KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

KNAW-Agenda Grootschalige Onderzoeksfaciliteiten
Format nadere uitwerking van een ingezonden voorstel

I. VOORSTEL ALGEMEEN

Acroniem	ADVANT
Naam van de infrastructuur	**AD**vanced **V**ideo **AN**alysis **T**ool (formerly: MAD)
Hoofdindiener Naam contactpersoon	Prof. dr. Tom Koole
Organisatie	Rijksuniversiteit Groningen
Functie	Professor
Adres	P.O.box 716, 9700 AS Groningen
Telefoon	+31(0)503638548
Email	tom.koole@rug.nl
Mede indiener(s) Max 5 totaal. Rest alleen naam + organisatie	
Naam contactpersoon/ mede-indiener	Prof. dr. Antal van den Bosch
Organisatie	Radboud Universiteit
Functie	Professor and Research director Centre for Language Studies, Faculty of Arts, Radboud University
Adres	Postbus 9103, 6500HD Nijmegen
Telefoon	+31(0)24-3611647
Email	a.vandenbosch@let.ru.nl
Naam contactpersoon/ mede-indiener	Prof. dr. Sandra van Dulmen
Organisatie	NIVEL/Radboud University Medical Center
Functie	Professor and Research coordinator
Adres	PO Box 1568, 3500 BN Utrecht
Telefoon	+31(0)30-2729703
Email	s.vandulmen@nivel.nl
Naam contactpersoon/ mede-indiener	dr. Wyke Stommel
Organisatie	Radboud Universiteit
Functie	Assistant professor
Adres	Postbus 9103, 6500HD Nijmegen
Telefoon	+31(0)24-3612482
Email	w.stommel@let.ru.nl
Naam contactpersoon/	dr. Tessa van Charldorp

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

mede-indiener	
Organisatie	Universiteit Utrecht
Functie	Assistant professor
Adres	Trans 10, 3512 JK Utrecht
Telefoon	+31(0)627369341
Email	t.c.vancharldorp@uu.nl
Board of advice / user group	dr. Khiet Truong (University of Twente) dr. Gregory Mills (University of Groningen) dr. Onno Crasborn (Radboud University) prof. dr. Julia Noordegraaf (University of Amsterdam) Han Sloetjes (Max Planck Institute, the Language Archive, Nijmegen) The consortium further consulted: Micha Hulsbosch, Johan Oomen (Beeld & Geluid), Francisca de Jong (UU, CLARIAH), Jan Odijk (UU, CLARIAH), Els Stronks (UU) & Nicoline van der Sijs (RU) (DESIDERIA), Henk van de Heuvel (RU), Asli Özyurek (RU), Paul Trilsbeek (MPI – TLA), Kobin Kendrick (MPI), Arjan van Hessen (Twente, UU, CLARIAH), Dirk Heylen (HCI, Twente), Roeland Ordelman (HCI, Twente & Beeld en Geluid), Gerben Westerhof (psychology, Twente), Miranda Olff, (trauma studies, AMC), Arnold Smeulders (computer vision, UvA), Grzegorz Chrupala & Afra Alishahi (computational linguistics, Tilburg), Joost de Wit (Media distillery), Christian Heath & Paul Luff (King's College London, UK), Arnulf Depperman & Thomas Schmidt (Institut für Deutsche Sprache, Germany).

Samenvatting
Geeft korte samenvatting van deze faciliteit in termen van werking, wetenschappelijke voordelen etc. (max 350 woorden).

ADVANT (Advanced Video Analysis Tool) is a shared facility in which multimodal interaction data can be played, stored, transcribed, and annotated. It can be used for a wide variety of research questions in the humanities and social sciences. Detailed levels of transcription and annotation can be done at video (image) and audio (sound) level, both automatically and manually. All data (raw, transcribed and annotated) can be searched through at both the video and audio level including multi-level searches.

Since there is a huge increase in video and audio data in society (for example, mobile phone video recordings and street cameras) there is a growing need to analyze such data more quickly and efficiently than ever before. Data samples, even those for which informed

K O N I N K L I J K E   N E D E R L A N D S E
A K A D E M I E   V A N   W E T E N S C H A P P E N

consent is acquired, are often inaccessible to others, difficult to share or analyze in collaboration with others, also due to privacy and security issues. The ADVANT facility offers a transparent way in which raw and annotated data can be shared with the scientific community. Various kinds of automatic analyses can be run, both on the level of images and sound (e.g., speech recognition, emotion detection, etc.). Not only after-the-fact, but also 'live' collaborative analyses will be possible. This will increase transparency and reliability of this type of research worldwide, within a safe environment. The facility will also support interaction and multimodal analyses on a much larger scale than ever before, allowing validation techniques for experimental research.

Various automatic analysis techniques that are becoming usable and sufficiently accurate on real-world data will be combined within the ADVANT facility, including speech recognition and computer vision. We will use CLARIAH standards in order to be compatible with the CLARIN/CLARIAH services and infrastructure. Data can also be stored in collaboration with CLARIAH. The ADVANT facility will include a desktop-based front end for the automatic annotation. Sharing data and notes, doing real-time collaborative data sessions and automatically annotating as well as archiving and publicizing data will be done through a back end of integrated servers. Controlling these tasks on the servers will be done through the front end so that the user only needs one tool for all tasks.

Kernwoorden
Geef maximaal 8 kernwoorden die de faciliteit typeren.
Audiovisual data analysis tool, collaborative & automatic annotating, searchable, interaction, multimodality, transcription.

II.   VOORSTEL INHOUDELIJKE UITWERKING

A. SCIENCE AND TECHNICAL CASE
Volledig nieuwe faciliteit of verbetering van reeds bestaande
– *Beschrijf in hoeverre het hier een geheel nieuw idee betreft of een verbetering of opvolging van een reeds bestaande faciliteit.*
The ADVANT facility is a new idea. Our survey among researchers using video data has shown that there are no facilities like the one we are proposing anywhere in the world. However, we intend to make use of the codebase of existing software such as ELAN as well as CLARIAH services and infrastructure. The front-end of the facility will be desktop based (not web-based), but data will be stored on a central server to facilitate collaboration between researchers (e.g., combining software such as Arbil and LAMUS).

