Noord 2004), the same accuracy can be expected, after some effort, for other text genres. A recent overview of the Alpino system is van Noord (2006).

Q-go Amsterdam provides self-service applications for companies with a large customer base, making extensive use of natural language processing techniques. For its customers, including large international companies including Postbank, T-Com, La Caixa, Telefonica, KLM, Q-go software answers hundreds of thousands of customer questions every day, basing the results upon a syntactic and semantic analysis of user input and customer content. The technology has been developed in-house. Q-go spends 30% of its revenue on R&D and employs a number of computational linguists. The choice for Q-go as an industrial partner in the DAISY consortium is motivated by existing deployment opportunities in real life systems running for Q-go customers from the financial and governmental sector.

9. Short CV Principal Applicant(s)

Marie-Francine Moens obtained a Ph.D. in Computer Science in 1999 at the Katholieke Universiteit Leuven. Since 2002 she is an associate professor at this university. She is author of ca. 100 international publications in the fields of text summarization, information extraction and information retrieval, among which are two international monographs. She is a member of numerous program committees of international conferences and workshops. In 2004 she co-chaired an international workshop on text summarization as part of the *42nd Annual Meeting of the Association for Computational Linguistics*. Since 2004, she is member of the roadmap committee of the *Document Understanding Conference* (DUC), organized by the *National Institute of Standards and Technology* in the USA. In 2005 she co-chaired a workshop on question answering at the *Nineteenth International Joint Conference on Artificial Intelligence* (IJCAI 2005) and in 2007 the *Dutch-Belgian Information Retrieval Workshop* (DIR 2007).

Gertjan van Noord is an associate professor (UHD) at the University of Groningen. He was theme-group leader of the NWO Priority Programme on Language and Speech Technology. In 1999, he received an NWO Pionier grant for research on 'Algorithms for Linguistic Processing'. He is the key architect of the Alpino parser for Dutch. He was programme chair of IWPT2003. In 2005 and 2006, van Noord was the chair of the EACL. He has contributed to the STEVIN projects D-Coi, IRME, and LASSY.

Bart Bos is Q-go's Director of Research and Development. After studying at the Technical University in Delft, he has worked seven years at KPN Research. At the New Interactive Services Department, Bart has lead the project to facilitate chipcard transactions over the Internet. He has been working for Q-go since May 2000.

10. Literature

Selection of publications of the partners

Deschacht, K. & Moens, M.-F. (2007). Text analysis for automatic image annotation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, June 23rd–30th*, 2007.

Malouf, R. & van Noord G. (2004). Stochastic attribute value grammars. In *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and Statistical Modeling for Deep Analyses.*

Moens, M.-F. (2000). *Automatix Indexing and Abstracting of Document Texts* (*The Kluwer International Series on Information Retrieval 6*). Boston: Kluwer Academic Publishers.

Moens, M.-F (2006). Using patterns of thematic progression for building a table of content of a text. *Journal of Natural Language Engineering* 12 (3), 1-28.

Moens, M.-F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context* (*The Information Retrieval Series* 21). Berlin: Springer.

Moens, M.-F., Angheluta, R., & De Busser, R. (2003). Summarization of texts found on the World Wide Web. In W. Abramowicz (Ed.), *Knowledge-Based Information Retrieval and Filtering from the Web*. Boston, MA: Kluwer Academic Publishers.

Moens, M.-F., Angheluta, R. & Dumortier, J. (2005). Generic technologies for single- and multi-document summarization. *Information Processing & Management* , 41(3), 569-586.

Moens, M.-F. & Szakowicz, S. (Eds.) (2004). *Text Summarization Branches Out*. New Brunswick: Association for Computational Linguistics.

Moens, M.-F., Uyttendaele, C. & Dumortier, J. (1997). Abstracting of legal cases: The SALOMON experience. In *Proceedings of the Sixth International Conference on Artificial Intelligence & Law* (pp. 114-122). New York: ACM.

Shieber, S.M., Pereira, F. C. N., van Noord, G. & Moore, R. C.(1990). Semantic-head-driven generation. *Computational Linguistics*, 16 (1).

Van Noord, G. (2004). Error mining for wide-coverage grammar engineering. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 447-454). East Stroudsburg, PA: Association for Computational Linguistics.

Van Noord, G. (2006), At last parsing is now operational. In *TALN 2006 Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*. pp. 20-42. Leuven.

Van Noord, G. (2007). Using self-trained bilexical preferences to improve disambiguation accuracy. In: *IWPT 2007*, Prague, Czech Replublic, June 21-22nd, 2007.

International literature on the research of the proposal

Aouladomar, Farida (2005). Towards answering procedural questions. In F. Benamara, M.-F. Moens & P. Saint-Dizier (Eds), *IJCAI-05 Workshop on Knowledge and Reasoning for Answering Questions* (pp. 21-28).

