A Multimodal Multi-Device Discourse and Dialogue Infrastructure for Collaborative Decision Making in Medicine

Daniel Sonntag and Christian Schulz

German Research Center for AI (DFKI)
Stuhlsatzenhausweg 3, 66123 Saarbruecken, Germany

Abstract. The dialogue components we developed provide the infrastructure of the disseminated industrial prototype Radspeech—a semantic speech dialogue system for radiologists. The major contribution of this paper is the description of a new speech-based interaction scenario of Radspeech where two radiologists use two independent but related mobile speech devices (iPad and iPhone), and collaborate via a connected large screen installation using related speech commands. With traditional user interfaces, users may browse or explore patient data, but little to no help is given when it comes to structuring the collaborative user input and annotate radiology images in real-time with ontology-based medical annotations. A distinctive feature is that the interaction design includes the screens of the mobile devices for touchscreen interaction for more complex tasks rather than the simpler ones such as a mere remote control of the image display on the large screen.

1 Introduction

Over the last several years, the market for speech technology has seen significant developments [7] and powerful commercial off-the-shelf solutions for speech recognition (ASR) or speech synthesis (TTS). For industrial application tasks such medical radiology, we implemented a discourse and dialogue infrastructure for semantic access to structured and unstructured information repositories [13]. The infrastructure is based on the assumption that in order to support a rapid dialogue system engineering process for domain-specific dialogue applications, an ontology-based approach should be followed for all internal and external processing steps.

The idea of semantic web data structures [1] has provided new opportunities for *semantically-enabled user interfaces*. The explicit representation of the *meaning* of data allows us to (1) transcend traditional keyboard and mouse interaction metaphors, and (2) provide representation structures for more complex, collaborative interaction scenarios that may even combine mobile and terminal-based interaction [11]. The collaborative speech-based interaction scenario in a multiparty setting for medical decision-making, namely in radiology, will be the focus of this paper. We relied on a semantic web toolbox for ontology-based dialogue

engineering. In previous implementation work of this large-scale project (THESEUS[1]), we provided a technical solution for the two challenges of engineering ontological domain extensions and debugging functional modules [14].

In this paper, we basically provide two new contributions. First, we provide distinctive features of our new dialogue infrastructure for radiology and explain the first speech-based annotation system for this task. Second, we discuss the radiology interaction system in greater detail and explain the implemented dialogue sequences which constitute a running demo system at our partner hospital in Erlangen. Thereby we also focus on the special technical components and implementation aspects that are needed to convey the requirements of dialogical interaction in a medical application domain. With traditional user interfaces in the radiology domain (most of which are desktop-based monomodal keyboard input systems), users may browse or explore patient data, but little to no help is given when it comes to structuring the collaborative user input and annotate radiology images in real-time with ontology-based medical annotations. To meet these objectives, we implemented a distributed, ontology-based dialogue system architecture where every major component can be run on a different host (including the graphical interface and audio streaming on mobile devices). This increases the scalability of the overall system.

In earlier projects [15, 8] we integrated different sub-components into multimodal interaction systems. Thereby, hub-and-spoke dialogue frameworks played a major role [9]. We also learned some lessons which we use as guidelines in the development of *semantic* dialogue systems [5]; the whole architecture can be found in [10]. Thereby, the dialogue system acts as the middleware between the clients and the backend services that hide complexity from the user by presenting aggregated ontological data. One of the resulting speech system, RadSpeech (http://www.youtube.com/watch?v=uBiN119_wvg), is the implementation of a multimodal dialogue system for structured radiology reports.

2 Special Radiology Task Requirements and Implementation

In the MEDICO use case, we work on the direct industrial dissemination of a medical dialogue system prototype. Recently, structured reporting was introduced in radiology that allows radiologists to use predefined standardised forms for a limited but growing number of specific examinations. However, radiologists feel restricted by these standardised forms and fear a decrease in focus and eye dwell time on the images [2, 16]. As a result, the acceptance for structured reporting is still low among radiologists while referring physicians and hospital administrative staff are generally supportive of structured standardised reporting since it eases the communication with the radiologists and can be used more easily for further processing.

