

1 Project Title & Acronym and Abstract

Project title: Dutch Language Online Media Analysis

Acronym: DuOMAn

Abstract: When marketing campaigns or policies on sensitive or broad-ranging issues need to be defined or revised, access to the opinion of the target group is vital. An explosion in online content—both edited and user-generated—has vastly increased the range of opinions potentially available to media analysts and the general public alike, but efficient and effective access methods are needed to unlock this potential. The DuOMAn project will carry out an ambitious research agenda that will result in the development of a set of Dutch language resources and tools for identifying and aggregating sentiments in online data sources. DuOMAn aims to transform the volumes of online information that threaten to leave media analysts information-bound into aggregates of attitudes organized by topic by employing classification, information extraction, and cross-document linking. DuOMAn will provide media analysts and members of the general public with focused access to opinionated information on people, products and topics through an online demonstrator for the general public and through integration of the tools and resources it develops into the workflow of professional media analysts. Key research contributions include sentiment-oriented lexical resources and advancement in the areas of automated sentiment analysis, parsing, and entity detection and coreference resolution. Applied research on robustness and adaptability receives central emphasis. DuOMAn builds on large-scale pilots realized by the hosting institute, which have met with considerable public resonance. DuOMAn will result in resources, natural language processing methods, and public demonstrators that consolidate the resources and tools developed by DuOMAn and will be released for public and professional use.

2 Principal Investigator/Co-ordinator

Prof. dr. Maarten de Rijke (ISLA, Universiteit van Amsterdam) is the principal investigator and co-ordinator.

3 Composition of the Research Team

Role	Name	Affiliation
Applicant	prof. dr. M. de Rijke	University of Amsterdam (UvA)
Professional end user	R. Franz	TrendLight, NL
Lexical tools	T. Spaan	GridLine, NL
Co-applicant	dr. D. Ahn	Postdoc, University of Amsterdam
Software development	NN	Programmer, University of Amsterdam
Syntactic parsing	dr. G. van Noord	Rijksuniversiteit Groningen (RuG)
Coreference resolution	dr. V. Hoste	University College Ghent (HG)

The ILPS group (part of ISLA, Intelligent Systems Lab Amsterdam) of the University of Amsterdam (UvA) coordinates the project and hosts the envisaged demonstrators. TrendLight provides the professional use case for the project, as well as essential training data. The envisaged solution requires the use of syntactic parsing; the Alpino parser developed by the Rijksuniversiteit Groningen (RuG) will be used, which will require domain adaptation. Lexical resources also need to be developed; here, GridLine's expertise in developing and validating lexicons will be used. Finally, coreference resolution modules are needed, which is University College Ghent (HG)'s expertise.

4 Requested Budget

The project duration is 3 years. The envisaged starting date is September 1, 2007. The total requested amount for this project is 440,447 €.

5 STEVIN Priorities and Type of Project

This is an *application-oriented* project, and should be assessed as such. It is aimed at the development of *language technology* to facilitate Dutch language online media analysis. The resources and algorithms to be developed will be integrated in online demonstrators aimed at the general public and in the workflow of TrendLight's media analysis practices. Evaluations will be conducted at the component level (in terms of "traditional" classification and/or retrieval metrics), as part of online demonstrators, and within the context of the envisaged professional end-user—the media analyst. As to the STEVIN priorities addressed by this proposal, in the area of *language technology* they include *text pre-processing* as well as *semantic analysis*. The project also addresses the following priorities in the area of *speech & language applications*: *monolingual or multilingual information extraction* and *other applications*.

6 Description of the Proposed Research Project

We detail the proposed research, describing the main scientific and engineering aspects and organizing these into work packages, but deferring a detailed timelining to §7.

6a Scientific aspects and innovative power

6a.1 Background

Media analysis is undertaken on behalf of a client to determine what the media coverage of the client's domain of activity is and how the client is being portrayed. Media analysts collect output from sources relevant to the domain and look at both the themes and the actors being covered, as well as the nature of the coverage. TrendLight analysts use a coding method based on the “net method” of Kleinnijenhuis (VU) to build a network of support and criticism relations between actors, with regard to specific topics. These relations, together with the roles of the actors involved, are used within the reputation model of Van Riel (EUR) to determine the reputation of the client. At its core, media analysis is concerned with filling the template: “stakeholder X supports/is critical of Y on topic Z ,” where X and Y are actors (individuals or groups) and Z can be just about anything. This information is coded in a database using a dynamic coding scheme that depends on the domain being investigated and current media trends.

The media landscape is changing, and it is doing so in two important ways. First, online news sources are playing an increasingly important role—originally, many were derivatives of traditional news sources (e.g., online versions of newspapers), but today's offering of online news sources is increasingly web specific, i.e., without being a derivative of traditional media counterpart. The second challenge comes from the rise of the participatory web: the traditional consumers of media are turning around to produce it themselves. Blogs, sharing sites, and internet fora all provide ordinary people the opportunity to express their opinions to a global audience—and they are taking advantage of it. These streams of user-generated content provide an opportunity—and challenge—for media analysts: an enormous amount of new data to analyze and a glimpse into the minds of the masses.

Media analysis on a web scale—on both news sources and user generated content—may be impossible without tools that can facilitate or even partially automate the process. The primary research goal of the DuOMAN project is to use technologies from the fields of Language Technology (LT) and Information Retrieval (IR) to develop a sentiment mining toolkit for media analysts who are interested in discovering and measuring the sentiment on the web—both in news sources and in user generated content.

6a.2 Research objectives

DuOMAN aims to support the workflow of the following five-step media analysis methodology (Fig. 1):

1. Data collection and retrieval: collecting documents for analysis, and extracting textual information;
2. Classification: isolating documents relevant to the domain of interest;
3. Coding: identifying categories of information in relevant documents and recording them;
4. Interpretation: interpreting the coded information and writing the research report;
5. Act: determine and/or revise one's strategy based on the interpretation.

This is the media analysis methodology followed by TrendLight, the media analysis partner in the project. Within the DuOMAN project, we will focus on steps 2 and 3. Step 3, coding, is at the heart of the media analysis process. There are three categories of coded information:

- *Actors*: people, companies, government organizations, etc.
- *Themes*: specific aspects of a domain; when coded, these may be related to a particular actor and may also have a *tenor*, i.e., a positive or negative orientation.
- *Opinions*: the opinions of interest to us from the point of view of media analysis are *support* and *criticism*; when coded, these must indicate both the holder and the target of the opinion.

While the identification of opinions is the ultimate goal of the DuOMAN project, the identification of themes is just as important for media analysis, and the identification of actors is an important prerequisite to identifying both of the other categories of information. In the following paragraphs, we lay out how our research goals for the DuOMAN project relate to the core steps 2 and 3 depicted in Fig. 1.

Pre-processing The first component of a media analysis toolkit is responsible for acquiring the documents to be processed and for extracting the textual content of these documents. Our primary aims here are to achieve coverage that is reasonably complete and fresh, while as much as possible eliminating spam and (near-)duplicate documents.

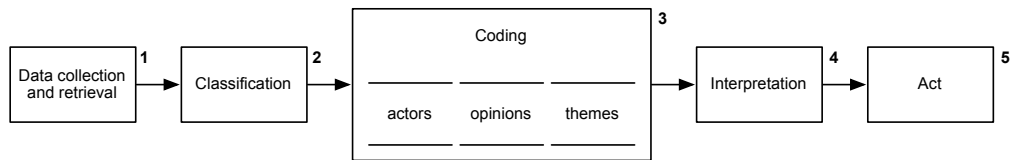


Figure 1: A five-step media analysis methodology.