Barzilay, R. & Lapata, M. (2006). Aggregation via set partitioning for natural language geberation. In *Proceedings of NAACL/HLT 2006.*

Barzilay, R. & Elhadad, M. (1999). Using lexical chains for text summarization. In I. Mani and M.T. Maybury (eds.), *Advances in Automatic Text Summarization* (pp. 111-121). Cambridge, MA: MIT Press.

Beeferman, D., Berger, A. & Lafferty, J. (1999). *Statistical models for text segmentation. Machine Learning*, *Special Issue on Natural Language Learning*, 34 (1-3), 177-210.

Branavan, P.D. & Barzilay, R. (2007). Generating a table-of-contents: A hierarchical discriminative approach. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, June 23rd–30th*, 2007.

Carroll, J. and Oepen, S. (2005) *High efficiency realization for a wide-coverage unification grammar.* In: Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP05). Lecture Notes in Computer Science (3651). Springer, Berlin, pp. 165-176.

Chali, I., Kolla, M., Singh, N. & Zhang, Z. (2003). The University of Lethbridge text summarizer at DUC- 2003. In D. Radev and S. Teufel (Eds.), *Proceedings of the Text Summarization Workshop and 2003 Document Understanding Conference May 31 and June 1, 2000* (pp. 148-152). Gaithersburg, MD: NIST.

Choi, F.Y.Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 26-33).

Clarke, J. & Lapata, M. (2006a). Constraint-based sentence compression: An integer programming approach. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics.*

Clarke, J. & Lapata, M. (2006b). Models for sentence compression: : A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics.*

Dorr, B., Zajic, D. & Schwartz, R. (2003). Hedge: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL Text Summarization Workshop and Document Understanding Conference (DUC 2003)* (pp. 1-8). East Stroudsburg, PA: Association for Computational Linguistics.

Galley, M. & McKeown, K. (2007). Lexicalized markov grammars for sentence compression. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technology 2007.*

Hachey, B. & Grover, C. (2005). Automatic legal text summarization: Experiments with summary structuring. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law* (pp.75-84). New York: ACM.

Hajičová, E. (1994). Topic/focus and related research. In P.A. Luelsdorff (Ed.), *The Prague School of Structural and Functional Linguistics* (pp. 245-275). Amsterdam: John Benjamins.

Hearst, M.A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23 (1), 33-64.

Hovy, E.H. (2002). Automated text summarization. In R. Mitkov (Ed.), *Oxford University Handbook of Computational Linguistics*. Oxford, UK: Oxford University Press.

Knight K. & Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139 (2002), 91-107.

Kosseim, L. & Lapalme, G. (2000). Choosing rhetorical structures to plan instructional text. *Computational Intelligence*, Blackwell, Boston 2000.

Mani, I. & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval* (1-2), 35-67.

Mann, W. & Thompson, S. (1988). Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8 (3) (pp. 243-281).

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA*:* The MIT Press.

Marsi, E.C., & Krahmer, E.J. (2005). Explorations in sentence fusion. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG)* (pp. 109-117). Aberdeen, UK.

McDonald, R. (2006). Discriminative sentence compression with soft syntactic constraints. In *Proceedings of the 11th EACL Conference.*

McKeown, K., & Radev, D.R. (1995). Generating summaries of multiple news articles. In E.A. Fox, P. Ingwersen & R. Fidel (Eds.), *Proceedings of the 18th SIGIR Conference* (pp. 74-82). New York: ACM.

Mehay. D., De Busser, R. & Moens, M.-F. (2005). Labeling generic semantic roles. In *Proceedings of the Sixth International Workshop on Computational Semantics* (*IWCS-6*) (pp. 175-188). Tilburg, The Netherlands: Tilburg University.

Moldovan, D. & Girju, R. (2005). Learing the semantics of noun compounds. In H. Bunt, J. Geertzen & E. Thyse (Eds.), *Proceedings of the Sixth International Workshop on Computational Semantics* (*IWCS-6*) (pp. 200-212). Tilburg, The Netherlands: Tilburg University.

Morris, J. & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17 (1), 21-43.

Radev, D.R., McKeown, K., & Hovy, E. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, 28 (4), 399-408.

Teufel, S. & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics* 28, 4, 409-445.

Turner, J. & Charniak, E. (2005). Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.*

Vandeghinste V. & Tjong Kim Sang E. (2004). Using a parallel transcript/subtitle corpus for sentence compression. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 231-234).

Vandeghinste, V. & Pan, Y. (2004). Sentence compression for automated subtitling. A hybrid approach. In *Proceedings of ACL-workshop on Text Summarization. Barcelona* (pp. 89-95). ACL.

Velldal, E. & Oepen, S. (2006). Statistical ranking in tactical generation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, July 2006.