[1] This work is part of THESEUS-RadSpeech (see www.dfki.de/RadSpeech/) to implement dialogue applications for medical use case scenarios. It has been supported by the German Federal Ministry of Economics and Technology (01MQ07016).

We implemented the first mobile dialogue system for radiology annotations, which is tuned for the standardised radiology reporting process. Our solution not only provides more robustness compared to speech-to-text systems (we use a rather small, dedicated, and context-based speech grammar which is also very robust to background noise), it also fits very well into new radiology reporting processes which will be established in Germany and the U.S. over the next several years: in structured reporting you directly have to create database entries of a special vocabulary (according to a medical ontology) instead of text. The semantic dialogue system presented by RadSpeech should be used to ask questions about the image annotations while engaging the clinician in a natural speech dialogue. Different semantic views of the same medical images (such as structural, functional, and disease aspects) can be explicitly stated, integrated, and asked for. This is the essential part of the knowledge acquisition process during the speech dialogue: the grammar of the ASR system only accepts the annotations of a specific grammar which stems from the used medical ontologies; this allows us to reject arbitrary annotations and recognitions with low probability which makes the system very reliable. Upon touching a region on the interaction device, the ASR is activated. After recognition, the speech and gesture modalities are fused into a complex annotation using a combination of medical ontologies. For disease annotations for example, the complete Radlex (http://www.radlex.org) terminology can be used, we but we also use an OWL-version of ICD-10 [4] and FMA [3]. With this dedicated grammar, the annotation accuracy of single term annotations is above 96%, whereby multi-term annotations (three annotations in one speech command) are difficult to handle (informal evaluation).

Another central requirement is the need for different graphical user interfaces and contents on the mobile devices and the screen. Currently, radiology working stations must feature an FDA clearing (http://www.fda.gov/) meaning that only cleared (mobile) devices can be used for active diagnostic purposes. Following this sub-requirement, we can use the FDA-cleared iPad (or iPhone) for diagnostic purposes and the big screen for non-diagnostic ones. As a result, the image series should only be manipulated and annotated on the mobile interaction devices, whereas key images are displayed on the big screen, thereby allowing to synchronise individual annotations stemming from multiple FDA-cleared devices. A very nice feature of the resulting interaction scenario which takes on this special requirement is the effect that, on the mobile device, we can implement the multimodal setting with a mobile image series viewer which runs through the slices (see, e.g., the commercial DICOM app MIM, http://www.mimsoftware.com). The ASR activates upon touch, and the manipulation of the images can be done using touch instead of trying to do all of these things using speech and the big touchscreen—thereby making a virtue of necessity.

In addition to ASR, dialogue tasks include the interpretation of the speech signal and other input modalities, the context-based generation of multimedia presentations, and the modelling of discourse structures. According to the utility issues and medical user requirements we identified (system robustness/usability and processing transparency play the major roles), we provide for a special rule-

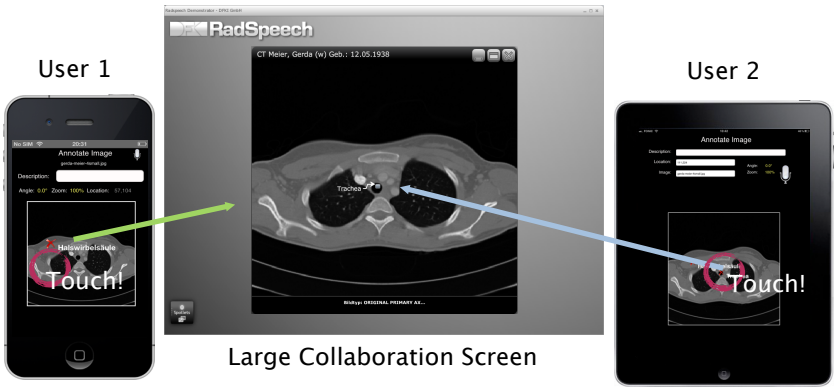

Fig. 1. Multimodal speech dialogue scenario with multiple input/output devices

based fusion engine of different input modalities such as speech and pointing gestures. We use a production-rules-based fusion and discourse engine which follows the implementation in [6]. Within the dialogue infrastructure, this component plays a major role since it provides basic and configurable dialogue processing capabilities that can be adapted to specific industrial application scenarios (e.g., the co-ordination of pointing gestures and ASR activation on the medical images). More processing robustness is achieved through the application of a special robust parsing feature in the context of RDF graphs as a result of the input parsing process. The domain-specific dialogue application is able to process the following medical multi-user-system dialogue on multiple devices (the cancer annotation is replaced by a simple anatomy annotation for illustration):