Linguistic resources In order to build a toolkit for Dutch media analysis, we need to bring together and develop a variety of Dutch linguistic resources. The most important resource that to be built is a sentiment lexicon for Dutch, together with algorithms for extending this lexicon for domain-specific terms using data-driven methods. Tools for actor, or entity, recognition will also be crucial. Entity recognition is an important component of the toolkit as the LT analogue to actor coding, but it will also be a key tool in projecting document-level codes back into documents to generate training data for fine-grained relation extraction. Finally, we plan to use a broad-coverage Dutch parser for syntactic analysis of both edited and non-edited content.

Sentiment analysis The core of the DuOMAN toolkit will be the IR and LT algorithms for identifying sentiment at the document and expression level. The recent outburst of research on sentiment analysis has brought with it a rich terminology. We stick to the following usage, derived largely from [20]. *Subjectivity* relates to an individual’s private state and attitudes and consists of *opinions*, which include beliefs and judgments, as well as affect and emotion. We are concerned with judgments, and in particular, *evaluative* judgments. Beliefs, affect, and emotion lie outside the scope of the project. A *sentiment* is a relation in which a *source* (also: *sender*) has an evaluative judgment regarding a *target* (also: *receiver*). We are especially interested in two types of sentiment: *criticism* (negative) and *support* (positive). We distinguish two types of sentiment expressions: *direct subjective* (also: *explicit*) expression explicitly attributes a sentiment to a source; *expressive subjective* (also: *implicit*) expression conveys a sentiment by the source’s choice of words. The evaluative factor of a word is called *semantic orientation* [18].

The document-level sentiment analysis tasks we will tackle are classifying documents as subjective or objective and of positive polarity or negative polarity. The expression-level sentiment analysis task we will be concerned with is the extraction of support and criticism relations. For the first phase of the project, we will focus on transforming TrendLight’s existing manually assigned codes into document labels and textual annotations suitable for supervised machine learning and developing algorithms for document classification and relation extraction, using the transformed annotations for training. For the second phase of the project, we will apply these algorithms to the blog data that we will be collecting, evaluate their efficacy (through manual assessment), and continue their development.

Aggregation In order to cope with sentiment mined at a Web scale, mined relations must be aggregated for presentation to analysts. The challenge here will be to develop algorithms for cross-document entity coreference, partially on the basis of TrendLight annotations. In addition to aggregation, it will be important to provide human analysts an indication of the salient themes surrounding sentiment peaks. To this end, we will also develop tools based on language modeling techniques for detecting and tracking topics and trends in text.

6a.3 Overall strategy

In this section we present our overall approach to the sentiment analysis problem and break it down into the work packages that will constitute the DuOMAN project. Before doing so, however, we highlight an important issue that affects several work packages: the use of so-called codes provided by TrendLight’s media analysts for annotation purposes and to generate training material for machine learning.

Using document-level annotations. The information coded by a media analyst for a document are distilled from the document on the basis of reading the entire document and are explicitly *not* associated with particular words or phrases occurring in the document. Thus, this information constitutes *document-level* annotations. Examples taken from a recent case study on former minister Rita Verdonk include:

source	source actor type	sentiment	target	theme
Author	letter writer	support	Verdonk	general amnesty
Aboutaleb	director	criticism	Verdonk	political advantage murder Van Gogh

In some cases, as in the first example above, these document-level annotations indicate the author’s sentiment and apply to the document as a whole. We will use such document-level annotations as labels for distinguishing subjective from objective articles and for distinguishing positive from negative articles.

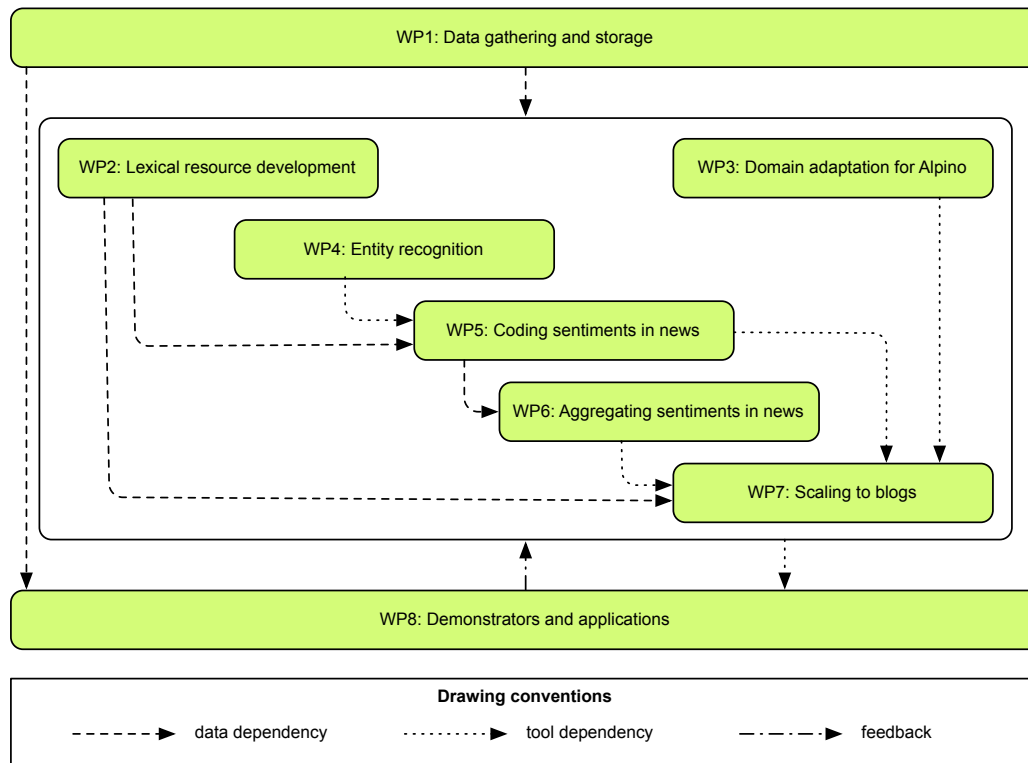


Figure 2: Project organization

In other cases, as in the second example above, document-level annotations may correspond to a specific phrase or sentence in the document. For language processing tasks (e.g., recognizing entities or extracting relations), it will be a bigger challenge to make these annotations usable. Training data for most language processing tasks consists of annotations that are directly associated with spans of text, usually words or phrases. To turn TrendLight’s document-level annotations into span-level annotations requires *projecting* them into the text: determining which span(s) correspond to the annotation by matching coded actor names or descriptions with mentions in the text. Matching named mentions is straightforward, but we may need to match coreferring non-named mentions (particularly to generate training data for relation extraction), so projection will be crucially dependent upon entity recognition.

In sum, an important part of the project will be devoted to turning the coded document-level information produced by media analysts into annotations suitable for machine learning. This transformation will depend on the particular task and on the coding procedure relevant to the task.

Project structure. The core of the envisaged support tools for online media analysis is concerned with steps 2 and 3 in Fig. 1 (“Classification” and “Coding”). Basic approaches to these steps are available from the partners. Making a major leap forward in the sophistication of these tools requires research efforts beyond the state-of-the-art. These efforts are informed by applied research aimed at online demonstrators and support tools for professional end users.

