

1 Project Title & Acronym and Abstract

DAESO: Detecting and Exploiting Semantic Overlap

The well-known fact that similar information can be expressed in many different ways is one of the major challenges in building robust NLP applications. It is commonly assumed that such applications can be improved with knowledge of how natural language expressions relate to each other, for instance in terms of paraphrases (same semantic content, different wording) or entailments (one expression implied by the other). DAESO investigates the detection of semantic overlap between Dutch sentences and the exploitation of this knowledge in a range of NLP applications. For this purpose, tools will be developed for the automatic alignment and classification of semantic relations (between words, phrases and sentences) for Dutch, as well as for a Dutch text-to-text generation application which fuses related sentences into a single grammatical sentence, which may be a generalization, a specification or a reformulation of the input sentences. To facilitate development and testing of these tools, an annotated monolingual Dutch parallel/comparable corpus of 1M words will be developed, consisting of pairs of texts that express comparable information. The utility of the resources and tools will be demonstrated in the context of three applications: (1) question-answering systems (improved recall, more complete answers), (2) information extraction (improved recall), and (3) summarization (beyond extraction: sentence compression, sentence fusion, anaphora resolution).

2 Principal Investigator

dr. E. Kraahmer, Tilburg University

3 Composition of the Research Team

Applicants	dr. E. Kraahmer prof. dr. W. Daelemans prof. dr. M. de Rijke drs. J. Zavrel	Tilburg University (UvT), NL Antwerp University (UA), B University of Amsterdam (UvA), NL Textkernel, NL
Personnel	dr. E. Marsi, postdoc dr. V. Hoste, postdoc dr. ir. E. Tjong Kim Sang, postdoc NN, researcher student-assistants	Tilburg University Antwerp University University of Amsterdam TextKernel Tilburg University

There is a long-standing and successful collaboration between the language technology research groups in Antwerp and Tilburg in the field of machine learning for NLP. The various members of the research team have extensive experience with joint collaborations in language technology projects (e.g., the recent IMIX¹ and Prosit² NWO projects). For the current project proposal, the participants have complementary capabilities, all of which are needed. Tilburg has experience with phrasal alignment (Marsi and Kraahmer, 2005b), natural language generation (Marsi, 1998, 2001; Theune et al., 2001), text-to-text generation techniques (Marsi and Kraahmer, 2005a) and anaphora generation (Kraahmer et al., 2003). Antwerp has experience with sentence alignment and summarization (Daelemans et al., 2004), anaphora resolution (Hoste, 2005), and question-answering (Buchholz and Daelemans, 2001). Amsterdam adds extensive research and know-how on Question Answering (Jijkoun et al., 2004), including experience with predicting textual entailment (Jijkoun and de Rijke, 2005). Textkernel is a company, specializing in text Mining and Information

¹<http://www.nwo.nl/imix>.

²<http://ilk.uvt.nl/prosit/>.

Extraction.³ During the project, the members of the research team will communicate through regular face-to-face meetings, as well as via the usual communication channels (e-mail, telephone).

4 Requested Budget

The project will start mid 2006 and will run for 3 years. The estimated costs are as follows:

Position:	Time:	Salary:	Overhead:	Bench fee:	Total:
postdoc UvT	36 months 1.0 fte	179,068	50,988	33,000	263,056
postdoc UA	20 months 1.0 fte	118,099	23,620	13,280	154,999
postdoc UvA	3 months 1.0 fte	15,365	3,841	2,750	21,957
researcher TextKernel	3 months 1.0 fte	15,365	3,841	2,750	21,957
student assistants UvT	3 × 12 months 0,25 fte	24,885	0	0	24,885
Total:					486,853

5 STEVIN priorities

DAESO directly addresses the following STEVIN priorities:

Resources aligned parallel corpora; richly annotated monolingual Dutch corpora

Research semantic analysis (i.e. detecting semantic overlap)

Applications automatic summarization; text generation; Q&A solutions.

6 Description of the Proposed Research Project

6.1 Scientific Aspects and Innovative Power

6.1.1 Problem Statement

Consider the following pair of sentences from news articles on the web:

1. *Intel posted a reward on eBay this month to find a copy of the April 1965 issue of Electronics, in which Gordon Moore predicted that the complexity of computer chips would roughly double every two years.*
2. *Intel had been searching for a mint copy of the issue which contains the Moore's Law article - famously predicting that the number of components on a processor would double every two year with costs falling commensurately.*

These sentences show an example of **semantic overlap**: they express partially overlapping information, but in rather different ways. The well-known fact that similar information can be expressed in many different ways is one of the major challenges in building **robust** NLP applications. It has been shown that such applications can be improved with knowledge of how natural language expressions relate to each other, for instance in terms of **paraphrases** (same semantic content, different wording) or (unidirectional) **entailments** (one expression implied by the other) (e.g., (Echihabi et al., 2005)).

Automatic summarizers, for example, typically rank sentences according to their informativity and extract the top n sentences, depending on the required compression rate. Although the sentences are essentially treated as unrelated, they typically are *not*. Extracted sentences may

³<http://www.textkernel.com/>.

have semantic overlap, resulting in unintended redundancy. This is particularly problematic in the case of **multi-document summarization**, where sentences extracted from related documents are very likely to express similar information in different ways. This issue has been addressed in recent exploratory work on **sentence fusion** (Barzilay, 2003), where semantically overlapping sentences are fused by means of text-to-text generation techniques, thereby eliminating redundant information. Evaluation results show that this improves multi-document summaries (Barzilay and McKeown, 2005).

We think that one of the ways in which this promising line of research on sentence fusion can be extended is to not only detect semantic overlap between expressions, but also to classify the **semantic relation** that holds between them. In the above sentence pair, for example, expressions such as *famously predicted* and *roughly double* are more **specific** than *predicting* and *double* respectively. Given this knowledge, an automatic summarizer can omit specific expressions in favor of general expressions. In combination with omitting non-overlapping information, this may ultimately lead to the generalization:

*Intel had been searching for a copy of the issue which contains the Moore's Law article
- predicting that the complexity of computer chips would double every two years*

The other way round, preferring specific information in combination with including all non-overlapping information would lead to the specification:

Intel posted a reward on eBay this month to find a mint copy of the April 1965 issue of Electronics which contains the Moore's Law article, in which Gordon Moore famously predicted that the complexity of computer chips would roughly double every two years with costs falling commensurately.

The latter could be applied in, for instance, **multi-document merging**, in which the goal is a minimal but lossless merge of documents.

Another area where detecting semantic overlap may help is **Question answering**, especially when the answer is more than a single factoid, but rather a collection of facts (as in the “other” type of questions at the TREC-QA track). The search process of QA may in fact be formulated as detecting semantic overlap between the question and potential answers (Punyakank et al., 2004; Bouma et al., 2005). However, there are additional ways to exploit semantic overlap. QA systems often start with searching for potentially relevant text snippets, which are likely to be related. For instance, given a question like “*When was Moore's Law published?*”, the answer is not found literally in either of the two sentences above. However, if a QA system is able to exploit the alignment between *a copy of the April 1965 issue of Electronics* and *a mint copy of the issue which contains the Moore's Law article*, it can in principle infer the correct answer. Moreover, sentence fusion techniques would allow for combining partial answers to obtain full answers, or selecting more specific answers in favor of more general ones.