K O N I N K L I J K E   N E D E R L A N D S E
A K A D E M I E   V A N   W E T E N S C H A P P E N

ELAN

For the structure of the facility we will use ELAN as a starting point. ELAN already has the functionality to simultaneously play back synchronized video files and to annotate videos on a large number of tiers (analysis layers). New about ADVANT is that the tool (which integrates ELAN) will be highly adaptable to fit different kinds of data sets and many fields of research and their methods. This adaptability is not limited to the user interface, but also includes functions for transforming data between several coding systems and functions for database connections, data archiving, data publishing, version control systems, collaboration tools and other back end tools located on a central server. ADVANT aims to also build a back end of seamlessly integrated tools such as the ones mentioned above. The tool will have an intelligent search system that allows for combined searches (e.g., *eyes wide open* combined with "oh" and video example based searches). Users will be able to plug in tools for automatic annotation of both audio and video, including speaker recognition, speech recognition and computer vision tools for recognizing human behaviour (which is under development, but steadily evolving in computer science).

CLARIAH

We intend to use CLARIAH standards, making it compatible with all existing (and future) CLARIN / CLARIAH services. ADVANT could be integrated with the Media Studies domain from CLARIAH (see below). Data management for ADVANT can then be arranged in collaboration with CLARIAH.

Science Case

– *Geef een algemene introductie van de wetenschappelijke waarde van de faciliteit.*

Research of human communicative and language behaviour is increasingly concerned with video data. Language and social interaction researchers study multimodal aspects of face-to-face interaction in authentic video data, gesture researchers work with experimental video data, social scientists study facial emotion displays, computational linguists want to develop tools for the automatic annotation and analysis of language use in video data, and researchers in the medical field study healthcare interactions as a way to identify (communication) training needs and to evaluate, for instance, effects of (policy) measures, protocol or interventions. At the same time, video data are increasingly becoming available as footage from television stations stored in archives such as the Dutch 'Beeld en Geluid' (Sound and Vision), and institute through social media such as Youtube, Facebook, and more recently Periscope and Twitch, to a great extent filled by members of the public who use their mobile phones to film or audio-record public and private events as well as medical interactions. The ADVANT facility aims to establish a necessary condition for researchers of multimodality to use Dutch video data to contribute to the internationally developing field of the study of multimodality.

Social interaction is a multi-semiotic enterprise, in which language as itself a multi-semiotic mode (lexicon, syntax, prosody) is used in combination and interaction with embodied

tokens such as body posture, eye gaze, head movement, and limb movement such as hand gestures. In spite of this multi-layered nature of communicative utterances, language users succeed in assembling these different semiotic channels into single, integrated interpretations. For the investigation of language-in-use, it is vital that we discover how these different semiotic systems contribute to these interpretations.

Tomasello (2008) argues that phylogenetically and ontogenetically embodied interaction is the basis for language use but the big question is how language and embodiment are related for adult speakers at the present state of evolution. Do we only speak when we cannot achieve our goals non-verbally? Recent research on Italian suggests for example that the unmarked manner for doing requests is the embodied rather than the linguistic format (Rossi, 2015). In medical settings, the content captured by spoken interactions often does not coincide with embodied signals emitted by healthcare providers or patients, making it vital to facilitate and study the synergy between different layers as well as the sequential nature within (dyadic, triadic or group) healthcare interactions. How do we combine or distribute spoken and embodied means and to what ends and what meanings?

The ADVANT video analysis tool will enable researchers to search for and gather specific visual phenomena or combinations of visual phenomena with specific linguistic elements, or with specific interactional positions such as end of turns. In smaller and larger data corpora, the ADVANT facility will be able to collect all instances of the object of inquiry. Moreover, ADVANT will allow researchers to explore systematic relations between the object of inquiry and annotated information on other tiers, both for the purpose of qualitative and for quantitative analyses. Finally, ADVANT and the attached video corpora will establish a test-base for the development of forms on automated annotation and analysis. This is even more important because the current coding practices require specific coder expertise, and are therefore very time-consuming and expensive.

The scientific fields for which the ADVANT facility aims to establish a necessary condition for doing research on multimodal interaction are:

Language and social interaction
A first major challenge in the field on language-and-social-interaction is the question of action formation: how do language users recognise an utterance as a certain action such as a request? This is, among others, very relevant and timely in medical settings, where the goal is to reach shared decision-making which requires a thorough understanding of each interactant's implicit and explicit messages. While earlier research suggested that language users orient for this to syntactic and prosodic characteristics, recent research has shown that this is not the case: both interrogative and declarative formats, with both rising and falling final intonations are understood as requests (Heritage, 2012). So far, however, embodied aspects of interaction have not played a major role in the study of action formation.

A second major challenge for this field of investigation is the issue of intersubjectivity: how

do participants in interaction establish and maintain common understanding? For this purpose, language users have a number of particles available such as mh, yes, oh, that can be used to claim understanding, but also embodied tokens such as different types of head nods (Whitehead, 2011). The big question is how we use these different tokens to indicate different objects of understanding (e.g. action, referent, proposition, emotion), different scopes of understanding (e.g. a prior turn constructional unit, an entire activity such as an explanation), and different epistemic stances (ranging from certain to uncertain). It is obvious that facial expressions can display epistemic stances, but research is yet far from a systematic analysis of this relation. At the same time Rossano (2012) showed that we have only begun to understand how embodied tokens contribute to the organization of talk when he discovered the much less obvious phenomenon that eye-gaze direction is used to indicate the understanding that an activity has come to a close.

Gesture and sign language studies
Corpus-based studies in the analysis of visual communication in the sense of hearing people's gestures and deaf people's sign languages are still in their infancy, because of the extremely time-consuming nature of manual annotation. In sign linguistics, data sets tend to be very small, while in gesture studies, experimental paradigms are more prominent. Both fields could see significant progress in their understanding of the role of manual and non-manual signals if they were based on larger corpora. Most crucially, to answer the larger question of the relation between sign and gesture, similar methodologies are needed. Improving the speed of annotation would enable such ecologically valid parallel data sets, enabling answers to burning questions like whether sign languages always find their source in the manual and non-manual gestures of the surrounding hearing population (Sandler et al. 2005; Senghas et al. 2004; Coppola et al. 2013; Novack et al. 2015).