Fig. 2 shows the project’s organisation and Table 1 provides an overview of its work packages. WP2: LINGUISTIC RESOURCE DEVELOPMENT is devoted to linguistic resource development. WP3: DOMAIN ADAPTATION FOR ALPINO adapts the Alpino parser to the domains covered by the project. Recognizing named entities and resolving coreferences are issues to which WP4: ENTITY RECOGNITION is devoted. WP5: CODING SENTIMENTS IN THE NEWS and WP7: SCALING TO BLOGS will use the resources and tools from WP2–WP4. In WP6: AGGREGATING SENTIMENTS IN THE NEWS we collect the information gathered in WP5 and present it in such a way that end users can search and browse it effectively.

Work packages WP2–WP7 are structured so as to minimize the risk of delays in the overall development. Each realizes tools and resources that are passed on to the demonstrator. The interaction between WP2–WP7 is mostly accomplished by the exchange of data and automatically generated metadata, and baseline versions of these resources (or of the tools generating them) will exist from the start of the project.

The demonstrator system will support end users in real-life settings. To this purpose, WP8: DEMONSTRATORS AND APPLICATION consolidates the tools and resources developed in DuOMAN’s R&D activities to create or extend

Table 1: Work package overview

WP nr	WP title	Participants involved	Start month	End month
1	Data gathering and storage	UvA	1	36
2	Lexical resource development	Gridline, UvA	1	36
3	Domain adaptation for Alpino	RUG, UvA	4	12
4	Entity recognition	GridLine, HG, UvA	1	9
5	Coding sentiments in news	Trendlight, UvA	10	17
6	Aggregating sentiments in news	Trendlight, UvA	18	20
7	Scaling to blogs	Trendlight, UvA	21	32
8	Demonstrators and applications	Trendlight, UvA	1	36

online facilities for naive end users; the WP interacts with the media analysis partner in the project (TrendLight BV) to ensure integration of the project’s results in their workflow.

Both the research and development activities and the demonstrators need content—the actual news (and later: blogs). **WP1: DATA GATHERING** addresses this need.

6a.4 Detailed work description

In this section, we provide detailed descriptions of each work package, including plans for *evaluation*, anticipated *risks and remedies*, resources to be *used*, and *deliverables*. Note that the listed deliverables are primarily resources and tools and will, in each case, be accompanied by an explanatory report.

WP1. Data gathering The objective of this work package is to address the selection, acquisition, archiving and indexing of contents of importance to the DuOMAN project and its online demonstrators. Three types of content are essential to the project: news content with document-level annotations by TrendLight, online news content, and blogs. The annotated data from TrendLight for which the UvA group holds a license will be indexed, analyzed, and used for machine-learning experiments at the UvA site; for other TrendLight data, UvA will work on site at TrendLight. UvA’s existing crawling infrastructure, which has been in operation since early 2005 and is being used as part of UvA’s online demonstrators, will be used to obtain online news and blogs. Our aims of completeness and freshness in this task are shared by other ongoing activities at UvA and are assigned high priority by the group.

For indexing and search purposes, DuOMAN will use Indri (<http://www.lemurproject.org/indri/>), an effective and efficient open source language modeling information retrieval tool that supports structured query operators. The UvA group has used Indri in many previous research and demonstration projects, on a wide range of topics, including biomedical search, blog search, expert finding, question answering, and web mining. To store annotations—document-level or word- or phrase-level—the project will use the XIRAF-based multi-dimensional markup framework developed at UvA [1]. This solution facilitates efficient storage and retrieval of multiple levels of annotation, allowing the project to swap any level of annotation for a new one (generated, perhaps, with an improved tagger), without requiring that all annotation be re-generated.

Spam detection—both of spam blogs, or *splogs*, and comment spam—is a research area that will have to be addressed by this work package. Splogs have been estimated to make up 10–20% of the blogspace, and comment spam has been estimated at 40% of blog comments [7, Ch. 2.3.3, Ch. 5.4]. What constitutes spam is not always clear-cut: in the 2006 TREC blog track, more than 10% of retrieved posts from assumed splogs were assessed as relevant [8].

- *Risks & remedies*: **Risk**: Annotated TrendLight data cannot be licensed to UvA. **Remedy**: UvA has a license for the Twente News Corpus for newspaper data; for data for which UvA cannot obtain a license, it will work on site at TrendLight. **Risk**: Online news content owners object to crawling. **Remedy**: Stick to established (non)redistribution protocols.
- *Start month/end month*: 1/36
- *Project partners involved*: TrendLight, UvA
- *Uses*: Indri for indexing; XIRAF for multi-dimensional markup; news aggregators developed for *VerkiezingsKijker*, blog aggregators developed for *BlogKijker*.
- *Deliverables*: (run during the entire project)

D.1.1: Define and set up indexing structure	D.1.2: Acquire annotated data from TrendLight
D.1.3: News aggregation and indexing	D.1.4: Blog aggregation and indexing
D.1.5: Blog data delivery at the project end	

WP2. Lexical resource development The objective of WP2 is two-fold: to develop algorithms for measuring the semantic orientation of Dutch words—i.e., the degree to which they express a positive or negative evaluation—and to use these algorithms to semi-automatically build a domain-independent Dutch sentiment lexicon. This semantic network of sentiment terms will provide information about linguistic properties, semantic orientation labels and thesaurus relations.

Algorithms for automatically measuring semantic orientation of terms fall into two basic groups: WordNet-based algorithms and data-driven algorithms that mine corpora. The basic approach of the first group starts with a small seed set of strongly evaluative terms and uses WordNet synonymy and antonymy relations to propagate semantic orientation from these seed terms to related terms [5, 14]. The second group is more varied and includes supervised methods that use statistical methods to extract potentially subjective terms from documents and sentences manually labeled for subjectivity [19], as well as bootstrapping methods that use web co-occurrence statistics with a small seed set to score terms for semantic orientation [18] or that use extraction patterns to extract terms similar to seed terms [16].

The first step in WP2 will be to use both kinds of methods to add information about semantic orientation to a large set of terms in Dutch WordNet. We plan to do this in a semi-automatic manner, using automatic methods to provide candidate information that will be filtered by humans. The Dutch WordNet is hampered by its relatively small size (44k synsets/56k words vs. 117k synsets/155k words for English), lower density of relations (including near absence of the antonymy relation), and lack of glosses. Nevertheless, it has a structure similar to the English WordNet and is currently being extended [10] so we should be able to adapt algorithms such as [5] to it, albeit at a lower level of efficacy. This expected reduced efficacy is a major reason that we must take a semi-automatic approach. For data-driven methods, we have access both to TrendLight annotations and data as supervision for methods such as [19], as well as large Dutch corpora for bootstrapping methods such as [18] and [16].

In addition to semi-automatic construction of a domain-independent resource, we will experiment with the data-driven methods on domain-specific corpora to evaluate their efficacy for identifying domain-specific sentiment-related terms. Furthermore, GridLine will extend the basic set of sentiment words in WordNet by applying their toolkit for domain-specific vocabulary analysis on the available corpora.

The primary contribution of WP2 will be the semantic orientation labels for a large set of terms in the Dutch WordNet, possibly extending it with additional sentiment-bearing expressions. The secondary contribution will be an evaluation of data-driven methods to extend such labels to domain-specific terms.