Other potential areas of application include NLP tasks such as **coreference resolution** and **word sense disambiguation**, and NLP applications such as **Information Retrieval** and **Information Extraction**, and **Authoring** (intelligent merging of multiple versions of a document). In fact, detecting semantic overlap may well be regarded as a generic NLP task on a par with tasks such as word sense disambiguation and named entity recognition. Although we are convinced that detecting and exploiting semantic overlap is crucial for progress in most areas of NLP, many open questions remain, and in particular for Dutch, resources and tools for detecting semantic overlap are not available.

6.1.2 Research Goals and Results

We propose to develop tools for

1. automatic alignment and classification of the alignment in terms of semantic relations at the level of words, phrases and complete sentences for Dutch

2. text-to-text generation for automatic fusion of semantically overlapping expressions into grammatically correct expressions which constitute specifications, restatements or generalizations of the original input.

Since our approach is fundamentally data-driven, we intend to create a large aligned monolingual corpus of parallel/comparable Dutch text. We also include a substantial component of external evaluation by applying these tools and resources in different applications. More specifically, the DAESO project will address the following resources, tools and applications.

6.1.2.1 Resources

We deliver a **monolingual aligned corpus** of 1M words which contains parallel and comparable text. The parallel texts – book translations and autocue-subtitle pairs – express roughly the same information with a strong relation between the sentences, i.e. approximately the same number of sentences, occurring in approximately the same order. In the comparable texts – news articles and answers from QA system – substantial parts of information expressed in one version may be lacking from the other version, and information may occur in a different order or may be divided differently over sentences.⁴

All corpus material is tokenized, part-of-speech tagged, lemmatized and syntactically parsed. Parallel/comparable text is aligned at the sentence, phrase and word level. Alignments are additionally labeled in terms of elementary semantic relations (see below). Half of the corpus is manually aligned (500k), the other half automatically. The design of the corpus is shown in the following Table:

Text type:	Manual alignment:	Automatic alignment:	Total:
book translation	125k	200k	325k
autocue-subtitles	125k	75k	200k
news articles	225k	200k	425k
answers from QA	25k	25k	50k
Total:	500k	500k	1000k

Book translations The first source for parallel text consists of existing independent modern Dutch translations of the following three books:

1. two translations of *Le Petit Prince* by de Saint-Exupery: (de Saint-Exupère, 2000) and (de Saint-Exupère, 1960)
2. two translations of *The Origin of Species* by Darwin: (Darwin, 2001) and (Darwin, 2002)
3. two translations of *Essais* by de Montaigne: (de Montaigne, 2004) and (de Montaigne, 2001)

Autocue-subtitles The second source for parallel text consists of autocue and subtitles from NOS broadcast news, which is part of the Twente News Corpus. The texts were already aligned at the sentence level in the Atranos project at CNTS in Antwerp⁵. Typically, the subtitles are compressed versions of the autocue, where the former are entailed by the latter. However, since in many cases both sentences are only marginally different, we will use only those sentence pairs which differ in at least 50% of the words.

News articles Comparable text comes from pairs of news articles about the same topic. These are obtained directly from the two main Dutch press offices, ANP and Novum.

⁴To avoid verbosity we will use the term “parallel corpus”, even though it contains comparable text as well.

⁵<http://atranos.esat.kuleuven.ac.be/>.

Answers from QA The last source of comparable text comes from answers from QA systems. We will use two of the QA systems from the IMIX Demonstrator: the ROLAQUAD system from Tilburg University, and the QADR system from the University of Groningen. These systems address questions in the medical domain, especially about Repetitive Strain Injury. We take the n-best answers to ‘open type’ questions (e.g., “What causes RSI?”), where the answer consists of a full sentence (rather than a named entity).

6.1.2.2 Tools

Four related tools will be developed and delivered.

Alignment Annotation Tool Within the Imogen project (part of the IMIX project), a graphical tool has been developed which supports manual alignment of dependency trees, as well as labeling of the alignment relations, by a human annotator; see section 12 for a screenshot. This tool, called Gadget (‘Graphical Alignment of Dependency Graphs and Equivalent Tokens’) will be extended to support alignment of comparable text.

Aligner This tool performs the same annotation automatically. That is, it aligns two related sentences at the word and phrase level, and classifies the alignments in terms of semantic relations.

Paraphrase Extractor Given a corpus of aligned text, this tool automatically extracts a list of paraphrases at the lexical and phrasal level.

Sentence Fuser This application automatically fuses two aligned sentences into a new sentence, which can be a paraphrase, a specification or a generalization of the input sentences.

6.1.2.3 Applications

The usefulness of these tools is externally validated in the context of three, related applications (see also section 6.6).

Summarization At the sentence level, we evaluate to what extent paraphrases extracted from the corpus using the Paraphrase Extractor improve sentence compression for subtitling. In the context of multi-document summarization, we evaluate to what extent sentence fusion improves summaries (in comparison with extracts). Special attention is given to the problems with anaphora that hinder current summarization systems, because anaphors from the source texts may end up lacking an antecedent in the summary. We hypothesize that alignments with parallel texts may be helpful, because an anaphoric expression in one sentence may be fully expressed in the aligned sentence.

Question-answering We investigate to what extent paraphrases and alignment may improve QA output in terms of accuracy. Also we evaluate whether sentence fusion of potential answers improves the quality of QA output compared to normal output.

Information Extraction Similarly, we investigate to what extent the project results improve information extraction on news texts (improved recall).

6.1.3 Approach

6.1.3.1 Alignment and classification of semantic relations

Alignment of parallel text forms the basis of statistical machine translation (Knight and Marcu, 2005) and is an active area of research with many interesting applications (Véronis, 2000). Most work has so far addressed alignment of multilingual text at the sentence and word level. We intend to join recent efforts to align at the syntactic or phrasal level (Pang et al., 2003; Imamura, 2001; Gildea, 2003; Aue et al., 2004; Quirk et al., 2004; Barzilay et al., 1999; Dolan et al., 2004; Herrera

et al., 2005) from the perspective of monolingual parallel text. Our general approach is as follows. Given a pair of related sentences S and S' , both sentences are syntactically analysed, and the resulting trees are automatically aligned. Next, machine learning is used to classify the semantic relation between aligned phrases. The alignment and classification algorithms are developed on the basis of a manually-annotated monolingual parallel corpus.

Following Barzilay (2003) and Barzilay and McKeown (2005), we align sentence pairs in terms of their **dependency structures**. Each node u in the dependency structure for a sentence S is associated with a – possibly discontinuous – string of words $\text{STR}(u)$. Aligning node u from the dependency analysis of S with node v from the analysis of S' therefore indicates that there is a semantic relation between $\text{STR}(u)$ and $\text{STR}(v)$, i.e., between the respective substrings associated with u and v .