Social and medical sciences
With ADVANT researchers of the NIVEL Databank Communication in Healthcare can more thoroughly investigate (determinants of) effective communication in healthcare. It is well-known that – apart from those of medical technical interventions – medical communication too can have health effects which are comforting or helpful for many patients. Patients, for instance, report less pain when their care provider expresses positive outcome expectations when prescribing pain medication or when (s)he provides more elaborate procedural information about a treatment (Mistiaen et al, 2015). Apart from that, there is also evidence for the health effects of providing empathy (Fogarty et al, 1999; Di Blasi et al, 2001). In order to implement such effects, we need more insight into which aspects of the spoken and embodied communication make a difference; the therapeutic value of, especially, embodied communication has so far not been investigated much, simply because of the lack of the right infrastructure like ADVANT offers. The possibilities offered by ADVANT for recognizing and analyzing emotions and the meaning of (the combination of), a.o., facial expressions and characteristics of talk, allow a yet unknown and under-researched field of research in the field of communication in healthcare, relevant for researchers, trainers and

health professionals in the Netherlands and beyond. After all, especially those implicit and more or less conflicting (spoken and embodied) communicative utterances determine whether or not patients understand, recall and follow through the advices being given by their care provider and if the information provided is being understood as intended.

Affective Computing and Social Signal Processing of Human-Machine Interaction
In the field of media studies, emotion detection is a crucial factor for analysing the meaning and impact of film, television programs and other media productions. Automated emotion detection could facilitate the attribution of genre as well as the understanding the emotional responses of viewers of media content (Tan 2014).

In the emerging domains of Affective Computing (AC) and Social Signal Processing (SSP), the goals are to develop emotionally and socially intelligent machines. The development of automatic recognizers of human emotional and social behaviour using machine learning techniques plays an important role in achieving these goals. To that end, a huge amount of multimodal data needs to be annotated and analysed for human social and emotional behaviours in both human-human and human-machine interactions. One of the challenges in creating emotion annotations is the relatively high level of subjectivity involved. Humans do not always agree on the observed emotion. In order to reach a sufficient level of inter-annotator agreement, annotators often sit together and analyse the annotations to reach consensus.

Computational linguistics
The benefits of ADVANT are various. Advant will facilitate the development of novel computational models of language (e.g. of dialogue acts, domain and genre-specific language modelling, language learning) that are based on fine-grained analyses of large data-sets. For example – in the largest study of its kind to date, Deb Roy placed cameras and microphones throughout his home, yielding a near-complete audio-visual record of the first two years of his son's life, in particular his interaction with his caregivers (Roy et al., 2012). For each word that was acquired by his son, the recordings were used to generate "wordscapes" – composites of automatically generated maps and human transcriptions – that correlated where in the house the word was used with the embodied actions of the caregivers and child. A key finding was that these wordscapes are highly predictive of when a word is learned: Unique configurations of embodied and spoken behaviour drive acquisition.

Also, ADVANT contributes toward the applied research goal of improving language technology applications. For example, currently one of the key challenges is developing automated agents that assist with the running of meetings. Projects such as CALO (Cognitive Assistant that Learns and Organises) (Tur et al. 2010) aim to create agents that transcribe and analyse meetings in real-time, allowing them to contribute to the meeting by providing relevant information at key moments in the interaction. In addition, such agents should be able to replay and summarize the meeting, and also help the decision-making

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

process by facilitation and coaching. This challenge is exacerbated by the multi-party and multi-modal nature of meetings. Yet this multiplicity of channels can also be used to support automated analysis: Patterns of speech overlap can be used to infer distribution of roles and power-relations; collaborative turn-completions can be indicative of agreement on joint decisions, and topic-shift can be substantiated by embodied patterns such as postural sway.

However, these computational challenges can only be met if comparisons be readily made within- and across- many related corpora, both by human researchers and by machine-learning algorithms.

– *Beschrijf de wetenschappelijke voordelen en verwachte doorbraken.*

- ADVANT offers the opportunity to collect, store, annotate and analyze data on a large scale, allowing the creation of larger collections of interactional and multimodal phenomena and thus allowing large-scale interaction research.
- Data can be transcribed and annotated much quicker with the aid of automatic transcription and annotation tools;
- Allows for increased transparency within the fields of multimodal analysis, interaction analysis and corpus analysis;
- Combinations of actions (spoken and embodied) can be related to each other;
- The data will be stored and made accessible online allowing researchers worldwide to access the raw and annotated data. References can be made to ADVANT in scientific publications and analyses can be repeated.

– *Beschrijf hoe deze faciliteit zich verhoudt tot alternatieve faciliteiten/onderzoeksmethoden.*

ADVANT is a tool to play back, analyze (both manually and automatically) and search through video data and to collaborate with other researchers. With its focus on video data, it complements initiatives for text-focused tools such as Desideria. The objective is that the tool is embedded in the Media Studies/Audiovisual Data focus domain of CLARIAH (Work Package 5): the data sources (audio-visual) and tools for automatic analysis largely overlap with the objectives of this domain. Tools in this domain are nascent (e.g., in the LISTEN-project, currently run by UvA and Beeld & Geluid) and will be more advanced in the future, tying in with computer vision developments. A collaboration between ADVANT and Media Studies / Beeld & Geluid would further this development, create a context for humanities and social science collaborations and warrant researchers' access to new techniques. What we would add to the Media Studies domain in CLARIAH is the focus on human social interaction and behavior from both a linguistic and social scientific perspective. This implies the Media Studies domain would be expanded with a new research community: language and

social interaction. In the context of ADVANT, we would develop a new corpus of ordinary social interaction – note a Dutch corpus of this kind does not exist yet - which could be stored, conformed to CLARIAH standards, at the Beeld & Geluid data center.