- *Evaluation*: Since all of the components to be generate in WP1 will be used in a semi-automatic manner, with a human annotator filtering their output, component evaluation will happen in tandem with the filtering process. Evaluation of the utility of the resulting lexical resources will be part of WP5 and WP7, where they will be used for sentiment analysis.
- *Risks & remedies*: Risk: Automated approaches fail to deliver any useful candidates. Remedy: more human annotation up front to generate larger seed sets for bootstrapping-based automatic approaches (this increased manual work will have the important side effect of providing a larger sentiment lexicon for downstream work packages if nothing else works out).
- *Start month/end month*: 1/9
- *Project partners involved*: GridLine, UvA
- *Uses*: Dutch WordNet, GridLine’s toolkit for domain-specific vocabulary analysis (lexicon building, term extraction, morphological analysis, semantic relation extraction), general domain corpora (e.g., DCoi, CGN, INL-corpora), domain-specific corpora (news, blogs)
- *Deliverables*: (due at Milestone 2, with updates for later Milestones)

D.2.1: Algorithms for measuring semantic orientations of semantic words and phrases, exploiting both analytic and data-driven approaches. (UvA)	D.2.2: Tool implementations of the algorithms from D.2.1. (UvA/GridLine)
D.2.3: Domain-independent sentiment lexicon of at least 4000 terms (UvA/GridLine)	D.2.4: Domain-specific sentiment lexicons, as needed. (UvA/GridLine)

WP3. Domain adaptation for Alpino The objective of this work package is to adapt Alpino to blog material, for improved robustness and accuracy.

Within the DuOMAn project syntactic analysis is necessary for several components. Within the entity recognition task (WP4), syntax-based features such as grammatical functions have proven useful for machine learning of coreference relations [4]. For the sentiment relation extraction task (WP5 and WP7), we expect that syntactic

features will be even more important. [13] use dependency features for extracting sentiment relations from the MPQA corpus, and much of the literature on relation and event extraction attests to the importance of syntax (see, e.g., [12] for relation extraction, and [2] for event extraction).

For syntactic parsing, the project will use Alpino [9]. Alpino is a linguistically motivated, wide-coverage grammar and parser for Dutch in the tradition of HPSG, which outputs dependency structures of the type developed in CGN, D-Coi and LASSY. Heuristics have been implemented to deal with unknown words and word sequences and ungrammatical or out-of-coverage sentences (which may nevertheless contain analysable fragments). The Alpino system includes a POS-tagger which greatly reduces lexical ambiguity, without an observable decrease in parsing accuracy. Disambiguation is performed on the basis of a Maximum Entropy model, trained on a treebank of over 7000 newspaper sentences.

In order to obtain better accuracy on blogs, Alpino will be adapted in various ways. For improved robustness, error mining techniques [9] will be applied to quickly identify which aspects of blogs are difficult to handle for the parser. The system can be adapted as required by adapting its knowledge sources (lexicon, grammar, tokenization component). To improve its disambiguation accuracy, the disambiguation component will be retrained on blog material, which involves manually annotating a collection of 2000 blog sentences. Finally, the POS-tagger will be adapted to blog material.

- *Evaluation*: Comparison of the base-line Alpino system and the adapted Alpino variant. To this end, a further 500 blog sentences will be manually annotated syntactically.
- *Risks & remedies*: No specific risks are foreseen
- *Start month/end month*: 4/12
- *Project partners involved*: RuG, UvA
- *Uses*: Alpino (available under GNU Lesser Public License)
- *Deliverables*: (due at Milestone 3)
 - D.3.1: Annotated corpus of 2500 blog sentences, using CGN/D-Coi/LASSY guidelines
 - D.3.2: Variant of Alpino with all domain adaptations in place
 - D.3.3: Report on the domain adaptations that have been implemented, and a formal evaluation on the annotated test data

WP4. Entity recognition The objective of this work package is the adaptation of trainable components for entity mention detection and coreference resolution in order to identify actors, or entities, that are potential sources or targets of sentiments.

Entity recognition involves two subtasks: entity mention detection and coreference resolution. For entity mention detection, we have a trainable Dutch named-entity tagger which uses machine learned classifiers to identify names and determine the type of entity to which they refer. We will make use of the TrendLight actor coding in two ways: to construct domain-specific gazetteers on the basis of actor and actor type codes, and to project actor codes back into the documents with which they are associated to generate training data.

For this latter task, we will build a basic tool for projecting actor codes to formally identical entities. As a second step, we will enrich the projection tool with linguistic devices for recognizing formally different instances of the same entities, exploiting morphological and syntactic correspondences, as well as semantic information (i.e., synonym, antonym and hypernym relations in the Dutch WordNet). The output of the Alpino parser will also be used to identify non-named entity mentions. As a last step, the tool will be made suitable for code projection in domain-specific documents. The resulting tool will be used to create an annotated training corpus for the tagger.

Coreference resolution is needed to link non-named mentions of a given entity to its named mentions. The machine-learning-based coreference resolution system developed in the STEVIN Corea project will be used as the baseline architecture. The system uses a pairwise classification approach in which distance, morphological, lexical, syntactic and semantic information about the candidate anaphor, its candidate antecedent, and the relation between the two is combined to decide on the presence of a coreferential link between two noun phrases. These pairwise classifications are then clustered into coreferential chains. A modular learning approach is used in which a separate module is developed and optimized per NP type.

The current baseline system will be adapted at different levels. The generic system will be tailored to the output of the entity recognition. Since the current coreference resolution system is trained and optimized on the small Knack news magazine corpus [4], we will first validate its performance against the automatically generated training data, of which a small part will be hand-annotated. In order to develop a system tailored to the task, a blog corpus will be annotated with coreferential relations following the annotation guidelines as developed in the STEVIN Corea

project (estimated time: 25,000 words at 500 words/hour). At the information source level, the system will use the syntactic information resulting from the Alpino parse; it will also be adapted to capture first-person expressions of sentiment, e.g., in direct speech, allowing for number disagreement.

- *Evaluation*: For entity mention detection: fine-grained evaluation using standard metrics against annotations automatically generated by projection and coarse-grained evaluation against TrendLight codes. For coreference resolution: evaluation of the baseline system and the tailored system on a manually annotated corpus (25,000 words). Evaluation of the effect of coreference resolution on the task of sentiment analysis (entity detection versus entity detection+coreference resolution).
- *Risks & remedies*: Risk: Impossible to project actors cleanly. Remedy: Use annotations as gazetteers for entity types, and do additional manual annotation for both entity mentions and coreference.
- *Start month/end month*: 1/9
- *Project partners involved*: GridLine, HG, UvA
- *Uses*: Domain-specific corpora (WP1), Alpino, UvA named-entity tagger, HG coreference resolution tool
- *Deliverables*: (due at Milestone 2)
 - D.4.1: Actor code projection tool. (GridLine/UvA)
 - D.4.2: Entity mention detection system trained on result of D.4.1. (UvA)
 - D.4.3: Corpus (50,000 words) annotated with coreferential relations following Corea guidelines. (HG)
 - D.4.4: Adapted version of the coreference resolution system. (HG)

WP5. Coding sentiments in the news WP5 will tackle the meat of the automated sentiment analysis problem for edited content. The primary objective of this work package is to identify sentiment relations between actors, as coded according to the methodology used by TrendLight’s media analysts.