1 **U1:** "Show me the CTs, last examination, patient XY."
2 **S:** Shows corresponding patient CT studies as DICOM picture series and MR videos.
3 **U1:** "Show me the internal organs: lungs, liver, then spleen and colon."
4 **S:** Shows corresponding patient image data according to referral record on the iPad.
5 **U1:** "Annotate this picture with Heart (+ pointing gesture on the iPad)
6 **S:** "Picture has been annotated with Heart."
7 **U1:** "Show it on screen."
8 **S:** "Shows patient XY on the large screen, automatically rendering the picture with the heart annotation in the foreground."
9 **U2:** "and Heart chamber (+ pointing gesture on the iPhone)"
10 **S:** Adds the second annotation on screen.
11 **U1:** "Synchronise annotations with my iPad".
12 **S:** "Shows new annotation on the iPad".
13 **U2:** "Search for similar patients."
14 **S:** "The search obtained this list of patients with similar annotations including 'Heart' and 'Heart chamber'."
15 **U1:** "Okay."

Our system then switches to the comparative records to help the radiologist in the differential diagnosis of the suspicious case, before the next organ (e.g., liver) is examined in the collaborative session of the two doctors. The semantic search for similar cases is implemented by a SPARQL engine which computes semantic similarities between the ontology concepts on the images and the image series in the databases (see [12]).

3 Multimodal Interaction in the Multi-Party Setting

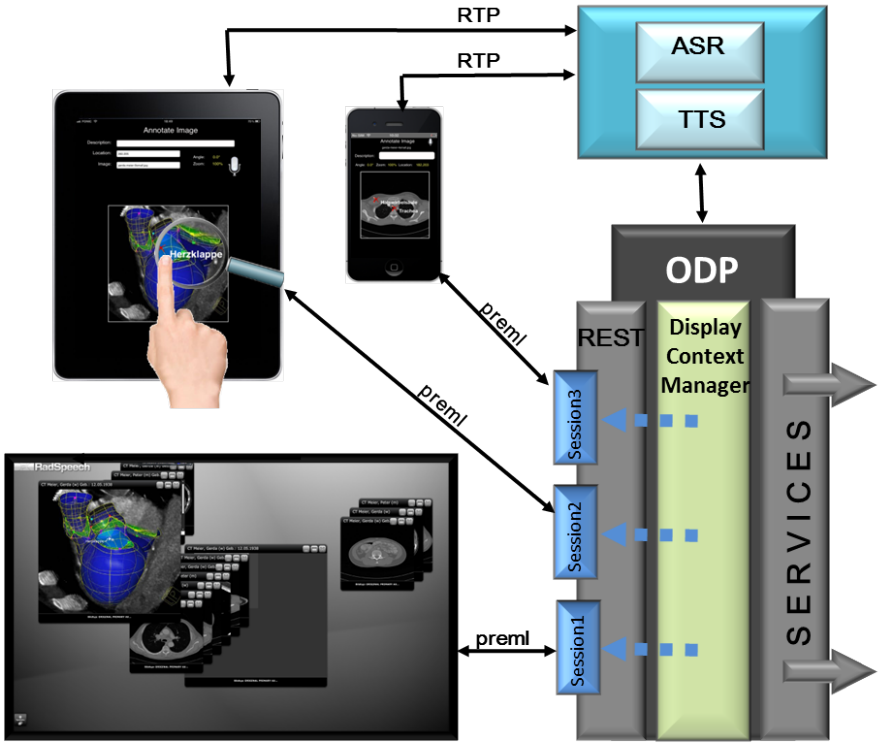

Fig. 2. The multi-party/multisession infrastructure: two active users on iPad and iPhone

For the collaborative scenario we need to be able to model the activity of each user that is connected to the infrastructure. The challenge in this setting is that, in our infrastructure, the input/output communication assigned to every individual user must be processed separately in one individual dialogue session. This architectural decision was made in the initial setting to cope with (deictic) dialogue references in the dialogue history and allow for a coherent representation of a specific session's working memory. In addition, we handle multi-party dialogue input by multiple devices. As a result, a single dialogue session has been restricted to a single user. Accordingly, a multi-session operation is our answer to the new multi-user requirement (towards the direction that one user indicates something and the second can refer to it (future work)). In figure 2, the most relevant parts of the implementations concerning the multi-party scenario are displayed.