In the international field of language and interaction, very few researchers use technical facilities for data analysis. Renowned researchers like Christian Heath and Paul Luff (King's College London) use manual, hand-written notes and transcriptions rather than software for data analysis. ADVANT could thus serve an international research community. In Germany, the Institute for German Language (IDS) is currently trying to obtain funding from the Leibnitz Gesellschaft for the development of a tool for manual transcription of video data (based on EXMARaLDA), serving the linguistics research community (sociolinguistics, interactional linguistics, conversation analysis). This initiative takes place under the umbrella of CLARIN-D. They have large, annotated audio corpora (in German) at their disposal. Contrary to ADVANT, the project does not include automatic analyses. We have agreed to collaborate with the IDS in co-development of (sub-)tools if this would prove fruitful. We Also have also established contacts with Brian Due of the University of Copenhagen where a video analysis software is being developed in the context of ethnographic research.

The MPI-tools ELAN, Arbil, Lamus and Annex are tools for building a manually annotated audio/video corpus; the present tools build on these tools. The EXMARaLDA system is similar to the MPI tools. Both tool sets are not as integrated and flexible as ADVANT aims to be. Furthermore, the functions for collaboration and automatic annotation are much more advanced in ADVANT.

Technical case
– *Geef op hoofdlijnen een technische beschrijving van de faciliteit. Hoe zit de faciliteit in elkaar en hoe werkt het?*

The ADVANT facility aims to offer a complete and intuitive workflow for researchers doing video and audio analysis for the humanities, social and medical sciences and other domains of use. The facility includes:

1. a central service for

    1. gathering raw data (audio and video), either recorded by researchers, patients or healthcare providers, or uploaded by members of the public, e.g. using a mobile phone;

    2. organizing the data in corpora and subcorpora;

    3. exchanging and sharing data with related projects;

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

4. collaboration for manual annotation including version control, working on the same data with two or more people and comment on fragments, annotations, transcriptions, metadata etc.;

5. an intelligent search engine for metadata and transcriptions and annotations, including search by example (are there similar fragments to the one given by the user) and the possibility to combine features (e.g., eyebrows raised and "oh") and analyse communication sequences.

2. a system for automatic generation of metadata (e.g. where did the event take place inside or outside, the number of participants), transcriptions and annotations using state of the art audio and video analysis tools and supplemented with other data from questionnaires or otherwise.

3. a flexible tool for manual annotation of audio and video data. The flexibility includes the adaptation of the user interface (e.g. key board short cuts for codes/symbols), the possibility to turn functions on and off and a well designed plug-in system for adding new functionalities

4. a system for archiving data for future generations

Components 1.1 to 1.5 will be implemented as one package on a server. This server will act as the central system with connections to a server for automatic analysis (2) and archiving server (4), and the tool for manual annotation (3).

Main components 1 to 3 are implemented in a modular fashion and they will communicate with each other using well defined and well described formats and protocols. This creates the possibility to use each component independent of the complete facility. The formats will be developed in accordance with CLARIAH guidelines.

The central idea of our facility is flexible and streamlined manual and automatic annotation of audio-visual data in a collaborative way. Below we will further describe the three notions of data, collaboration and annotation.

1. THE DATA ITSELF

The central part of the facility will manage the research data. The tasks of this part are storing the data so that it can be used for automatic and manual analysis, allow for a wide range of collaboration functions, and manage archiving and publication.

Types of data

The ADVANT facility is designed to make analysis of videos of human behaviour a smooth and productive process. We intend to allow for an event-based analysis meaning that several media files of one event can be analysed in parallel. That means that the tools in our facility

must be able to handle multiple media files at once.

The reason for focusing on video is threefold. First, video is already an abundant digital medium and we expect that the use of video data will increase in the years to come. We aim to tap into the immense collection that can be found on the internet already. Second, video recordings can be made without much intrusion. Cameras can be placed in the corner of a room or even almost out of sight, so that subjects are not aware they are recorded (as long as it falls within the remit of the medical ethical agreement and fits the informed consent given by those being recorded). Third, humans experience the world with senses of which sight and hearing are important ones. These senses are often implemented first in human-machine interaction.

Besides video data, the ADVANT facility will also be able to analyse audio data. This type of data has been used in human behaviour research for many years. There are many tools for automatic analysis (e.g. speech recognition) and manual transcription of audio.

Other data types could be eye movements, psychophysiological data such as heartbeat and skin conductance (van Bruinessen et al, 2016), data from activity trackers and questionnaires, but also from news sites, (micro)blogs, and social media, since this data can also be coupled to an event.

Finally, the outcome of the analysis takes the form of transcriptions and annotations, and metadata.

2. THE POSSIBILITY TO COLLABORATE WITH THE DATA

Collaboration in the ADVANT facility takes several forms. First, we will enable data sharing, working on the same data by multiple agents[1], and version control. Second, we will implement a way to stick notes to any part of a transcription. Third, our facility will have a function to conduct a so-called data session.

Data sharing, editing by multiple agents and version control

To be able to work on the same data, access to that data by multiple agents must be organised well. We opt for a central storage facility that can be accessed from anywhere. This facility inflicts access control to users as configured by the administrator of a project.

This facility also makes it possible for multiple agents to work on the same data (parallel editing). There are several things to consider. First, should it be possible to work on the same data at the same time? If so, should changes made by one agent, be made immediately visible in the editor of another agent (e.g. in Transana)? Problems arise when two or more agents are editing the same element, possible resulting in incorrect data. Also, when an automatic analysis tool is working on data, it is rarely equipped to incorporate changes while analysing.

---

1 Agents refers to both human annotators and automatic annotators.

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

Finally, what does parallel editing mean for version control? If one version gets two separate new versions, they need to be merged into one new version. Merging is not always trivial and may need additional human effort and expertise.

If we choose to not enable parallel editing, data can only be edited by one agent at a time. Editing should be denied for an agent if another agent is editing the data. The advantage is that version control is very straightforward. There is only one linear string of consecutive versions as opposed to a branching version structure involving merging.