In contrast to the earlier work, which only aligns similar phrases, we also classify the semantic relation that holds between these phrases. One of the advantages of this extension is that it enables different kinds of sentence fusion later on. We propose five potential, mutually exclusive, relations between nodes (with illustrative examples, taken from sentences 1 and 2 above):

1. u **equals** v iff $\text{STR}(u)$ and $\text{STR}(v)$ are literally identical
Example: “Intel” equals “Intel”;
2. u **restates** v iff $\text{STR}(u)$ is a paraphrase of $\text{STR}(v)$ (same information content but different wording),
Example: “the complexity of computer chips” restates “the number of components on a processor”;
3. u **specifies** v iff $\text{STR}(u)$ is more specific than $\text{STR}(v)$,
Example: “every two year with costs falling commensurately” specifies “every two years”;
4. u **generalizes** v iff $\text{STR}(v)$ is more specific than $\text{STR}(u)$,
Example: “roughly double” generalizes “double”;
5. u **intersects** v iff $\text{STR}(u)$ and $\text{STR}(v)$ share some informational content, but also each express some piece of information not expressed in the other,
Example: “a copy of the April 1965 issue of Electronics” intersects “a mint copy of the issue”

In an exploratory study (Marsi and Krahmer, 2005a), we have shown that human annotators can align sentences and assign relations to the aligned phrases with a high agreement (an F-score of .95 on alignment, and a κ for relations of .92), indicating that the task is well-defined and feasible. These 5 relations will form the basis of the annotation scheme to be developed in the beginning of the project.

For **automatic alignment** existing tree alignment algorithms can be used (Meyers et al., 1996), where the Dutch part of EuroWordNet (Vossen, 1998) will be used to incorporate information on lexical relations (synonyms, hyperonyms or hyponyms) and information from the parallel corpus to incorporate information on paraphrases.

For the **automatic classification** of the semantic relations, machine learning techniques will be used, where alignments between node pairs are classified on the basis of lexical-semantic relations between the nodes, their corresponding strings and, recursively, on previous decisions about the semantic relations between the respective daughter nodes. A preliminary small-scale investigation (Marsi and Krahmer, 2005b) using the Tilburg Memory-based learning package TiMBL (Daelemans et al., 2004) on a small data set suggests that, given sufficient amounts of labeled data (as will be developed in this project) automatic classification of the 5 aforementioned semantic relations is a feasible task.

Special attention will also be given to the automatic correction of parsing errors. Syntax-based approaches to alignment may be hindered by such parsing errors, and these may propagate through the entire processing stage (e.g., aligning certain phrases may become impossible if the dependency structure is flawed). However, we conjecture that the alignment itself can be used to filter out incompatible parses of parallel text, at least when the two sentences are indeed closely

related. More specifically, the dependency parser can be used in such a way that it does not output the single most plausible analysis, but the N-best ranked analyses. Taking these N-best analyses of both sentences and trying to align all of them may reveal that certain analyses can be aligned better than others, and are thus more likely to be correct.

6.1.3.2 Sentence fusion and text-to-text Generation

Once sentences have been aligned and semantic relations have been assigned, the **sentence fusion** module is able to merge them into a new sentence expressing shared information without introducing redundancy. One of our new contributions in this area is that we classify the semantic relations between aligned phrases, as discussed in the previous section, and on this basis better decisions can be made about which information should be included in the output text. In this way we can output paraphrases, generalizations or specifications of the original input texts (cf. the examples in section 6.1.1). Exploratory research has shown that this is a feasible and promising approach (Marsi and Kraemer, 2005a).

Sentence fusion can be seen as a specific instance of **text-to-text generation**: applications which generate text from text rather than from some semantic or knowledge representation as in more traditional NLG (Reiter and Dale, 2000). An important question in text-to-text generation is how to guarantee that the generated text is grammatical and informationally correct, since detailed linguistic specifications (e.g., regarding tense and aspect) are not straightforwardly available. These problems play a central role in **statistical NLG** (Langkilde and Knight, 1998), which is inspired by the use of statistical language modeling in MT and ASR. It is a hybrid approach which uses a rule-based grammar to generate surface variants and a statistical n-gram model to rank the variants. Filtering generation output with the language model allows the grammar to remain very small and shallow, and also allows underspecified input.

We intend to employ similar techniques for the purpose of sentence fusion, using a shallow grammar to map (merged) dependency trees to surface strings and a statistical language model to filter out ungrammatical variants. The use of n-gram models has been explored to some extent by Barzilay (2003), but more recent work on statistical NLG suggest better results can be obtained with lexicalized or tree-based statistical models (Bangalore and Rambow, 2000; Daume III et al., 2002). A closely related area which we want to explore in this project is that of symbolic machine learning for surface generation (Vargas, 2003). Given the extensive experience with memory-based NLP in Tilburg in Antwerp, we are especially interested in building memory-based NLG systems.

6.2 Economic Aspects

In this project we investigate to what extent automatic alignment and classification of semantic relations may improve applications such as automatic summarization, information extraction, and question answering. These applications are of much interest to companies which deal with large amounts of textual data. A number of such companies in the Netherlands have explicitly expressed their interest in the current project, including TextKernel (one of the applicants). Other companies for which the outcomes of the current project are of interest include the press offices ANP and Novum, the news provider NU.nl (6 million pageviews per weekday), the publishing firms Atlas, Atheneum-Polak & van Gennep, Boom, Donker, and Nieuwezijds, and language technology companies Q-go and AskNow which provide automatic question-answering tools. All these companies take part in the DAESO user group (see Table 1).

6.3 Contribution to the STEVIN-programme

6.3.1 Innovations & Relevance for Dutch

The proposed resources and tools are not yet available for Dutch and fit in well with the Stevin priorities (see section 5). Research on Dutch regarding statistical NLG, sentence fusion, and multi-document summarization is still in its infancy.

Company	Contact-person	Function
ANP	Rob Jacobs	Head ANP-Business
Novum Nieuws	Marcel van de Hoef	Editor in chief
NU.nl	Rogier Swagerman	Editor in chief
Uitgeverij Donker	Willem A. Donker	Publisher
Uitgeverij Atlas	Tamara Doornink	Editor
Uitgeverij Nieuwezijds	Michiel ten Raa	Publisher
Uitgeverij Atheneum-Polak & Van Genneep	Frederieke Doppenberg	Editor
Uitgeverij Boom	W. van Gils	Publisher
AskNow	Rob Boeyink	Computational linguist
Q-go	Stefan Rijnhart	Computational linguist

Table 1: DAESO User group

From an international perspective, our proposal contains various innovations, including semantic classification of alignment relations, using alignments for automatic correction of parsing errors as well as for coreference resolution, and flexible sentence fusion taking semantic relations into account.

In the area of applications, it is worth emphasizing that full sentence fusion of potential answers from a QA system has not seriously been tried and that its contributions too (multi-document) summarization are far from sufficiently explored. Also, although there are a number of (word) alignment tools in the area of MT, there appear to be no publicly available graphical tools for manual alignment of dependency trees or graphs.

6.3.2 Knowledge dissimination

Our primary means for knowledge dissimination will be presentations at scientific conferences (e.g. ACL, EACL, EMNLP, HLT, COLING, CLIN) and publications in proceedings and journals (e.g., *Computational Linguistics*, *Artificial Intelligence*, *Natural Language Engineering*, etc.). Wherever possible (as our work is on Dutch), we will also participate in relevant competitions such as the recent Pascal challenge on recognizing textual entailment.

Secondly, we will inform the members of the user group and other interested parties about the practical aspects of our work on a regular basis. Finally, we will build online demos of the tools developed within the project.

6.4 IPR and Standards

6.4.1 Duplication

There are no parallel *monolingual* corpora for Dutch which are aligned at the phrasal level. The only resource that we are aware of is that of auto-cue and subtitle texts, which were aligned at the sentence level in the Atranos project. There are no graphical tools for automatic alignment and fusion of Dutch text. The systems for question-answering, sentence compression, coreference resolution and information extraction are all existing applications to which we merely add our tools and resources in order to measure their contribution to the performance. The remaining application – multi-document summarizer – has not been developed for Dutch (as far as we know). A number of Dutch research groups work on summarization, but this is limited to extraction-based approaches from single documents.