Sentiment analysis is an active field of research, with a wide range of tasks and methods. We classify this research into work that focuses on identifying the evaluative orientation of subjective text, such as product reviews, etc., and work that focuses on extracting individual sentiments from a range of text types. Work of the first type treats sentiment analysis as classification. The basic task is to determine whether a subjective document is positive or negative, and there are two classes of approaches to this problem. One approach is to treat this as a standard document classification task with two categories: positive and negative, and to use standard machine learning techniques, possibly with sentiment-oriented features [15]. The other approach is purely lexical: terms in a document are assigned a semantic orientation value (from an existing resource or computed using methods such as those described under WP2), and these values are combined to determine a semantic orientation value for the entire document [18]. In addition to document-level polarity classification, document-level subjectivity and sentence-level versions of these two tasks have also been studied [8, 15, 19].

Work on relation extraction for sentiment analysis has been spurred in recent years by the Multi-Perspective Question Answering (MPQA) corpus, in which sentiment relations between sources and targets are identified at a fine grain [20]. This corpus has been used for supervised machine learning to identify, inter alia, relations between sources and propositions [13] (cf. [11], who use their own corpus for this task). Little work, however, addresses the extraction of source-target relations. [14] extract source-target relations using a combination of entity recognition and word counting, but they assume that targets are already identified.

The overall goal of WP5 is to extract sentiment relations as coded by media analysts: criticism or support relations between source and target. While these relations are uniform in form, with a source, a target, and a polarity, they can arise from different textual expressions. Some encode the relation between the author of a subjective document and its topic, while others encode the relation between the agent of a specific speech act and its topic. Furthermore, while the latter category of relations are largely confined to limited segments of documents, the former may be characteristic of an entire document or only an excerpt. Since the annotations as we receive them from TrendLight are associated only to documents and not to text spans, our initial efforts in WP5 will be directed toward using them as labels for document-level subjectivity and polarity classification. We will compare lexical approaches (using the resources developed in WP2) and machine learning approaches.

Our ultimate goal is to extract sentiment relations, and not just sentiment labels, and to that end, we will experiment with methods for projecting TrendLight annotations into documents, using the entity recognition module developed in WP4, together with lexical resources developed in WP2. Given projected annotations, we will focus on sentiment relation extraction methods using entity-pair classification, which is standard in the literature on ACE relation extraction [12] but which has not been explored for sentiment relation extraction, in conjunction with automatically identified opinion expressions.

- *Evaluation*: Manual validation of codes as document labels; manual validation of projected codes as annotations. Document-level classification tasks: recall against TrendLight codes, manual precision evaluation. Relation extraction: recall against TrendLight codes, manual precision evaluation.
- *Risks & remedies*: Risk: Poor projection performance. Remedy: Manual validation extended to manual annotation; sentence-level sentiment/subjectivity classification may be necessary as filter. Risk: Poor classification/extraction performance (likely to vary by domain). Remedy: Assume that results are to be used in semi-automatic way (candidates to be filtered by human analysts); Have evaluation measures per domain to provide an indication of expected performance.
- *Start month/end month*: 10/19
- *Project partners involved*: TrendLight, UvA
- *Uses*: Domain-specific corpora (WP1), sentiment lexicons (WP2), entity recognition tools (WP4), Alpino
- *Deliverables*: (due at Milestone 2)
 - D.5.1: Projection tool to generate training data for relation extraction on basis of TrendLight codes and entity recognition module
 - D.5.2: Algorithms for subjectivity and sentiment classification, based on two both lexical and machine-learning methods
 - D.5.3: Algorithms for sentiment relation extraction
 - D.5.4: Test data derived from manual validation and assessment tasks

WP6. Aggregating sentiments in the news The objective of this work package is to aggregate sentiments mined from news documents and to discover the themes related to these sentiments.

If we are to mine sentiments on a Web scale, the mined document-level sentiments have to be aggregated before being presented to an analyst for review. Furthermore, since we expect that an unadorned list of source-target source/criticism relations will not be digestible for a human analyst, the relations must also be enriched with themes. In short, opinion summaries, with both sentiment relations and the topics out of which these sentiment arise, must be generated. The objective of WP6 is to experiment with summarization of sentiment relations extracted from edited content for which we have full manual annotations regarding not only sentiment but also themes in order to develop and evaluate algorithms that can be deployed on a larger and more noisy scale in WP7.

There are two sub-tasks for WP6: aggregating extracted sentiment relations on the basis of coreferring sources and/or targets (or, for infrequently occurring sources/targets, on the basis of actor type) and detecting topics in the documents from which aggregated relations are extracted. The first sub-task covers relatively unexplored territory. [17] address the problem of determining coreference for sources within a single document, but they are concerned with how to take advantage of the partial coreference annotation present in the MPQA corpus to train a machine learning-based coreference system (as described for WP4). Since we already intend to use a full coreference system for entity recognition (see WP4), our concern in WP6 is with the multi-document coreference task.

For multi-document coreference resolution, we will start from the within-document coreference resolution which forms coreference chains for all entities in the document (WP4). This information will be enhanced with information about the surrounding context, other possibly disambiguating named entities and keywords in the text, etc. This resulting information is converted into a bag of words feature vector which serves as input for a combined classification and clustering approach [6].

In order to discover the themes surrounding extracted sentiment relations, as well as to discover possibly new themes, we will use UvA's language-model-based topic detection tool. This tool—which has been used for a number of theoretical and applied purposes [3, 7]—will automatically mine the document collection and sub-collections associated with aggregated sentiment relations for potentially new themes (and sub-themes), which can then be presented to analysts for addition to the coding frame.

The agenda to be pursued in this work package has received scant attention in the literature. The primary contributions of the work package—developing methods for aggregating sentiments and discovering the related themes—will thus, in any case, be innovative in their application to these problems.

- *Evaluation*: Multi-document coreference: evaluation against TrendLight codes or, if necessary, artificially created data. Theme detection: manual assessment of identified themes against TrendLight-coded themes.
- *Risks & remedies*: Risk: Lack of ambiguity (thus no interesting research). Remedy: Create artificial data same-name data (to introduce artificial ambiguities). Risk: Too much ambiguity. Remedy: Manual evaluation of bad clusters can used as training data.
- *Start month/end month*: 18/20
- *Project partners involved*: TrendLight, HG, UvA

- *Uses*: Domain-specific corpora (WP1), sentiment lexicons (WP2), entity recognition tools (WP4), sentiment extraction tool (WP5), Alpino, UvA language-model-based topic detection tool
- *Deliverables*: (due at Milestone 2)
 - D.6.1: Multi-document coreference tool
 - D.6.2: Sentiment aggregation tool
 - D.6.3: Sentiment-related theme detection tool

WP7. Scaling to blogs The objective of this work package is to scale up the components built in the previous work packages to the Dutch blogspace. This concerns two main aspects: porting our sentiment coding work (WP5) to blogs, and porting our aggregation work (WP6) to blogs.

Major blog search and analysis engines (such as Technorati and BlogPulse) provide tracking services for the blogosphere. UvA's *MoodViews* uses aggregation to identify, present and explain sudden changes in the sentiment of the blogging community. Simple aggregation and peak explanation was implemented in UvA's electoral search engine *VerkiezingsKijker*. Mishne [7] provides a number of key observations on the differences between blogs and edited mainstream content that must be kept in mind as we move from news to blogs. First, language use in blogs diverges significantly from that in edited content. Second, unlike the mainstream media, which generally addresses topics of general interest, blogs cover an enormous range of topics, often refer to private experiences, and are often addressed to a limited audience. We expect that these factors should contribute to significantly higher referential ambiguity of names and other entity mentions. Third, the focus of blogs on private experiences means a higher degree of subjectivity in blogs. Of course, this subjectivity is precisely why we think blogs can be of great interest to media analysts. Fourth, blogs have structure that news articles do not. Unlike news articles, blog posts do not need to be self-standing and may rely on earlier posts for important content. Being hypertext, blogs and blog posts also have link-based structure. Many blogs have a comment feature, in which readers can write reactions to posts. Finally, spam (both splogs and comment spam) is a problem unseen in edited content. Some of these properties of blogs are challenges for sentiment analysis, while others may provide opportunities that we will explore.