There are several options for using existing software for our files sharing module. In software development Git is widely used. Git is especially equipped for parallel editing and merging programming code, but it is being expanded towards controlling large files while allowing full version control with Git Large File Store[2]. As developments move quickly, a subtask in ADVANT is to explore options such as Git Large File Storage and other open source versioning systems for large datasets.

Subversion could be another good candidate for it allows for so called locks. A locked file means that users cannot upload a new version of that file. A lock could indicate that the file is not to be edited by anyone other than the agent that locked the file.

Notes

When analysing video or other data, some things cannot be part of the metadata, transcription or annotations, e.g. considerations for later reference, noting examples of interesting parts, making to-do's and asking other researchers questions about certain parts.

Our facility will be able to assign notes to every identifiable part of the data such as media file, every meta-datum, transcription layer, annotation and time interval. A note can have a type (e.g. consideration, example, to-do etc.), subject, body text, time stamp, sender (the latter two automatically assigned) and receivers.

Notes will be kept apart from the data, and they will be searchable.

Sticking notes to parts of the data and the fact that the facility has version control means that notes refer to a certain version of the data. To be able to keep notes in consecutive versions, parts of the data must be identifiable across versions. A possible mechanism is to use identifiers for every part of the data while not reusing identifiers, even after deleting a part.

Data sessions

For much research it is necessary for researchers to sit together to work on certain data in order to deliberate and to further the analysis. We propose a tool in which these so-called data sessions could be done over the internet. The deliberation part could be done through

---

[2] https://git-lfs.github.com/

existing video streaming services such as Skype. Pointing, highlighting, drawing over and editing of the data, however, is managed by our facility. As we pointed out in de section about data sharing, it is undesirable to give people the possibility to edit the data in parallel. Therefore, we propose a broadcasting mechanism in which there is one agent that can edit data and for instance make drawings on video stills. Viewers will see live what the broadcasting agent is doing. Broadcasting control can be handed over to another participant if needed. Moreover, the broadcast can be recorded so that it can also be used as a teaching aid.

As data sessions are done live over the internet, the data that are broadcast should be limited. This also means that it will not be a broadcast of the computer desktop similar to desktop sharing sessions, as this takes too much bandwidth of the internet connection. Special attention should be paid to the simultaneous play back of the video.

Crowd sourcing

By building a very flexible set-up of the front end as well as detailed access control of the data, our facility will be very suitable for crowd sourcing projects. Since all data communications, e.g. between the front end and the central service, are well defined (in a so-called application programming interface or API), it allows for easily connecting our facility to crowd sourcing services such as Amazon Mechanical Turk and Crowdflower.

3. ANNOTATION OF THE DATA

Manual annotation

For transcribing audio and video many tools are already available. Although most tools are very similar, they are targeted at a specific transcription/annotation method, are meant for general purpose and loaded with functions, which makes them hard to use, or it takes expert knowledge to make them work your way.

Below, we describe a few of the most widely used tools.

TranscriberAG can handle many media file formats and long files, and can play back synchronized media files. The tool uses a default transcription convention, but this can be altered by a so called corpus administrator (expert user). Annotation text may have markup for extra visibility of information. It has speaker dictionaries that can be used in multiple transcriptions. The main disadvantage of TranscriberAG is that it has an elaborate installation process that could be overwhelming for non-technical user.

Transana can cope with text, (still) images, and multiple synchronized audio and video files. The tool mainly supports the Jeffersonian notation for transcribing. Transana has elaborate functions for gathering examples and making notes. In the multi user version changes to a transcription are shared immediately with others, allowing for a basic data session. The main

disadvantage is that it is not free for anyone to use.

ANVIL has an adaptable coding scheme, including colour coding of parts of annotations. It can make links between levels of annotation. It has a corpus management tool on board. A disadvantage is that it uses the Java Media Framework, which could give problems for media play back.

NITE XML is a very adaptive toolkit for viewing linguistic data. It has an elaborate data model with cross-links and a sophisticated search engine. The disadvantages lie in the fact that it needs expert knowledge to build the tool you envision. So far, elements for media transcription, such as a wave viewer, are not yet part of this toolkit.

EXMARaLDA consists of several tools for transcription, annotation and corpus management including sophisticated search. One of the tools, FOLKER, is a simplified version of Partiture, the transcription and annotation tool.

ELAN is an elaborate tool for audio and video transcription. It has an adaptive data model consisting of annotation layers (called tiers). It can play back multiple synchronized video files and time-series data. It has a sophisticated search engine. The main disadvantage is that it has so many functions it may come across as bloated and difficult to use for relatively simple tasks.

The developers of the latter two tools have determined that their tools are highly compatible. Both have shown interest to donate their codebase, experience and/or time for the development of the ADVANT facility.

Although all the tools described above are excellent for many tasks, many users prefer a tool that is tailored for the specific tasks within their project. Therefore, the ADVANT facility incorporates a highly adaptable tool for all possible tasks: manual annotation of time aligned data as well as visual context, e.g. object identification in videos, corpus management including editing metadata, managing automatic annotation tools, collaboration, etc.

Automatic annotation

There are many tools for automatic annotation of audio and video. One of the first that comes to mind is speech recognition. Speech recognition is widely used in mobile platforms (Siri on the iPhone, Google Speech on Android) and desktop computers. Although the audio signal needs to be relatively clean (no noise, no husky voice, no accent) it could help the transcription and annotation process in some cases. For instance, in a consulting room, patient and healthcare provider usually interact in a quiet environment without much background noise: the automatic tool does the bulk of the work after which the annotation only has to be corrected.

Other automatic annotation tools analyse gesture or facial expression. This could help research on human interaction beyond the scope of speech. The automatic analysis of

K O N I N K L I J K E   N E D E R L A N D S E
A K A D E M I E   V A N   W E T E N S C H A P P E N

emotions can, for instance, help to monitor the often only bodily expressed emotions in patients. These tools combined could be implemented in a agent for human-machine interaction (e.g. companion robots).