6.4.2 IPR

The corpus can become available for research purposes via the *TST-centrale*. We have contacted the publishers of all six books mentioned in section 6.1.2.1 and have obtained oral permission to

use their translations for this purpose. A similar permission has been obtained from both press offices, ANP and Novum, as well as from NU.nl for use of their news articles. We are confident that these agreements can be formalized in case the project is funded. The autocue-subtitle material is derived from the Twente News Corpus, which is already available for research purposes. With respect to the answers from the IMIX QA systems: these can be considered as citations from text material (e.g. medical encyclopedia and web sites, which are available for research purposes), without copying or distributing substantial parts of the full text.

Source code of the tools developed during the project can be distributed by the *TST-centrale* under an *Open Source* license approved by the Open Source Initiative⁶ (OSI), preferably under the GNU Public License⁷ (GPL), which guarantees that they become available to the R&D community in a non-discriminative way. We will make an effort to rely on open source software as much as possible (e.g. the Python programming language, the wxWidgets GUI toolkit, the Alpino parser, the Graphviz graph layout program, the MEAD summarization system). However, it is to be expected that sometimes we will have to resort to standard software with a more restrictive license (for instance, in the case of Eurowordnet and the Tilburg Memory-based Learner), in which case we will investigate the possibilities of appropriate licenses.

6.4.3 Standards

For the syntactic analysis we will use dependency trees in the same format as used in the syntactic annotation of (CGN) Spoken Dutch Corpus.⁸ The corpus will be delivered in XML; we will adhere to one of proposed standards like the XCES Corpus Encoding Standard.⁹

6.5 Coordination and Project Management

Daily project coordination and management will be done in Tilburg by the main applicant. Supervision of the student-assistants lies in the hands of the PostDoc in Tilburg. See 3 for distribution of expertise and communication between partners.

6.6 Evaluation, validation and success criteria

Evaluation and validation form an integral part of the project. During the first phase of the corpus annotation, all text will be analysed by multiple annotators and their agreement will be monitored. We aim at a high level of interannotator agreement: F-score $> .90$ on alignment and $> .85$ on semantic relation labelling. Marsi and Kraemer (2005b) have shown that these figures can easily be obtained for parallel translations. The XML format of the corpus will be validated automatically with standard tools.

The main tools (for alignment and sentence fusion) will be evaluated both *internally* (do they perform well?) and *externally* (do they offer an added value in the context of an application?). The internal evaluation of the tool for alignment and semantic relation classification will be performed on the manually aligned part of the corpus using standard cross-validation, aiming at an F-score $> .8$ for both subtasks. The internal evaluation of the sentence fusion module will be done experimentally: human participants judge the *fluency* (is the generated sentence grammatical and 'pleasant' to read?) and the *faithfulness* (is the relevant information from the original sentences correctly represented). This will be measured using standard 5-point Likert-scales, aiming at an average > 4 .

The external evaluation will test the added value of the tools and resources for three applications (summarization, information extraction (IE) and question-answering (QA)), so that we obtain a better understanding of the strengths and weaknesses of the approach (it is likely that

⁶<http://www.opensource.org/>.

⁷<http://www.gnu.org/copyleft/gpl.html>.

⁸<http://www.tst.inl.nl/cgn.htm>.

⁹<http://www.cs.vassar.edu/XCES/>.

the techniques may work better for one application than for another). Existing IE and QA applications will be used (Textractor¹⁰ and Quartz¹¹ respectively). Since no multi-document summarizer exists for Dutch one will be developed in the project, based on the MEAD software.¹² In all three applications, the benefits of the DAESO modules will be evaluated by comparing their performance to a baseline system without these modules. This will be a combination of standard objective and subjective metrics, which is important since these measures do not necessarily correlate; in the case of QA, for instance, it might be that fusion of relevant answer sentences leads to a more complete answer, but is not preferred by users since there no longer is a single source from which the answer derives. For all metrics measuring external performance we aim at a statistically significant improvement over the baseline systems.

7 Work programme

The work programme contains 3 work packages (WPs) consisting of various sub-packages. An overview is given in Table 2. For IPR and standards, see section 6.4, for a risk analysis, see section 7.4.

7.1 WP1: Corpus construction

WP1A: Data collection and preparation

PERIOD year 1: month 1–6

EXECUTOR UvT postdoc, student assistant

TASK The text material to be included in the corpus (see section 6.1.2.1 for details) is obtained in electronic form from the providers. All text material is automatically tokenized. Next, the Alpino parser is used for POS tagging, lemmatization and dependency parsing (Bouma et al., 2001). Sentences from the parallel translations are automatically aligned at sentence level using standard algorithms (Manning and Schütze, 1999), autocue-subtitle pairs are already aligned). Sentences from the comparable material are aligned using MEAD clustering tools, which allows each sentence to become aligned with a (possibly empty) set of similar sentences. All alignments at sentence level are manually checked (student assistant). The Gadget annotation tool is adapted, optimized, properly documented, and ported to Windows (to facilitate annotation by student assistants). A first version of the annotation guidelines will be written, describing the task and discussing reference examples of the relevant notions.

DELIVERABLES A validated XML corpus (1M words), all sentences are aligned and syntactically analysed. Final, documented version of the Gadget annotation tool.

WP1B: Manual annotation

PERIOD year 1: month 7–12

EXECUTOR student assistants & UvT postdoc

TASK Student assistants are employed for the manual annotation task: alignment of words and phrases and semantic classification (500k words). The corpus is annotated iteratively: the first 10% (5k words, evenly distributed over the four components) is annotated by at least two annotators and it is checked iteratively whether the inter-annotator agreement is sufficiently high. Following this, the remainder of the corpus will be annotated, without overlap between annotators.¹³

¹⁰<http://www.textkernel.com/news.php>.

¹¹<http://ilps.science.uva.nl/~qa/>.

¹²<http://www.summarization.com/mead/>.

¹³Annotation is estimated to take 12 to 15 weeks, but student assistants cannot be employed full time.

DELIVERABLES A high-quality manual annotation of half the corpus (500k words). A conference paper is written to present the corpus to the community.

WP1C: Automatic annotation

PERIOD year 2: month 11

EXECUTOR UvT postdoc

TASK Using the tools for automatic alignment and semantic label classification developed in WP2A, the second half of the corpus (500k words) is automatically annotated.

DELIVERABLES An automatic annotation of the remaining part of the corpus (500k words).

WP1D: Corpus delivery

PERIOD year 2: month 12

EXECUTOR UvT postdoc

TASK The final versions of the corpus and the Gadget annotation tool are documented, packed, and delivered to the *TST centrale*.

DELIVERABLES Final, public release of Gadget tool and corpus.

7.2 WP2: Sentence fusion

WP2A: Alignment and labeling software

PERIOD year 1: month 8–12; year 2: month 1–10

EXECUTOR UvT postdoc

TASK The module for automatic alignment and labeling first aligns the respective dependency subtrees. Starting with the dynamic programming approach by Meyers et al. (1996), we experiment with alternative approaches from the large body of work on tree and graph alignment (Bunke, 2000; Shasha et al., 2002; Bille, 2003). Special attention will be paid to the alignment of negations, scope elements and anaphoric references. The second part of the module labels the alignments between dependency subtrees in terms of the semantic relations, using a suitable machine learning algorithm (e.g. memory-based learning) on the manually labeled corpus. The last months of this period are explicitly reserved for performance evaluations as well as modifications and optimizations based on the evaluation results.