With respect to this work package, we anticipate three major challenges (as an aside: spam detection and filtering will be addressed in WP1). First, the differences in language use between blogs and news mean that models trained on news content will be less effective on blogs (this has already been noted in the TREC blog track [8]). Second, the expected high degree of referential ambiguity will make aggregation more difficult. Finally, the sheer volume of content will make the presentation of results especially important. Effective solutions for these problems are bound to have implications far beyond the project.

Media analysts still need delineate the range of information that they would like to extract from blogs, but we do know what the basic tasks we will have to tackle here are. The first step will be to apply our linguistic processing tools: parsing and (domain-specific) entity recognition using models trained on TrendLight annotations for news content for the domain of interest. The high degree of subjectivity in blogs means that classification-oriented tasks (for both subjectivity and sentiment) will be important. Since we expect that models trained on news may not perform well on blogs, we will explore both lexical approaches and machine learning approaches. We also expect that the sentiment relations expressed in blogs will be heavily skewed towards those with author as source, which means that the sentiment mining task here is similar to the TREC blog opinion retrieval task. Thus, for those domains with a clear set of opinion targets, we will experiment with targeted opinion mining, using the combination of relevance-based retrieval and opinion classification that was most successful for this task [8].

The expected high degree of referential ambiguity means that coreference for aggregation will be especially challenging. For this task, we will experiment with exploiting link structure and the additional context that an entire blog offers to a specific post to help in disambiguation. As far as theme detection is concerned, the tools we have created for *MoodViews* and *VerkiezingsKijker* have already been successfully deployed for (Dutch) blogs.

- *Evaluation*: Evaluation will be in the form of manual assessment of the aggregated sentiments and related themes with respect to the documents from which the sentiments have been mined; student annotators in the early stage and TrendLight analysts in later stages. Assessments will be used in an iterative fashion to create training and test material for further experimentation.
- *Risks & remedies*: Risk: Blog data too noisy. Remedy: Invest more effort in spam detection. Shift focus from sentiment to relevance.
- *Start month/end month*: 21/32
- *Project partners involved*: TrendLight, UvA
- *Uses*: Domain-specific corpora (WP1), sentiment lexicons (WP2), Alpino adapted for blogs (WP3), entity recognition tools (WP4), sentiment extraction tool (WP5), multi-document coreference tool (WP6), sentiment aggregation tool (WP6), sentiment-related theme detection tool (WP6)

- *Deliverables:* (due at Milestone 3)
 - D.7.1: Sentiment retrieval tool for Dutch blogs: Retrieval of opinionated posts relevant to a specific given target
 - D.7.2: Sentiment mining tool for Dutch blogs: Extraction and aggregation of sentiment relations and discovery of themes related to a broad topic
 - D.7.3: Test set for sentiment mining derived from manual assessments

WP8. Demonstrators and applications This work package takes care of the demonstrators and of the transfer of technology to the professional end user. The objectives of the WP are (1) to demonstrate our resources, sentiment analysis and aggregation tools to the general public; (2) to test our research in real-life professional media analysis applications; (3) to obtain feedback from users; and (4) to engage in genuine develop-test-evaluate cycles within the life-span of the project itself so as to be able to release multiple versions of the online demonstrator.

Current news and blog search engines allow users to rank the results by relevancy or publication date. Some exceptions exist; e.g., Opinmind (<http://opinmind.com>) splits search results into positive and negative opinions, and ranks these side by side. Opinmind is limited to North-American blogs only. The demonstrators that we envisage within the DuOMAN project will go beyond, and differ from, Opinmind in a number of ways. Our demonstrators will be a combined Dutch-language (online) news and blog search engine, whose special feature will be facilities to not just rank search results by time or relevancy but to also enable sorting, grouping or filtering by sentiment. E.g., results on a topic can be grouped into *pro* or *con*, or restricted to *pro* only, etc. The initial baseline version of the system will offer a simple, data-driven implementation of this idea (on top of an Indri-based search engine), exploiting the UvA group's experience with *MoodViews* and *VerkiezingsKijker*, and the 2006 TREC blog track [7]. Later versions will incorporate the advances of WP5, WP6, and WP7.

The online demonstrators aimed at the general public will run on top of existing infrastructure at UvA. UvA's search engine log file storage and analysis tools will be adapted so as to provide the feedback needed by the project. Resources and algorithms that have been tested "in vitro" and as part of the online demonstrator will be put forward as components to be integrated into TrendLight's professional media analysis workflow. The planned bimonthly project meetings (see §6e) will be used for this purpose. UvA will deliver documented "academic" code and resources, leaving it to the TrendLight engineers to address the integration task.

- *Evaluation:* Through log file analysis, online feedback facilities, interaction with TrendLight analysts
- *Risks & remedies:* Risk: incremental updates reduce user satisfaction. Remedy: switch back to previous version and perform controlled user testing. Risk: demonstrator system not sufficiently responsive. Remedy: disable some of the non-core functionalities; identify problematic modules and improve responsiveness before the next release. Risk: limited uptake of resources and tools by TrendLight. Remedy: increase interaction between UvA and TrendLight; revise requirements of online demonstrators. Risk: demonstrator hosting/network facilities insufficient. Remedy: use UvA's cluster-based infrastructure.
- *Start month/end month:* 1/36
- *Project partners involved:* TrendLight, UvA
- *Uses:* Lemur/Indri for indexing and retrieval; Apache for web serving; MySQL for log analysis; builds on the infrastructure created for *BlogKijker*, *MoodViews* and *VerkiezingsKijker*, tools and resources developed in work packages WP2–WP7
- *Deliverables:* (D.8.1, 2, 3 due by Milestone 1; others by Milestones 2, 3, 4, respectively)
 - D.8.1: User requirements for both professional and general public users and system specification
 - D.8.2: Log analysis tools
 - D.8.3: Baseline online demo, v0.1 plus evaluation and transfer to TrendLight
 - D.8.4: Demo v1.0 plus evaluation and transfer to TrendLight
 - D.8.5: Demo v2.0 plus evaluation and transfer to TrendLight
 - D.8.6: Final consolidation and handover to TST Centrale and TrendLight

6a.5 The project's main contributions

The main scientific contributions of the DuOMAN project may be summarized as follows:

- Lexical resources for Dutch sentiment analysis. (Domain-independent Dutch sentiment lexicon, built on top of Dutch WordNet, and algorithms for extending this lexicon with sentiment orientation for domain-specific terms using data-driven methods.)

- Algorithms for information retrieval and information extraction tasks. (Transforming the output of professional media analysts into training data for machine learning algorithms for sentiment-oriented document classification and information extraction tasks; Classifying documents and snippets according to sentiment orientation; Extracting actors and opinion relations from documents; Aggregating mined opinion relations; Detecting topical trends related to mined opinion relations.)
- Test sets to evaluate classification and extraction algorithms.
- A data set consisting of Dutch language blogs crawled over a three year period.
- New language technology tasks. (Extraction of source-target sentiment relations; Aggregation of sentiment relations from multiple documents; Detecting themes related to aggregated relations.)