The way in which these tools can be used in the ADVANT facility is as follows. A user simply selects the media files he wants to have analysed. Then, based on that selection, all possible automatic annotation tools are offered for selection. Finally, the user is able to choose the way the output is handled, for instance whether it should become a transcription layer and whether it should be the input of another automatic analysis tool. In this way, tools can be concatenated in a complete workflow. Options for the implementation are Taverna for the workflow and the CLARIN tool Clam[3] for the deployment of automatic analysis tools.

The output of automatic analysis tools can be manually corrected within the front end of the facility (see above). If the user chooses to do so and if the automatic tool gives that possibility, corrected annotations can be fed back to an automatic analysis tool so that the tool can use the corrected data to improve its machine learning models giving better results in future sessions.

– *Beschrijf welke onderdelen/technieken beproefd zijn en welke geheel of gedeeltelijk nieuw?*

Our facility aims for the integration of many tools of which quite a few already exist. However, the integration of these tools and especially the adaptability of our facility are new. See "Technical case" for an extensive description.

Uitdagingen en risico's

– *Beschrijf de belangrijkste technische knelpunten en geef aan hoe deze opgelost zouden kunnen worden.*

There are many technical challenges. First of all, the tool for manual annotation needs a modular structure in order to be flexible. ELAN's modularity is currently insufficient and should be improved for the needs of ADVANT. The Nite XML Toolkit (http://groups.inf.ed.ac.uk/nxt/), a customizable tool to view linguistic data, could be a source of inspiration.

A second challenge is the aggregation of automatic audio and video analysis tools such as speech recognition, speaker recognition, facial expression recognition, gesture recognition and emotion recognition into one service. Often, a tool serves a small niche or is even only a proof of concept. No efforts are made to offer the tool to the (research) community in a standardized way. Many researchers would like to use these tools, but lack of technical

---

[3] http://proycon.github.io/clam/

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

expertise forces them to abandon that idea. We hope to unlock the potential of these tools and perhaps initiate new research into parallel transcriptions of multiple automatic audio and video analysis tools. For instance, speaker recognition can benefit from both audio (voice biometrics) and video (mouthings, gestures) data.

A third challenge is the handling of multiple synchronized media files. Due to a multitude of video and audio encodings, differences in frame rate and so called clock drift1, synchronizing multimedia is very difficult. Especially for researchers making annotations on a frame-by-frame basis, exact synchronization is crucial. ELAN uses platform specific video players and each of those players handles video differently. This makes working on different platforms inconsistent. We propose to aim at using a player that is platform independent, but still performs similar to platform specific software.

A fourth and special challenge is how to handle intellectual property rights (IPR), privacy (ethical questions) and data safety. We are aware of the undeniable relevance of these issues in the case of audiovisual data. The solution to these questions may differ per data collection. We will take all possible security measures and install password systems etc. depending on property rights of the data (e.g., media data) and privacy (data without informed consent from the participants, and data revealing personal or identifiable details). The facility will offer the option of rendering faces unrecognizable. With regard to data safety, we will make very clear and explicit arrangements with the archive in which data collections will be stored (e.g., TLA, Beeld en Geluid).

– *Beschrijf de belangrijkste risico's.*

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The first risk of a large project like the ADVANT facility is the inability to incorporate changes in the requirements. If we would make an elaborate blueprint that takes a long time to implement, we would not be able to adjust our plans to developments that influence our project. Therefore, we propose a more agile approach in which we produce new components of the facility on a regular basis. Every new component is functional and can be tested and used with other existing components. In this way, our facility steadily grows while it stays usable and therefore useful at all times.

A second area of risk is the maintenance of the facility, especially the responsibility thereof. In order to keep the facility running as a high quality service, a small team of experts is required to keep it up to date with new developments, adapt it to a growing amount of data, and resolve problems when they arise.

A third risk is the lack of available automatic audio and video analysis tools. Third parties will be asked to make them available. Possibly, very few parties are willing to do that and if they do their user policies must be compatible to the ones we offer to our users.
In case our facility lacks a specific tool we would intend to develop it.

B. INBEDDING

Hoe past dit voorstel in het (internationale) landschap van grote onderzoeksfaciliteiten?

ADVANT is unique in its focus on video annotation and analysis, and in the possibilities it creates for collaboration between researchers.

The ADVANT consortium has found that also internationally ADVANT is unique and potentially interesting for international researchers. There are interesting initiatives (e.g. at the Institut für Deutsche Sprache, Mannheim) that are however less ambitious since they aim at supporting only one type of research.

ADVANT will evolve in time, i.e. when other systems become available ADVANT will be updated. That way we can guarantee that we are always up to date with the infrastructure. This tool for video-analysis would exist next to a large number of facilities developed for the analysis of texts. ADVANT will be an addition to existing language research infractucture that is primarily concerned with the analysis of written language (e.g. the Desideria initiative). Through our collaboration within CLARIAH, and our contacts with the Netherlands Centre for Humanities and Technology (CHAT-NL), the ADVANT consortium can guarantee that we will make use of existing tools and facilities whenever possible, and will develop new tools in consultation of these research communities.

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

> – *Hoe wordt de nationale toegang gegarandeerd?*

Access control

The ADVANT facility is primarily targeted at scientific researchers. We aim to use a federated authentication service so that a researcher can access the facility with the credentials of the institute. Potential services are eduGAIN [https://wiki.edugain.org/EduGAIN] or CLARIN Service Provider Federation [https://www.clarin.eu/content/service-provider-federation].

Besides researchers, the facility should also be accessible to people outside the scientific community. For instance, projects may want to involve members of the public for certain annotation tasks, i.e. crowd-sourcing. Therefore, the facility should also incorporate a registration service.

National access to ADVANT is guaranteed in the following ways:

- It is offered through the CLARIN services website, in CLARIN standard formats, which makes it compatible with a large range of other program;
- It is stored on a server at the CLARIAH centre Beeld & Geluid (Institute of Sound and Vision, Hilversum, the Netherlands);
- It has been developed in close collaboration with various types of users from various research disciplines in the Netherlands so it is well-fitted to the wishes and work flows of these users (user centered design).