DELIVERABLE Documented and evaluated automatic alignment and semantic labeling software. A publication which explains the techniques and evaluates their performance.

WP2B: Paraphrase extraction tool

PERIOD year 2: month 1–3

EXECUTOR UvT postdoc

TASK We implement a module to automatically extract paraphrases from the aligned corpus, with an emphasis on multiword paraphrases, following an approach similar to (Lin and Pantel, 2001).

DELIVERABLE A tool to extract paraphrases from an aligned corpus.

WP2C: Merging and generation software

PERIOD year 3: month 1 –11

EXECUTOR UvT postdoc

TASK Once alignment and labeling is accomplished, a second module is responsible for merging and generation. Merging may involve modification of one dependency tree by substituting and/or inserting subtrees from the other dependency tree as well as pruning. We implement merging procedures supporting restatement, generalization and specification of the original input. The generation part implements the mapping from a merged dependency tree to a grammatically-correct surface expression, following recent advancements in statistical and phrase-based NLG (Langkilde and Knight, 1998; Pan and Shaw, 2005) The quality of the sentence fusion module is evaluated on two aspects: semantic ‘faithfulness’ (does the fused sentence express correct information?) and grammatical well-formedness (is the output sentence grammatically correct?). The software is adapted to take evaluation results into account, and a journal publication is written about the results. Finally, an on-line demo of the software will be developed.

DELIVERABLES Documented and evaluated automatic sentence merging and generation software. A publication.

WP2D: Software delivery

PERIOD year 3: month 12

EXECUTOR UvT Postdoc

TASK The final versions of the alignment and the merging and generation software modules are documented, packed, and delivered to the *TST centrale*.

DELIVERABLES Final, public release of software.

7.3 WP3: Applications

WP3A: Coreference resolution

PERIOD year 2: month 5–10

EXECUTOR UA postdoc

TASK Coreference relations between pronominal, common noun and proper noun nominal constituents are annotated for a part of the translations and news articles. We evaluate to what extent the information from the aligned parallel text contributes to automatic coreference resolution.

DELIVERABLES Publication about the usefulness of the developed resources for coreference resolution.

WP3B: Sentence compression

PERIOD year 2: month 8–10

EXECUTOR UA postdoc

TASK The usefulness of the paraphrase resources and sentence fusion methods for the purpose of sentence compression is investigated, building upon an evaluation environment available at CNTS (developed within the Atranos and MUSA¹⁴ projects).

DELIVERABLES An evaluated sentence compression demo. A publication.

¹⁴<http://sifnos.ilsp.gr/musa/>.

WP3C: Application in Multi-document Summarization

PERIOD year 2: month 11-12; year 3: month 1-12

EXECUTOR UA postdoc

TASK A multi-document summarization environment for Dutch is created, to be used as a baseline system. The environment will be based on the MEAD environment, and will be evaluated to obtain baseline scores. The different resources and tools developed in DAESO will be integrated in the multi-document summarizer, and their respective benefits will be empirically investigated. The results will be used to improve the system, and will also form the basis of a journal publication.

DELIVERABLES An evaluated Dutch multi-document summarizer. An on-line demo. Publication about the added value of the developed tools and resources for multi-document summarization.

WP3D: Application in Question-Answering

PERIOD year 3, month 9-11

EXECUTOR UvA postdoc

TASK The modules for automatic alignment and classification of semantic relations will be integrated in the Quartz QA system for Dutch. It is hypothesized that this will improve the recall (without deteriorating the precision too much). The sentence fusion module will be applied to combine similar, potential answer strings on ‘open’ questions. This is expected to enhance completeness while simultaneously reducing redundancy.

DELIVERABLES A paper about the benefits of the developed tools and resources for QA.

WP3E: Application in Information Extraction

PERIOD year 3, month 9-11

EXECUTOR TextKernel

TASK It will be investigated to what extent recall of the Textractor Information Extraction tool can be improved through alignment and semantic relation classification. Again, the hypothesis is that this may lead to an increased recall, hopefully without a serious decrease in precision.

DELIVERABLES A paper about the benefits of the developed tools and resources for IE.

7.4 Risk analysis

Even though the DAESO project addresses a new field of research, the chances of success are considered to be high since (1) the consortium covers all the required expertise (see section 3) and (2) the project can build on a solid basis developed in the IMIX-Imogen project. While the ideas for the current proposal are all new, there are some results from the Imogen project that can be reused: Gadget, including a preliminary version of an alignment algorithm (inspired by (Meyers et al., 1996)) to speed up manual labeling, an n -gram language models trained for filtering in text-to-text generation.

In addition, a number of existing NLP modules for Dutch will be reused such as the Memory-based part-of-speech tagger (Zavrel and Daelemans, 1999), the Alpino parser (Bouma et al., 2001) for dependency analysis and EuroWordnet to retrieve lexical-semantic relations (Vossen, 1998). For the external evaluation in applications, we can build on existing resources as MEAD (Radev et al., 2004), Quartz (Jijkoun et al., 2004), and Textractor.

Year 1														
ID	Work Package	Exec	1	2	3	4	5	6	7	8	9	10	11	12
WP1A	Data collection and preparation	UvT	x	x	x	x	x	x						
WP1B	Manual Annotation	UvT							x	x	x	x	x	x
WP2A	Alignment and labelling software	UvT								x	x	x	x	x

Year 2														
ID	Work Package	Exec	1	2	3	4	5	6	7	8	9	10	11	12
WP1C	Automatic annotation	UvT											x	
WP1D	Corpus delivery	UvT												x
WP2A	Alignment and labeling software	UvT	x	x	x	x	x	x	x	x	x	x		
WP2B	Paraphrase extraction tool	UvT	x	x	x									
WP3A	Coreference resolution	UA					x	x	x	x	x	x		
WP3B	Sentence compression	UA								x	x	x		
WP3C	Application in multidoc sum	UA											x	x

Year 3														
ID	Work Package	Exec	1	2	3	4	5	6	7	8	9	10	11	12
WP2C	Merging and generation software	UvT	x	x	x	x	x	x	x	x	x	x	x	
WP2D	Software release	UvT												x
WP3C	Application in multidoc sum	UA	x	x	x	x	x	x	x	x	x	x	x	x
WP3D	Application in QA	UvA									x	x	x	
WP3E	Application in IE	TK									x	x	x	

Table 2: Time schedule

One potential risk is that the raw text materials to be included in the corpus are not available in time. To reduce this risk to a minimum, the months before the project would start (if accepted) will be used to formalize the oral agreements and where possible to collect all the required texts in an appropriate electronic format in advance. Two important prerequisites for this project are the feasibility of the annotation task and of the sentence alignment and fusion modules; the pilot work of Marsi and Kraemer (2005a, 2005b) shows that the annotation task can indeed be performed with a high amount of agreement among annotators, and offers encouraging results about the feasibility of the software modules. Arguably, a final risk is that the developed tools and resources, contrary to expectation, might not result in a significant improvement in evaluation results. However, this is precisely one of the important empirical questions underlying this project, and even if it would turn out that the methods are less beneficial for some application than for others, this would generate important insights in the development of such applications.