The Ontology-based Dialogue System ODP represents the central part in the architecture and handles the communication among the external device components through multiple channels (i.e., handshaking/messaging among clients, controlling the speech server to listen to audio streams, and the like). In addition, it provides the multisession infrastructure based on a rule engine in order to instantiate several dialogue system sessions in the desired multi-device setting. At this point, we want to emphasise the fact that all peripheral devices (our mobile devices such as iPhones or iPads) are associated with one session for one device respectively, which is hold throughout the dialogue.

As a consequence, an event within one session will not directly affect the state of another session. In what follows, we will illustrate how we extend our infrastructure by implementing a multi-party-enabled Display Context Manager to meet the new requirements: to implement collaborative scenarios where actions on peripheral devices actually have an effect on other users (and corresponding dialogue sessions) connected to the dialogue system.

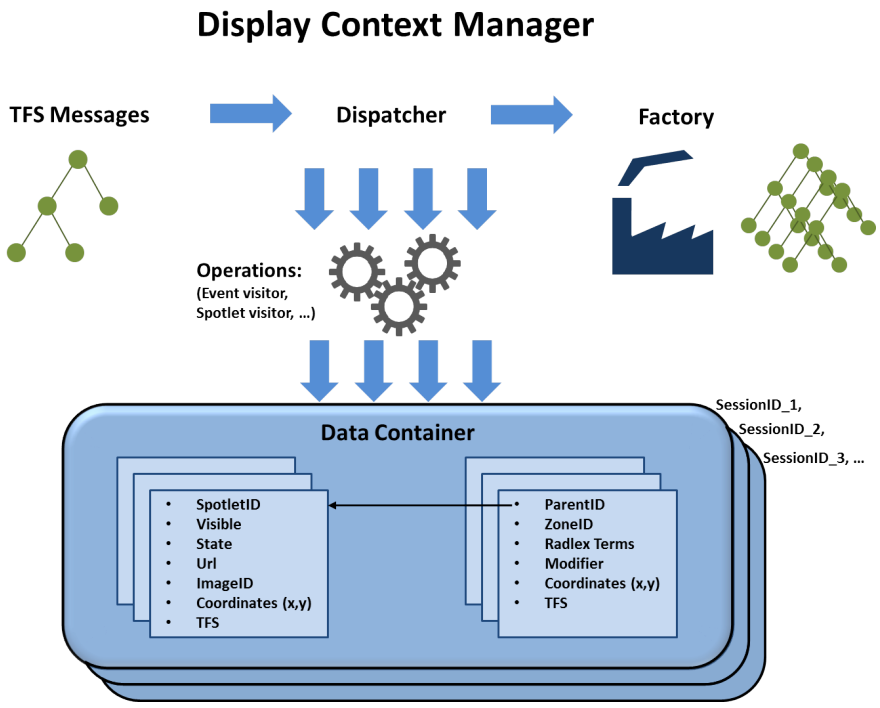

Fig. 3. The display context manager and data container

The Display Context Manager is in charge of dispatching the command messages which are also ontological instances, with an internal representation as