> – *Sluit het aan op reeds bestaande faciliteiten?*

Currently, the most well-known tool for the analysis of video is ELAN. ELAN, developed at the MPI Nijmegen, offers a wide range of functionalities, but has become less and less user friendly with every new addition. ELAN does not offer options for version control and researcher collaboration; in practice, researchers save their annotations on local computers and e-mail files to colleagues when working together on a project. Other software available for the analysis of video are Exmaralda, Nite XML, and Observer (Noldus). In contrast to ADVANT, these require technical expertise to fit them to specific research purposes.

> – *Zijn er voor zover bekend vergelijkbare ideeën (of al bestaande faciliteiten) in het buitenland? Zo ja, zou Nederland een aparte nationale faciliteit moeten hebben of betreft dit een internationale faciliteit op Europees of mondiaal niveau?*

In the field of research into language and social interaction, there is currently a German initiative from the IDS (Institut für Deutsche Sprache) to find funding for the development of a tool based on Exmaralda which includes the option of collaboration. Once ADVANT is ready to be developed, the German tool (in case it has been built) could serve as basis for the development of ADVANT. The IDS is open to the co-development of ADVANT. The

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

consortium that collaborates in the ADVANT project is particularly strong and broadly composed, consisting of researchers from different fields within the humanities, the social sciences, and the medical sciences.

Hoe past het voorstel bij de NL sterktes van onderzoek?

The ADVANT facility would strengthen several fields of research excellence in the Netherlands. Dutch research is internationally strong in experimental multi-modality research (e.g. Kramer, Swerts, Özyürek), computational linguistics (e.g. Van den Bosch, Bos, de Jong, Odijk, Vossen), medical communication research (e.g. van Dulmen, Bensing), media studies (e.g. van Dijck, Noordegraaf), and social interaction research (e.g. Koole, te Molder). These fields of research depend on the use of video data. The ADVANT facility will enable these fields to maintain and expand their present strengths.

Beschrijf de voordelen/belang voor NL indien zo'n faciliteit zou worden gerealiseerd. Dit mogen zowel wetenschappelijke als economische of maatschappelijke voordelen zijn.

ADVANT will offer new opportunities to different research areas such as multimodality studies, computational linguistics, medical communication research, media studies, and social interaction research, as well as to societal institutes such in the field of broadcasting, and institutions such as health care and police. Particularly:

1) The absence of an integrated video-analysis facility such as ADVANT as well as the international absence of the development of such a facility will make cutting-edge, large-scale research in the Netherlands possible in all the scientific areas mentioned above;
2) the ADVANT facility will encompass all sorts of cutting-edge technological developments and near-future possibilities that will put the Netherlands on the map within the scientific fields mentioned in this proposal;
3) the ADVANT facility will set an international standard for multi-modal video-analysis, both in research and in society;
4) ADVANT has the potential of becoming a Dutch export product.

C. ORGANISATIE en FINANCIËN

Organisatie

Geef aan welke partijen/expertise nodig zijn voor de ontwikkeling van deze faciliteit. Geef ook aan of en zo ja hoe deze al zijn betrokken.

In preparation of this proposal the ADVANT project has contacted relevant parties and experts that are necessary for the development of this facility. Contacts have been established with three different types of parties relevant for this project:
1. CLARIN/CLARIAH for collaboration and standards for data archiving and annotation. We have agreed for ADVANT to be included in the focus area of Media Studies (Work Package 5 of CLARIAH).

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

2. Researchers from different scientific fields who will want to use the ADVANT facility. Several researchers have expressed needs that have been included in the present design of the facility. The future development of ADVANT will remain to be supported by these members of the research community.
3. The Netherlands Centre for Humanities and Technology (CHAT-NL) may become a major environment for the consultation of and collaboration with potential users of ADVANT. The ADVANT consortium is involved in the talks about the development of CHAT-NL.
4. The ADVANT advisory and user group will be available for consultation and advice.
5. ICT-developers of tools that will be included in ADVANT. We are in close contact with the ICT staff of ELAN, and further with developers of video-analysis software such as face and emotion recognition. Also, the ADVANT project group includes expertise on automatic annotation as well as different types of search tools.

Beschrijf de (mogelijke) organisatiestructuur. Geef ook aan of er al een begin van organisatievorming is.

In terms of the organization of ADVANT, the University of Groningen 'Centre for Digital Humanities' will host the project and will collaborate with the other participating institutions, Radboud University, Utrecht University, NIVEL, and the Beeld en Geluid Institute. ICT and research staff will be located and supervised at these different institutions. In terms of institutional embedding, ADVANT will be a part of the CLARIAH focus area of Media Studies.

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

Financiering			
Geef een ruwe schatting van de benodigde kosten voor de realisatie van deze faciliteit. Maak onderscheid in kosten voor: 1) Ontwikkeling; 2) Bouw, 3) Exploitatie.			

Development				
Annotation tool				
3,5	ICT-developers	1.0 (5 years)	à 62.000 / year	1.085.000
Aggregation of automatic audio and video tools				
3,5	ICT-developers	1.0 (5 years)	à 62.000 / year	1.085.000
Collaboration service				
3,5	ICT-developers	1.0 (5 years)	à 62.000 / year	1.085.000
40	Test users	1.0 (2 days)	à 100 / day	8.000
2	Scientific staff : annotation testing (assistant professor)	0.6 (2 years)	à 60.000 / year	240.000
1	Scientific staff: analysis testing (assistant professor)	0.6 (2 years)	À 60.000 / year	120.000
3	Supervision (full professor)	0.1 (5 years)	à 12.000 / year	180.000
Maintenance				
	hardware	10 years	à 80.000 / year	800.000
3	ICT-developers	1.0 (10 years)	à 62.000 / year	1.860.000
2	Scientific staff (assistant professor)	0.1 (10 years)	à 10.000 / year	200.000

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

2	Supervision (full professor)	0.025 (10 years)	à 6.000 / year	60.000
Total				6.723.000

Geef een globale beschrijving van de business case. Hoe zou deze faciliteit gefinancierd kunnen worden? Ga hierbij in de drie fases: ontwikkeling, bouw en exploitatie.