8 International Perspective

There is a growing awareness in the language technology community of the importance of semantic overlap and paraphrasing, as can be seen for instance from the fact that various recent workshops address issues like the empirical modeling of entailment and paraphrasing (e.g. NLPRS2001, the IWP2003 and IWP2005 international workshops on paraphrasing, the PASCAL textual entailment contest, and the ACL05 Workshop on Empirical Modeling of Semantic Equivalence and

Entailment) and the recent release of the Microsoft Research Paraphrase Corpus¹⁵. In addition, summarization, information extraction and question-answering are all currently active areas of research, as can be witnessed from the number of recent dedicated workshops at conferences such as AAAI, ACL, and CL, addressing these applications, and the number of international competitions in these fields (CLEF, TREC, Pascal, etc.). The members of the current project proposal have excellent contacts in the international community in these fields.

Krahmer works with researchers from the Information Technology Research Institute (ITRI, University of Brighton) and from the University of Aberdeen on natural language generation in the context of an EPSRC project (TUNA¹⁶). Focus of attention of the collaboration are anaphora and co-reference, evaluation of NLG algorithms and text-to-text generation.

In the context of the PASCAL network of excellence (EU 6th framework), Daelemans participates in a project on semantic entailment coordinated by Ido Dagan (Bar Ilan University, Israel), with as further partners XEROX (XRCE, Grenoble) and IDIAP (Switzerland). In the domain of Information Extraction and Summarization, he is or has been involved in two EU 5th framework projects: (i) BioMinT¹⁷ on text mining from biomedical literature in cooperation with the University of Manchester (bioinformatics), the University of Geneva (computer science), the University of Vienna (OEFAI), and the Swiss Institute of Bioinformatics, and (ii) MUSA on sentence simplification for automatic subtitling in cooperation with the BBC, SYSTRAN, ESAT (Leuven), ILISP (Athens), and Lumière (Athens).

In CLEF, De Rijke is coordinator of the tracks on Dutch QA and cross-lingual web retrieval. He and his group participate in TREC, INEX, Senseval and Pascal. He has working contacts with, among others, Dublin (Jones), Edinburgh (Webber), Glasgow (van Rijsbergen), London (Lalmas), Duisburg (Fuhr), Trento (Magnini), Sheffield (Sanderson), Haifa (Dagan), Maryland (Monz) and Pisa (Peters).

9 Short CV Principal Applicants

Dr. Emiel Krahmer is a computational linguist by training. After completing his Ph.D. in 1995 (a version of his thesis was published by *CSLI Publications* in 1998) he worked for six years at the Technical University of Eindhoven (TU/e), mainly on Natural Language Generation. In 2001 he moved to Tilburg University, continuing his research. He has (co-)authored more than 80 publications, including articles in such journals as *Computational Linguistics*, *Natural Language Engineering*, *Speech Communication*, *Speech Technology*, *Journal of Memory and Language*, *Language and Speech*, *Journal of Semantics*, *Journal of Logic, Language and Information*, and the *IEEE Transactions on Professional Communication*. He has been involved in several NWO, EU and EPSRC projects, including a recent project in the IMIX programme (Interactive Multimodal Information Extraction). He is a member of the editorial board of *Computational Linguistics*.¹⁸

Prof. dr. Walter M.P. Daelemans received a Ph.D. in 1987 on an object-oriented model of Dutch morphology and phonology and its applications in language technology. Until 2005 he was affiliated as professor of Machine Learning and Language Technology to the Computational Linguistics group of the faculty of Arts of Tilburg University where he founded the ILK (Induction of Linguistic Knowledge) research group. Since 1999, he is full professor at the University of Antwerp (Linguistics Department), teaching Computational Linguistics and Artificial Intelligence courses. He is also co-director of the CNTS research centre where he is responsible for the language technology projects. The last decade, he has focused in his research on theoretical, methodological, cognitive, and practical aspects of the application of machine learning techniques to natural language processing problems, a research area he

¹⁵http://research.microsoft.com/research/nlp/msr_paraphrase.htm.

¹⁶<http://www.csd.abdn.ac.uk/~agatt/tuna/>.

¹⁷<http://www.biomint.org/>.

¹⁸See <http://fdlwww.uvt.nl/~krahmer/>.

helped pioneer in European computational linguistics. He was awarded the ECCAI fellowship in 2003, is currently associate editor of the *Journal of Artificial Intelligence Research*, is member of the executive board of the international *Association for Computational Linguistics (ACL)*, and has until now supervised 10 PhD theses and (co-)authored more than 200 publications.¹⁹

Prof. dr. Maarten de Rijke is professor of Information Processing and Internet in the Informatics Institute at the University of Amsterdam. He leads the Information and Language Processing Systems group. While young, this group has rapidly established itself as one of the leading research groups in information retrieval in Europe. His current focus is on intelligent web information access, with projects on vertical search engines, question answering, weakly or semi-structured documents, and multilingual information. He currently holds one of the prestigious Pionier grants, has published over 250 papers, has published or edited over a dozen books, is editor for various journals and book series, and coordinates the evaluation efforts of the Crosslingual Web Track at CLEF and of the Dutch Question Answering task within CLEF.²⁰

Drs. Jakub Zavrel worked for a number of years as a researcher in computational linguistics at Tilburg University and Antwerp University before founding the commercial R&D spin-off company *Textkernel* of which he is the CEO. Textkernel is developing new technology to produce advanced information extraction and text understanding engines, and to provide intelligent content management solutions for organisations with large volumes of textual data. Clients of TextKernel include CTB, WiseGuys and Unilever. Textkernel was founded in 2001 and has been profitable since 2003.²¹

Dr. Erwin Marsi wrote his PhD thesis about predicting intonation in spoken language generation systems for Dutch (2001). He worked for three years as a postdoc in the Prosit project on machine learning for prosody prediction. He is currently employed in the IMIX project for developing multimodal QA systems, where his tasks involve speech synthesis output and text-to-text generation. He is also the principal founder and maintainer of the Nextens open source TTS system for Dutch, and was one of the supervisors for the prosodic annotation of the Corpus Spoken Dutch.²²

Dr. Veronique Hoste worked as researcher for the NWO Prosit Project at CNTS (Antwerp University). In 2005 she defended her thesis *Optimization Issues in Machine Learning of Coreference Resolution*. She currently works as a postdoc researcher on the Stevin Core project.²³

Dr. ir. Erik F. Tjong Kim Sang received his PhD from the Computational Linguistics Group of the University of Groningen, The Netherlands (in 1998). He worked as a lecturer at the Linguistics department of Uppsala University in Sweden (1995-1998), before joining the CNTS - Language Technology Group at the University of Antwerp, Belgium in 1998. He was employed in two projects: the European TMR project Learning Computational Grammars (1998-2001) and the Flemish IWT project Automatic Transcription and Normalisation of Speech (2001-2004). Currently he works as a postdoc at the ISLA group of the University of Amsterdam, The Netherlands in the project FactMine (2004-2007) which is part of the NWO-sponsored IMIX programme²⁴

¹⁹See <http://www.cnts.ua.ac.be/~walter/>.

²⁰See <http://staff.science.uva.nl/~mdr/>.

²¹See <http://www.textkernel.com/>.