Typed Feature Structures (TFS). The corresponding TFS is then handed over to proper operational components possessing exclusive access to write on medical data records. The medical data that are subject to the expert's analysis and manipulation are located inside a data container, maintaining so-called *spotlets* and *zones*. Spotlets are containers for meta-information of patient images (e.g., DICOM meta data about the image recording process in the hospital such as date, time, image modality, and the patient's name). Zones are containers administrating the annotations associated with the spotlets. Medical data inside the container are instantiated as soon as the user retrieves patient images at the backend service by using the dialogue engine. In this sense, the life cycle of the data in the working memory is determined by the image retrieval process and the length of a session. However, a user has the option to commit annotation results of his or her diagnostic analysis to dedicated servers as backend services at any point during a session.

```
<object type="radspeech#ImageInputEvent">
  <slot name="odp#hasContent">
    <object type="medico#ImageAnnotation">
      <slot name="odp#isSelected"/>
      <slot name="medico#annotation"/>
    </object>
  </slot>
  <slot name="odp#action">
    <value type="String">
        <![CDATA[select_zone]]>
    </value>
  </slot>
  <slot name="radspeech#id">
    <value type="String">
        <![CDATA[1]]>
    </value>
  </slot>
  <slot name="comet#xCoordinate">
    <value type="Float">
        <![CDATA[252]]>
    </value>
  </slot>
  <slot name="comet#yCoordinate">
    <value type="Float">
        <![CDATA[190]]>
    </value>
  </slot>
</object>
```

```
<object type="medico#AnnotateTask">
  <slot name="odp#hasContent">
    <object type="medico#MedicoSpotlet"/>
  </slot>
  <slot name="odp#hasContent">
    <object type="medico#ImageAnnotation">
      <slot name="medico#annotation">
        <value type="String">
          <![CDATA[herzklappe]]>
        </value>
      </slot>
    </object>
  </slot>
  <slot name="medico#linked">
    <object type="medico#Modifier">
      <slot name="radspeech#modifier">
        <value type="String">
          <![CDATA[add_ann]]>
        </value>
      </slot>
    </object>
  </slot>
</object>
```

Fig. 4. TFS messages that represent different types of events which in turn invoke different classes of operations

Operations on data are categorised into different levels of intrusion. For instance, the deictic input on the user interface can be associated with a different operation than voice input which may contain the user's demand to attach an annotation to a medical image. In particular, the first operation is relevant to inform the Display Context about what the attentional focus of the user is (e.g., selecting medical images or performing image annotations), whereas the second operation performs data manipulation in a zone that belongs to some spotlet representing the selected medical image on the mobile device of the respective user.

Figure 4 (on the left) shows the corresponding TFS message that is transferred by a select gesture, while the TFS message (figure 4, on the right) encapsulates an annotation task triggered by voice. Please note, however, that the level of intrusion is independent of the input modality, a voice command may easily serve to change the attentional focus by saying "Open the patient's last image annotation," for example.

Table 1. Overview of the modelled collaborative interactions

Device Type	Gesture/Voice Input	Multisession Action
iPad	1-SwipeToRight(onImage)	propagate image to screen
	1-SwipeToRight(onMainview)	propagate all images to screen
	"Show the images on the screen."	propagate all images to screen
	1-swipeToLeft(onImage)	remove image from screen
	"Synchronise with the screen."	synchronise actions on screen
	"Stop synchronisation with screen."	desynchronise actions on screen
	doubleTap(mainviewFooterCenter)	close the patient file/images (both)
	longPress(ann)	delete annotation (both)
	select(ann)+ *"Delete annotation."*	delete annotation (both)
	doubleTap(imageviewFooterRight)	delete all annotations of image (both)
	drag(ann)	repositioning the annotation (both)
iPhone	1-SwipeToRight(belowImage)	synchronise actions on screen
	1-SwipeToLeft(belowImage)	desynchronise actions on screen
	"Synchronise with the screen."	synchronise actions on screen
	"Stop synchronisation with screen."	desynchronise actions on screen
	longPress(annotation)	delete annotation (both)
	select(ann)+ *"Delete annotation."*	delete annotation (both)
	drag(ann)	repositioning the annotation (both)

In order for the multi-session scenario to use inputs from different users, we have implemented a class of operations that has the permission to make manipulations even on data which do not belong to the same session. As pointed out in the lower part of figure 3, each data container is assigned to a session ID.

Depending on the type of operation, the Display Context Manager identifies the corresponding session ID that is connected to the data container. In this way, we are able to model a process that a user is able to perform actions on an iPad whereupon the display content changes and displays further related results on a big screen. Table 1 shows an overview of the basic multi-session interactions that support gesture and voice inputs for the setting where a mobile device propagates its contents to the big screen. For example, on the iPad the propagation of all manipulations of the images is only executable in the main view where all images are displayed. After manipulation, all annotation activities will be mirrored to and synchronised with the big screen.This refers to the actions in table 1 that are indicated by '(both)'.