Since the greater part of the costs of the ADVANT facility is personnel, the facility is largely dependent on subsidies. The development and the actual building of the facility are designed as a process of five years in which ICT personnel and scientific staff will collaborate. Potential subsidizing instances include KNAW, NWO, CLARIAH, and the EU, in particular the European Science Foundation (ESF). The exploitation, maintenance, and up-to-date keeping of the facility are scheduled as a ten-year process for which again the main costs are ICT and research staff. Again the instances mentioned above will be approached for funding. Should the ADVANT facility be selected by KNAW for its research agenda, we will gladly consult the board of KNAW to advise us on funding options.

D. VERDERE ONTWIKKELING

Beschrijf wat er moet gebeuren om deze faciliteit verder te ontwikkelen. Ga in op de belangrijkste knelpunten die opgelost moeten worden.

The ADVANT facility will be initially developed and tested over a period of 5 years. The development will include a restructuring of ELAN as the basis for ADVANT, and the implementation of different tools for the annotation and analysis of video data into ADVANT. Testing will be done with occasional test users, and further in two projects. In the first project, the annotation facilities will be tested by collecting a video corpus of everyday spoken Dutch and annotating this in ADVANT. In the second, the analysis facilities will be tested by integrating NIVEL data from different sources as facilitated by ADVANT in order to study the explained variance in relevant communication outcome measures (patient, visit and health outcomes). In addition, the validity of the different observational ADVANT tools from patient, clinician and researcher perspective will be studied. The researchers of the two test projects will collaborate in designing solutions for ethical dilemmas related to sharing privacy sensitive data.

Following the development of the facility, maintenance will take over. Since maintenance will be concerned with all elements of the facility, these activities will have to start as soon as elements of ADVANT are in place. Therefore, maintenance, although it logically follows development, will in practice overlap with the 5-year development period.

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The advisory board / user group will meet every year from the start of the project.

We expect that over time the research community will demand new functionalities. These may include new modules for new types of users, new data types or new types of collaboration. Also, automatic analysis tools are under constant development and new tools will emerge. ADVANT has to adapt to these developments, i.e. they will have to be incorporated to keep ADVANT a state-of-the-art facility. Therefore, the ADVANT team will have to keep close contact with users, scientists and developers. Furthermore, since the facility may store sensitive data there must be a constant effort to update security to the highest standards.

Finally, ADVANT may grow rapidly over time. The facility will have to be scaled to the amount of data and users.

Geef aan wat de ontwikkeltermijn voor deze faciliteit ongeveer zou kunnen zijn.

For a maintained version of ADVANT to be operational in 2025, the ideal starting year is 2017.

Development	2017-2022
Test research projects	2019-2021
Prototype	2022
Ready for external users	2025
Maintenance	2018-2028

References

Bruinessen, I.R. van, Ende, I. van den, Visser, N.C., Dulmen, S. van. (2016). The impact of watching educational patient-provider interactions in oncology care on analogue patients' physiological arousal and information recall. *Patient Education and Counseling* (in press)

Di Blasi, Z., Harkness, E., Ernst, E., Georgiou, A., Kleijnen, J. (2001). Influence of context effects on health outcomes: A systematic review. *Lancet* 357, 757–762.

Fogarty, L.A., Curbow, B.A., Wingard, J.R., McDonnell, K., Somerfield, M.R. (1999). Can 40 seconds of compassion reduce patient anxiety? *Journal of Clinical Oncology* 17, 371.

Goldin-Meadow, Susan. (2012). Homesign: gesture tot language In Roland Pfau, Markus Steinbach, & Bencie Woll (Eds.), *Sign Language. An international handbook* (pp. 601-625). Berlin/Boston: De Gruyter Mouton.

Heritage, J. (2012). 'Epistemics in Action: Action Formation and Territories of Knowledge.' *Research on Language and Social Interaction* 45:1-29

K O N I N K L I J K E   N E D E R L A N D S E
A K A D E M I E   V A N   W E T E N S C H A P P E N

Mistiaen, P., Osch, M. van, Vliet, L. Van, Blasi, Z. di, Bensing, J., Dulmen, S. van. (2016). The effect of patient-provider communication on pain: a systematic review. *European Journal of Pain* (in press)

Novack, Miriam A., Goldin-Meadow, Susan, & Woodward, Amanda L. (2015). Learning from gesture: How early does it happen? *Cognition*, 142, 138-147. doi:http://dx.doi.org/10.1016/j.cognition.2015.05.018X

Sandler, Wendy, Meir, Irit, Padden, Carol A., & Aronoff, Mark. (2005). The emergence of grammar: systematic structure in a new language. *Proceedings of the National Academy of Sciences*, 102(7), 2661-2665.

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, 305(5691), 1779-1782. Retrieved from <Go to ISI>://000224039500043X

Roy, B. C., Frank, M. C., & Roy, D. (2012). Relating activity contexts to early word learning in dense longitudinal data. In: *Proceedings of the 34th Annual Cognitive Science Conference*.

Rossano, Federico (2013) Gaze in conversation, in: J. Sidnell & T. Stivers (eds.) *Handbook of Conversation Analysis*, Oxford: Wiley,

Rossi, Giovanni (2015) *The Request System in Italian*, PhD dissertation: Radboud University Nijmegen

Tan, Ed. (2014) Engaged and Detached Film Viewing. *Cognitive Media Theory*, Paul Taberham and Ted Nannicelli, eds. New York: Routledge

Tomasselo, Michael (2010), *Origins of Human Communication*, Boston: MIT Press

Tur, Gokhan, Andreas Stolcke, Lynn Voss, Stanley Peters, Dilek Hakkani-Tür, John Dowding, Benoit Favre et al. "The CALO meeting assistant system." *Audio, Speech, and Language Processing*, IEEE Transactions on 18, no. 6 (2010): 1601-1611.

Whitehead, Kevin (2011) Some uses of head nods in "third position" in talk-in-interaction, *Gesture*, Volume 11, Number 2, 2011, pp. 103-122