²²See <http://ilk.uvt.nl/~marsj/>.

²³See <http://www.cnts.ua.ac.be/~hoste/>.

²⁴See <http://staff.science.uva.nl/~erikt/papers/>.

10 Literature

10.1 Selection of publications

1. Daelemans, W., van den Bosch, A., Zavrel, J. (1999), Forgetting exceptions is harmful in language learning, *Machine Learning*, **34**(1-3):11-41.
2. Daelemans, W. and A. van den Bosch (2005), *Memory-Based Language Processing*. Cambridge University Press. 2005.
3. Daelemans, W., Hoste, V., de Meulder F., Naudts B. (2003), Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In: *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, Berlin, Springer, pp. 84–95.
4. van Deemter, K., Krahmer, E. and Theune, M. (2005), Real versus Template-Based Natural Language Generation: A False Opposition?, *Computational Linguistics*, **31**(1): 15-23.
5. Kamps, J., M. de Rijke, and B. Sigurbjörnsson (2005), The Importance of Length Normalization for XML Retrieval, *Information Retrieval*, **8**(4):631-654.
6. Krahmer, E. (1998), *Presupposition and Anaphora*, CSLI Publications/Cambridge University Press, CSLI Lecture Notes Series, Number 89, Stanford, CA.
7. Krahmer, E., S. van Erk, A. Verleg (2003) Graph-based Generation of Referring Expressions, *Computational Linguistics* , **29**(1):53-72.
8. Marsi, E., Reynaert, M., van den Bosch, A., Daelemans, W. and Hoste, V. (2003) Learning to predict pitch accents and prosodic boundaries in Dutch. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 489-496, Sapporo, Japan.
9. Schlobach, S., D. Ahn, M. de Rijke, and V. Jijkoun (2005), Data-driven Type Checking in Open Domain Question Answering, *Journal of Applied Logic*, to appear.
10. Tjong Kim Sang, Erik F. (2002), Memory-Based Shallow Parsing, *Journal of Machine Learning Research* **2**: 559-594.

10.2 International Literature

1. Barzilay, R. (2003). *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Columbia University: PH.D.Thesis.
2. Echihabi, A., U. Hermjakob, E. Hovy, D. Marcu, E. Melz, D. Ravichandran (to appear). How to Select an Answer String? In T. Strzalkowski, S. Harabagiu (eds.), In *Advances in Textual Question Answering*, Dordrecht: Kluwer.
3. Langkilde, I. and K. Knight (1998). Generation that Exploits Corpus-based Statistical Knowledge. *Proceedings of COLING-ACL'98*, Morristown NJ, USA, pp. 704-710.
4. Knight, K. and D. Marcu (2002). Summarization beyond sentence extraction: A Probabilistic Approach to Sentence Compression. *Artificial Intelligence* **139**(1): 91-107.
5. Mani, I. and M. Maybury (1999). *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press.
6. Maybury, M. (2004). *New directions in Question Answering*, AAAI Press.
7. Manning C., H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

8. Pang, B., K. Knight, D. Marcu (2003). *Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences*. Proceedings of HLT-NAACL 2003, Edmonton, pp. 102-109.
9. Reiter, E. and R. Dale (2000), *Building Natural-Language Generation Systems*, Cambridge University Press.
10. Véronis, J. (2000). *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer.

11 Project budget details

Cf. Excel sheet.

References

- Aue, A., A. Menezes, R. Moore, C. Quirk, and E. Ringger (2004). Statistical machine translation using labeled semantic dependency graphs. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.
- Bangalore, S. and O. Rambow (2000). Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany.
- Barzilay, R. (2003). *Information Fusion for Multidocument Summarization*. Columbia University: Ph.D. Thesis.
- Barzilay, R. and K. McKeown (to appear 2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*.
- Barzilay, R., K. McKeown, and M. Elhaded (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, Maryland.
- Bille, P. (2003). Tree edit distance, alignment distance and inclusion. Technical report, IT University of Copenhagen.
- Bouma, G., J. Mur, G. van Noord, L. van der Plas, and J. Tiedemann (to appear 2005). Question answering for Dutch using dependency relations. In *Proceedings of the CLEF 2005 Workshop*.
- Bouma, G., G. van Noord, and R. Malouf (2001). Alpino: Wide-coverage computational analysis of Dutch. In *Computational Linguistics in The Netherlands 2000*, pp. 45–59.
- Buchholz, S. and W. Daelemans (2001). Complex answers: A case study using a WWW question answering system. *Natural Language Engineering* 7(4), 301–323.
- Bunke, H. (2000). Recent developments in graph matching. In *Proc. 15th Int. Conf. on Pattern Recognition*, Volume 2, Barcelona, pp. 117 – 124.
- Daelemans, W., A. Höthker, and E. T. K. Sang (2004). Automatic sentence simplification for subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1045–1048.
- Daelemans, W., J. Zavrel, K. Van der Sloot, and A. van den Bosch (2004). TiMBL: Tilburg memory based learner, version 5.1, reference guide. ILK Technical Report 04-02, Tilburg University.
- Darwin, C. R. (2001). *Het ontstaan van de soorten: door natuurlijke selectie ofwel het bewaard blijven van de rassen die in voordeel zijn in de strijd om het bestaan: de definitieve editie* (6th ed.). Amsterdam: Atlas. Translated by Ruud Rook.
- Darwin, C. R. (2002). *Over het ontstaan van soorten: door middel van natuurlijke selectie, of het behoud van bevooroordeelde rassen in de strijd om het leven*. Amsterdam: Nieuwezijds. Translated by L. Helleman.
- Daume III, H., K. Knight, I. Langkilde-Geary, D. Marcu, and K. Yamada (2002). The importance of lexicalized syntax models for natural language generation tasks. In *Proceedings of the Second International Conference on Natural Language Generation.*, Arden House, NJ,.
- de Montaigne, M. (2001). *Essays*. Amsterdam: Boom. Translated by Frank de Graaff.
- de Montaigne, M. (2004). *De Essays*. Amsterdam: Atheneum, Polak & Van Gennip. Translated by Hans van Pinxteren.