The second user or additional passive user groups might then inspect the results of the dialogue-based annotation and diagnostic analysis in real-time on the big screen. In particular, the operations that are executed within the session

dedicated to the iPad have not only access to the data container representing the display contents of the iPad, but also the data container that is responsible for the display content of the big screen. The synchronisation of manipulation behavior and TFS data between multiple data containers is achieved by an operation that enables instances of other sessions to obey to operations executed by the session in command. This means amongst other things that only the user who opens a session is allowed to make his or her actions shareable to other sessions.

Besides providing a mechanism to manipulate meta information of data containers regardless of the device the command is issued from, we also have to make sure that the result reaches the correct recipient among the sessions. Again depending on the type of operation, the Display Context Manager detects the corresponding working memory being associated with a particular session/device on the basis of the session ID. After the operation has been executed on the data in terms of updating its internal state, the dispatching mechanism selects a factory method to produce the appropriate TFS result.

Based on the identified working memory, the corresponding update rule inside the dialogue engine instance fires in response to the created TFS object that wraps the modified state of spotlets and zones.

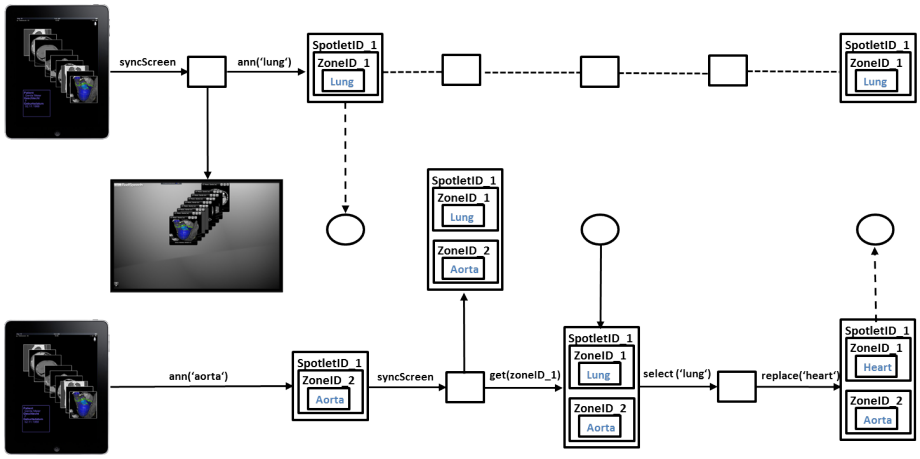

Fig. 5. Workflow of a collaborative scenario

The workflow of the collaborative scenario is shown in figure 5, where the behaviour of the multi-session operations between multiple devices is outlined. The chart demonstrates a collaborative interaction example where an annotation of the first user is overwritten/corrected (or potentially specified in more detail) by another user while using the shared view on the big screen.

First, user 1 (upper iPad) propagates all relevant images of the patient being treated to the big screen. Then, user 1 annotates the zone with the id 'ZoneID_1' of the image referring to 'SpotletID_1' with the term *lung*.

Meanwhile, another user (user 2, lower iPad) annotates the same image with the term *aorta* but in another zone. The propagation of the annotation event by the second user allows the Display Context Manager to unify the annotations assigned to the same image and display them both on the screen.

Subsequently, the second user disagrees with the annotation of the first user for illustration. First she pulls the annotations of the image on the screen to her device (which is implemented as an update operation similar to subversion systems), namely the annotation she wants to correct. Only at the point when the second user obtains the annotation of the first user on her own device, she is able to replace the annotation in question. In turn, this manipulation of the zone (replacing *lung* with *heart* by a voice command) will be reflected on the big screen. In this way, we obtained a clear "speech co-operation policy" and avoided too complex synchronisation behaviours, conflict solution strategies, and recovery mechanisms for unification failures. (Please note that the case with a remote client is slightly different; here the *syncScreen* function synchronises with the big screen and the remote iPad.) Our next steps will include the evaluation of the range of multi-session co-references and co-reference resolution strategies we ought to address when it comes to model more comprehensive collaborative multi-session scenarios.