6b Economic aspects

Within the DuOMAN project there is explicit cooperation with and support by companies. Building on a joint small-scale pilot with the UvA team, TrendLight makes four types of contribution to this project: problem statement, annotated data, integration effort, and assessment effort and time, thereby demonstrating a serious and tangible commitment to the project. Another key contribution is offered by GridLine. GridLine and UvA have previously worked together within an Innovatievoucher-funded project, where UvA provided consultancy and knowledge transfer documents on algorithms for domain-specific information extraction. The interest of GridLine in this project is related to the development of tools for the automatic analysis of domain-specific document sets. To support this, GridLine provides tools, knowledge and assistance for the development and evaluation of the lexical resources.

As to the prospects for spin-offs, within the Netherlands there are several start-ups in the area of online media analysis (including Buzzcapture, ZookMa). Also, the potential of online media analysis tools to provide near real-time feedback is of great interest to traditional media analysis companies. E.g., like many media analysis companies worldwide, TNS-Nipo, parent company of TrendLight, is actively exploring the potential of the participatory web.

Tracking dynamic online data streams (news, blogs, etc) offers a range of industrial and/or societal opportunities. Never before have we had access to this amount of consumer data, which is of great interest to marketers, policy makers, and strategists. Internationally, there are several commercial activities that bring modern language technology to bear on online media analysis. These include Corpora Software in the UK, and companies such as Umbria Inc and Nielsen BuzzMetrics (which combines BuzzMetrics, BlogPulse, Trendum, and Intelliseek).

6c Contribution to the STEVIN programme

The DuOMAN project contributes to all three goals of the STEVIN programme. First, it focuses exclusively on the Dutch language, both in the resources that it aims to build and in the tools that it plans to develop. Second, it promises to improve the position of the Dutch language in the modern information and communication society by creating analysis tools to help make sense of data generated by Dutch speaking (or rather: writing) web users. Third, it promises to make exploration of the attitudes of the Dutch blogspace possible for both the general public and professional media analysts. The project is a mixture of all layers that the STEVIN programme covers: basic resources, research and development, technology integration, and end users. Within DuOMAN there is a realistic perspective on knowledge transfer and network creation: the outputs of the project will be integrated into the workflow of professional end users (TrendLight's media analysts), and important parts of the resources created by the project will be administered by GridLine, using their lexicon tools. DuOMAN integrates the results and methods from several projects funded within the STEVIN programme (Cornetto, Corea, LASSY, D-Coi).

6d IPR and standards

The project partners UvA declare willingness to negotiate agreements with the TST Centrale on access, use and exploitation of the DuOMAN project outcomes by the R&D community. The UvA language modeling software is freely available for research purposes; earlier versions have been released as open-source distributions. The improvement/extension of UvA's language modeling tools made possible within DuOMAN will be made available in a new release of the toolkit at the end of the project. We will strive to make all resources developed within the project available under an open source license, such as the GNU Lesser General Public License. The scientific results achieved by the DuOMAN project will be published in papers at scientific conferences and in journals.

The project will use document-level annotations provided by TrendLight. If UvA does not have a license for the underlying documents, the machine learning required for WP2, WP4, and WP5 will be carried out on site at TrendLight. If a license is available, UvA will transfer the documents plus annotations to its own servers.

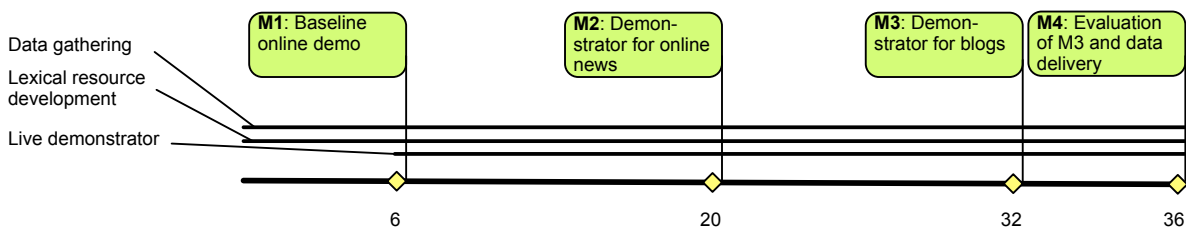


Figure 3: Overall timeline of the project, with milestones.

This project makes use of data crawled from the Internet. Crawling will respect the Robots Exclusion Protocol and general crawler etiquette. Crawled news content will be used for indexing purposes only and will not be redistributed or displayed in a form, which allows it to be reconstructed. News content will be deleted in case of request from the owner. DuOMAN’s demonstrators will present a link to the user allowing her to view news messages from the original source. When news content is no longer available from the original source, it will not be made available by the demonstrators. For blogs, largely the same policy will be used—except that at the end of the project the blog data crawled will be made available to the R&D community. (UvA has previous experience with releasing large web crawls as part of CLEF, under suitable licenses, based on examples from TREC.)

The project makes use of log data collected from the search engine for the purpose of improving the engine by analyzing user behavior. Users will be alerted that information is being collected. The data will be anonymized by software agents before it is analyzed. It will be stored on secure servers only and not be made publicly available.

To the extent possible the project builds on existing standards, e.g., for data exchange and lexical resource structuring and storage; the project does not explicitly aim to extend existing standards or develop new standards.

6e Coordination and project management

Structured into 8 work packages, the project will be managed by UvA; (see Table 1). Except for WP3 and WP4, all WPs are led by UvA. Synchronization of the WPs is monitored by UvA—for its own WPs, this is an integral part of the WP, and for WP3 and WP4 synchronization takes place within the setting of WP5 for which WP3 and WP4 produce input. There will be bi-monthly meetings with all project partners. During the meetings partners will report on the progress of their WPs. UvA will also report on the demonstrator progress and, together with TrendLight, identify resources and tools for integration within TrendLight’s workflow. In between the bi-monthly meetings, WPs to which multiple participants contribute will have weekly phone meetings. Standard collaboration software (wiki’s, TRac) will be used to support the project. Dr Janyce Wiebe (Pittsburgh) has kindly agreed to serve as External Advisor for the project, and she will provide on the general strategy of the project at least once a year.

6f Evaluation, validation and success criteria

Evaluation and validation are at the heart of the DuOMAN project, with a strong focus on component-level evaluation. Specific proposals for evaluation of the components developed in each work package are detailed in the work package descriptions in §6a.4. Validation facilities for WP2 will be provided by GridLine. For evaluating Alpino (WP3) and the coreference resolution tool of WP4 the WPs involved will develop ground truth. Facilities for evaluating the results of WP5, WP6, and WP7 will be through online demonstrators, user studies, and, of course, through TrendLight’s media analysis. Technical hosting and server facilities for the online demonstrators will be provided by UvA. UvA’s database backend will be used for the log analysis module needed by WP8. Success at the component level will be assessed in comparisons to state-of-the-art performance on similar data or tasks (if available) or in comparisons to “acceptable” baselines. Success of the demonstrators will be assessed with respect to deliverable D.8.1, the user requirements document, and we will use the bi-monthly meetings to reevaluate these requirements on a regular basis.