- de Saint-Exupéry, A. (1960). *De kleine prins*. Rotterdam: Donker. Translated by Laetitia de Beaufort-van Hamel.
- de Saint-Exupéry, A. (2000). *De kleine prins*. Rotterdam: Donker. Translated by Ernst van Altena.
- Dolan, W., C. Quirk, and C. Brockett (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING 2004*, Geneva, Switzerland.
- Echihabi, A., U. Hermjakob, E. Hovy, D. Marcu, E. Melz, and D. Ravichandran (2005). How to select an answer string? In T. Strzalkowski and S. Harabagiu (Eds.), *Advances in Textual Question Answering*, Chapter How to select an answer string? Kluwer.
- Gildea, D. (2003). Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the ACL*, Sapporo, Japan.
- Herrera, J., A. P. nas, and F. Verdejo (2005). Textual entailment recognition based on dependency analysis and wordnet. In *Proceedings of the 1st. PASCAL Recognition Textual Entailment Challenge Workshop*. Pattern Analysis, Statistical Modelling and Computational Learning, PASCAL.
- Hoste, V. (2005). *Optimization Issues in Machine Learning of Coreference Resolution*. Ph. D. thesis, Antwerp University.
- Imamura, K. (2001). Hierarchical phrase alignment harmonized with parsing. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, Tokyo, Japan, pp. 377–384.
- Jijkoun, V. and M. de Rijke (2005). Recognizing textual entailment using lexical similarity. In *Proceedings Pascal 2005 Textual Entailment Challenge Workshop*.
- Jijkoun, V., G. Mishne, and M. de Rijke (2004). How frogs built the Berlin Wall. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems, CLEF 2003*, Volume 3237 of *Lecture Notes in Computer Science*, pp. 523–534. Springer.
- Knight, K. and D. Marcu (2005). Machine translation in the year 2004. In *Proceedings of ICASSP*.
- Krahmer, E., S. van Erk, and A. Verleg (2003). Graph-based generation of referring expressions. *Comput. Linguist.* 29(1), 53–72.
- Langkilde, I. and K. Knight (1998). Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th conference on Association for Computational Linguistics*, Morristown, NJ, USA, pp. 704–710. Association for Computational Linguistics.
- Lin, D. and P. Pantel (2001). Discovery of inference rules for question answering. *Natural Language Engineering* 7(4), 343–360.
- Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.
- Marsi, E. (1998). A reusable syntactic generator for Dutch. In P.-A. Coppen, H. V. Halteren, and L. Teunissen (Eds.), *Computational Linguistics in the Netherlands 1997. Selected Papers from the Eight CLIN Meeting*. Amsterdam: Rodopi.
- Marsi, E. (2001). *Intonation in Spoken Language Generation*. Utrecht: LOT.
- Marsi, E. and E. Krahmer (2005a). Explorations in sentence fusion. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG05)*, Aberdeen, Scotland.

- Marsi, E. and E. Kraemer (2005b). Semantic classification by humans and machines. In *ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan.
- Meyers, A., R. Yangarber, and R. Grisham (1996). Alignment of shared forests for bilingual corpora. In *Proceedings of 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, pp. 460–465.
- Pan, S. and J. Shaw (2005, June). Instance-based sentence boundary determination by optimization for natural language generation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, pp. 565–572. Association for Computational Linguistics.
- Pang, B., K. Knight, and D. Marcu (2003). Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *HLT-NAACL*, Edmonton, Canada.
- Punyakanok, V., D. Roth, and W. tau Yih (2004). Mapping dependencies trees: An application to question answering. In *Proceeding of the Eighth International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, Florida.
- Quirk, C., A. Menezes, and C. Cherry (2004). Dependency tree translation: Syntactically informed phrasal smt. Technical Report MSR-TR-2004-113, Microsoft Research.
- Radev, D., T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhu (2004). Mead - a platform for multidocument multilingual text summarization. In *Proceedings of LREC 2004*, Lisabon, Portugal.
- Reiter, E. and R. Dale (2000). *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.
- Shasha, D., J. T.-L. Wang, and R. Giugno (2002). Algorithmics and applications of tree and graph searching. In *Symposium on Principles of Database Systems*, pp. 39–52.
- Theune, M., E. Klabbers, J. R. D. Pijper, E. Kraemer, and J. Odijk (2001). From data to speech: a general approach. *Nat. Lang. Eng.* 7(1), 47–86.
- Vargas, S. (2003). *Instance-based Natural Language Generation*. Ph. D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.
- Véronis, J. (Ed.) (2000). *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers.
- Vossen, P. (Ed.) (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Norwell, MA, USA: Kluwer Academic Publishers.
- Zavrel, J. and W. Daelemans (1999). Recent advances in memory-based part-of-speech tagging. In *VI Simposio Internacional de Comunicacion Social, Santiago de Cuba*, pp. 590–597.

12 Appendix

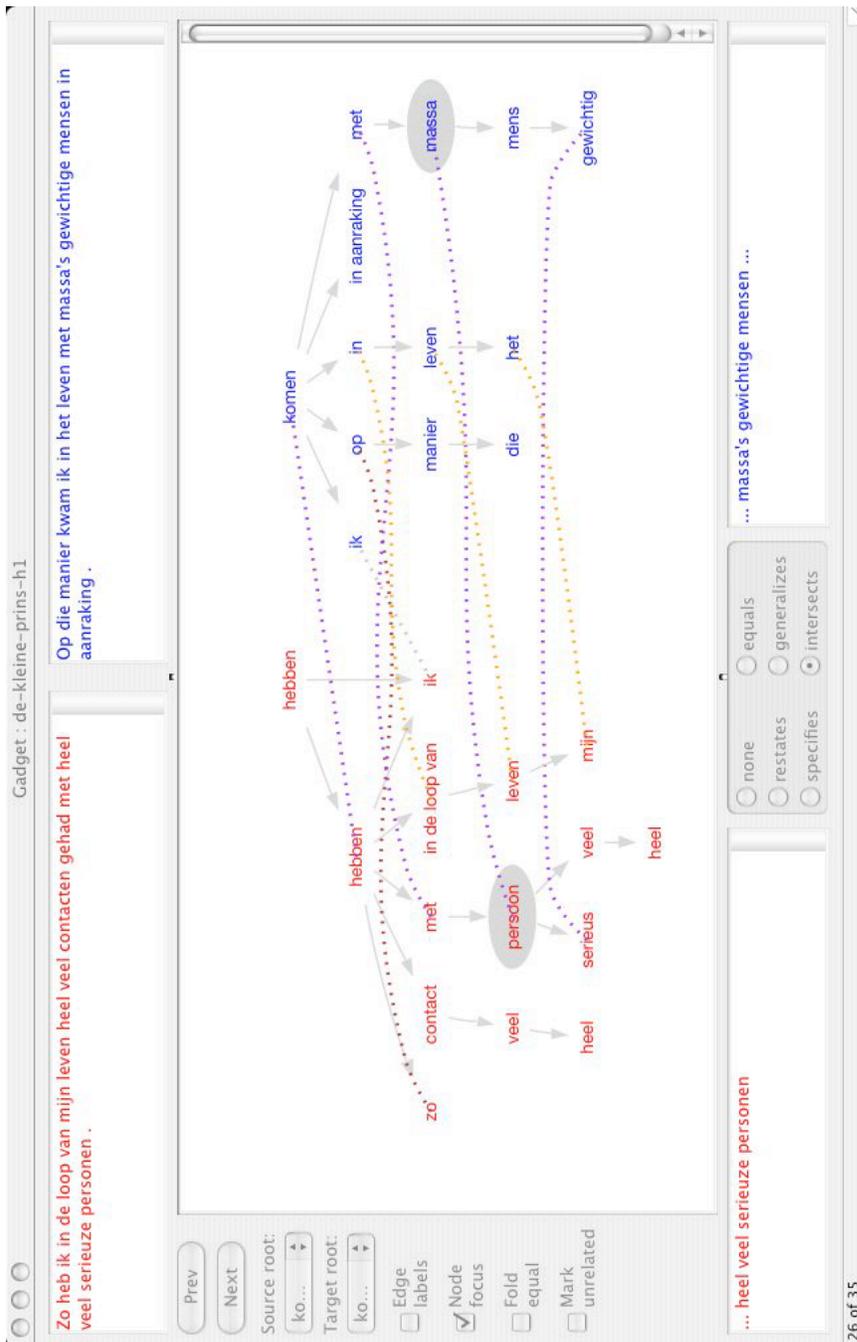


Figure 1: Screen shot of Gadget, the tool used for aligning dependency structures of sentences.