4 Conclusion

Today, medical images have become indispensable for detecting and differentiating pathologies, planning interventions, and monitoring treatments. Our dialogue platform provides a technical solution for the dissemination challenge into industrial environments, namely an application for a collaborative radiology scenario. Our new prototypical dialogue system provides two radiologist with the ability to, first, review images when outside the laboratory on mobile devices, and second, collaboratively annotate important image regions while using speech and gestures on multiple mobile devices while co-operating in front of a large synchronised touchscreen installation. Currently, the system is part of a larger clinical study about the acquisition of medical image semantics at Siemens Healthcare, the University Hospital in Erlangen, and the Imaging Science Institute (ISI). In future work, we will pursue the idea of multi-session dialogue management in order to allow for more complex user interactions such as "What do you think about this lesion? + pointing gesture (user 1)," user 2: "—it's a difficult case, but I think it's a subtype of Non-Hodgkin Lymphoma." Thereby, we would extend our first Radspeech scenario (http://www.youtube.com/watch?v=uBiN119_wvg) not only to the collaboration described here, but to the highly desired multi-session fusion scenario.

References

1. D. Fensel, J. A. Hendler, H. Lieberman, and W. Wahlster, editors. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, 2003.
2. F. M. Hall. The radiology report of the future. *Radiology*, 251(2):313–316, 2009.
3. J. L. Mejino, D. L. Rubin, and J. F. Brinkley. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. In *Proc. of AMIA Symposium*, pages 465–469, 2008.
4. M. Möller, P. Ernst, A. Dengel, and D. Sonntag. Representing the international classification of diseases version 10 in OWL. In *Proc. of the International Conference on Knowledge Engineering and Ontology Development (KEOD)*, Valencia, Spain, 25-28 October 2010.
5. S. Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.
6. N. Pfleger. FADE - An Integrated Approach to Multimodal Fusion and Discourse Processing. In *Proceedings of the Dotoral Spotlight at ICMI 2005*, Trento, Italy, 2005.
7. R. Pieraccini and J. Huerta. Where do we go from here? research and commercial spoken dialog systems. In *Proceedings of the 6th SIGDdial Worlshop on Discourse and Dialogue*, pages 1–10, September 2005.
8. N. Reithinger, D. Fedeler, A. Kumar, C. Lauer, E. Pecourt, and L. Romary. MIAMM - A Multimodal Dialogue System Using Haptics. In J. van Kuppevelt, L. Dybkjaer, and N. O. Bernsen, editors, *Advances in Natural Multimodal Dialogue Systems*. Springer, 2005.
9. N. Reithinger and D. Sonntag. An integration framework for a mobile multimodal dialogue system accessing the Semantic Web. In *Proceedings of INTERSPEECH*, pages 841–844, Lisbon, Portugal, 2005.
10. D. Sonntag. *Ontologies and Adaptivity in Dialogue for Question Answering*. AKA and IOS Press, Heidelberg, 2010.
11. D. Sonntag, M. Deru, and S. Bergweiler. Design and Implementation of Combined Mobile and Touchscreen-Based Multimodal Web 3.0 Interfaces. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)*, pages 974–979, 2009.
12. D. Sonntag and M. Möller. Unifying semantic annotation and querying in biomedical image repositories. In *Proceedings of International Conference on Knowledge Management and Information Sharing (KMIS)*, 2009.
13. D. Sonntag, N. Reithinger, G. Herzog, and T. Becker. *Proceedings of IWSDS—Spoken Dialogue Systems for Ambient Environment*, chapter A Discourse and Dialogue Infrastructure for Industrial Dissemination, pages 132–143. Springer, LNAI, 2010.
14. D. Sonntag, G. Sonnenberg, R. Nesselrath, and G. Herzog. Supporting a rapid dialogue engineering process. In *Proceedings of the First International Workshop On Spoken Dialogue Systems Technology (IWSDS)*, 2009.
15. W. Wahlster. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In R. Krahl and D. Günther, editors, *Proceedings of the Human Computer Interaction Status Conference 2003*, pages 47–62, Berlin, Germany, 2003. DLR.
16. D. L. Weiss and C. Langlotz. Structured reporting: Patient care enhancement or productivity nightmare? *Radiology*, 249(3):739–747, 2008.