7 Work Programme

The project will be developed in four R&D phases, each leading to a milestone and targeting specific demonstrators and/or evaluations; see Fig. 3; see Table 1 for an overview of the work packages.

In the start-up phase the emphasis of the engineering efforts will be on platform development and establishing a baseline online media analysis system (for naive end users). This system will build on existing tools and resources from UvA and TrendLight. The **Milestone 1** is reached when the first online demo goes live. The delivery date of the milestone (month 6) is chosen so that it is possible to take current tools and add a very basic, data-driven

sentiment analysis approach. In parallel with the baseline demo development, the scientific efforts will work towards **Milestone 2** (month 20), which incorporates the modules developed in WP2 and WP4–6. The focus of the milestone will be on *linguistic resources* and *aggregation*. Milestone 2 also includes an analysis of the use of the baseline demo. During the third phase, leading to **Milestone 3** at month 34, we port our sentiment analysis and aggregation solutions to work on user generated content. In the final consolidation phase, Month 33–36, with a **fourth and final milestone** at month 36, an evaluation of the M3 demonstrator will be conducted. We perform an end-to-end evaluation in cooperation with TrendLight of their workflow extended with DuOMAN’s tools and resources. The project’s code and resources will be prepared for integration by TrendLight and handover to the TST Centrale.

In each new R&D phase, the quality of the sentiment analysis and aggregation tools will be improved; where needed they will be made more robust, and additional functionalities will be added to the demonstrator system, leading to a new major milestone.

8 International Perspective

The ILPS group is a frequent participant in (and co-organizer of) international retrieval evaluation efforts such as CLEF, INEX, NTCIR, and TREC (including TREC’s blog track). The methodological and algorithmic lessons from these participations will feed directly into the demonstrators envisaged within the project. In addition, the ILPS group is active in organizing international conferences—SIGIR, the major international conference in information retrieval, will be held in Amsterdam in 2007, and, more specifically related to the DuOMAN project, the International Conference on Weblogs and Social Media will be held in Amsterdam in 2009.

9 Short CV Principal Applicant(s)

Maarten de Rijke (principal investigator) graduated from the University of Amsterdam in Mathematics and Computer Science (PhD, 1993) on the topic of modal logic. In 2004 he became full professor in “Information processing and internet” at the UvA. He has an interest in “information retrieval beyond the document”: retrieving answers, objects, opinions, experiences. He has written over 350 papers in refereed journals and conferences. He was editor of the *ACM Transactions on Computational Logic*, and is currently editor of *Foundations and Trends in Information Retrieval* and of the *Cambridge Studies in Natural Language Processing*. De Rijke leads a group of around 25 staff within ISLA, the Intelligent Systems Lab Amsterdam at the Informatics Institute of the University of Amsterdam.

David Ahn is a postdoctoral researcher in ILPS at the University of Amsterdam. He received his PhD in 2004 from the University of Rochester with a thesis on the semantics-pragmatics interface. He has worked on the NWO project ITEQA and has also contributed to information extraction and question answering efforts within ILPS.

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 ‘Algorithms for Linguistic Processing.’ He is the key architect of the Alpino parser for Dutch. In 2005 and 2006, van Noord was the chair of the EACL. He has contributed to the STEVIN projects D-Coi, IRME, and LASSY.

Veronique Hoste studied translation (School of Translation Studies, Antwerp) and computational linguistics (University of Antwerp). She worked at the CNTS-language technology group on optimization in machine learning of coreference resolution (PhD, 2005). In 2006, she founded the LT3 Language and Translation Team at the School of Translation Studies (Ghent), where she heads the language technology section. Her research is focused on machine learning of natural language. She was involved in the STEVIN Core project and the SoNaR project.

GridLine is an ICT-company that specializes in the construction and application of domain-specific word systems (e.g., thesauri) for companies and public institutions. GridLine provides a broad range of services, from product development to consultancy to implementation. Among GridLine’s products are specialized termlists and thesauri, thesaurus-based systems for document retrieval, and applications for dynamic term screening.

TrendLight is an independent research institute for media reputation. TrendLight uses state-of-the-art software solutions together with human analysis. TrendLight maintains strong ties with academic research in the areas of media research, statistical analysis, and software development. TrendLight’s customers include large multinationals as well as governmental and societal organizations.

10 Literature References¹

Selection of publications by applicants

¹As per the guidelines the references are organized in two groups of at most ten each.

- [1] D. Ahn, W. Alink, V. Jijkoun, M. de Rijke, P. Boncz, and A. de Vries. Representing and querying multi-dimensional markup for question answering. In *Proceedings EACL 2006 NLPXML-2006 Workshop*, April 2006.
- [2] D. D. Ahn. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics.
- [3] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *Proceedings 11th Meeting of the EACL (EACL 2006)*, April 2006.
- [4] V. Hoste. *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, Univ. Antwerp, 2005.
- [5] J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using WordNet to measure semantic orientations of adjectives. In *Proceedings LREC 2004*, volume IV, pages 1115–1118, 2004.
- [6] E. Lefever, T. Fayruzov, and V. Hoste. A combined classification and clustering approach for web people disambiguation. In *Proceedings of SemEval 2007*, 2007.
- [7] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, 2007.
- [8] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 Blog Track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*. NIST, 2007.
- [9] G. van Noord. At Last Parsing Is Now Operational. In *TALN 2006 Verbum Ex Machina, Actes de la 13e Conference sur le Traitement automatique des Langues naturelles*, pages 20–42, 2006.
- [10] P. Vossen, K. Hofmann, M. de Rijke, E. Tjong Kim Sang, and K. Deschacht. The Cornetto database: Architecture and user-scenarios. In *Proceedings DIR 2007*, pages 89–96, 2007.

International literature

- [11] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky. Automatic extraction of opinion propositions and their holders. In *Proc. AAAI Spring Symp. Exploring Attitude and Affect in Text*, 2004.
- [12] R. C. Bunescu and R. J. Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of HLT-EMNLP 2005*, pages 724–731, 2005.
- [13] Y. Choi, E. Breck, and C. Cardie. Joint extraction of entities and relations for opinion recognition. In *Proceedings of EMNLP*, 2006.
- [14] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the COLING 2004*, 2004.
- [15] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL-2004*, 2004.
- [16] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of CoNLL 2003*, 2003.
- [17] V. Stoyanov and C. Cardie. Partially supervised coreference resolution for opinion summarization through structured rule learning. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2006.
- [18] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL'02*, 2002.
- [19] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Comput. Ling.*, 30(3), 2004.
- [20] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3), 2005.

11 Project budget details

Name	Position	Time	Costs	Total
Dr David Ahn	Postdoc UvA	36 months, 1.0 fte	221,875	
NN	Programmer UvA	36 months, 0.5 fte	104,100	
	Benchfee UvA	36/36 * 16,500	16,500	
Dr. Gertjan van Noord	UHD RUG	3 months, 1.0 fte	22,500	
	Benchfee RUG	3/36 * 16,500	1,375	
Dr Veronique Hoste	UD HG	6 months, 1.0 fte	32347	
	Benchfee HG	6/36 * 16,500	2,750	
Dr Oele Koornwinder	GridLine	480 hrs @ 50€/hr	24,000	
<i>Total personnel:</i>				425,447
	Annotation effort	250 student hrs @ 20€/hr	5,000	
	User study	students	5,000	
	Data storage	10 Tb	5,000	
<i>Total other:</i>				15,000
Grand total:				440,447