# Towards a Multimedia Knowledge-Based Agent with Social Competence and Human Interaction Capabilities

Leo Wanner*, Josep Blat, Stamatia Dasiopoulou, Mónica Domínguez, Gerard Llorach, Simon Mille, Federico Sukno
*ICREA / Pompeu Fabra U, Barcelona, Spain
leo.wanner@upf.edu

Eleni Kamateri, Ioannis Kompatsiaris, Stefanos Vrochidis
Centre for Research and Technology, Thessaloniki, Greece
ekamater@iti.gr

Elisabeth André, Florian Lingenfelser Gregor Mehlmann
U of Augsburg, Germany
andre@informatik.uni-augsburg.de

Andries Stam, Ludo Stellingwerff
Almende, Rotterdam, The Netherlands
andries@almende.org

Lori Lamel, Bianca Vieru
Vocapia Research, Paris, France
lamel@vocapia.com

Wolfgang Minker, Louisa Pragst, Stefan Ultes**
U of Ulm, Germany
**U of Cambridge
wolfgang.minker@uni-ulm.de

## ABSTRACT

We present work in progress on an intelligent embodied conversation agent in the basic care and healthcare domain. In contrast to most of the existing agents, the presented agent is aimed to have linguistic cultural, social and emotional competence needed to interact with elderly and migrants. It is composed of an ontology-based and reasoning-driven dialogue manager, multimodal communication analysis and generation modules and a search engine for the retrieval of multimedia background content from the web needed for conducting a conversation on a given topic.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*multimodal*; I.2.7 [**Artificial Intelligence**]: Embodied Agents—*dialogue, social and emotional competence*

## Keywords

embodied agent, dialogue, multimodal communication, retrieval

## 1. INTRODUCTION

The amount of research on intelligent agents is on the rise; see, e.g., [41, 7, 46], also because of the increasingly high demand for intelligent agents. Intelligent agents are omnipresent, be it in manufactoring, network surveillance, trade, home assistance, or social interaction. Of particular relevance are social companion agents. An increasingly high number of elderly need help to carry out their daily life routines and lack social warmth and sympathy [47]; an increasingly high number of people abandon their ancestral cultural and social environments to move to countries with different language and culture and need support to get around and integrate; and so on. This calls for culturally sensitive versatile multilingual conversation agents. However, most of the current proposals in the field do not rise up to the challenge. Thus, they usually follow a predefined dialogue strategy (which cannot be assumed when interacting with, e.g., elderly); they do not take into account cultural idiosyncrasies of the addressee when planning their actions; they are not multilingual to be able to intermediate between a migrant and a native from the host country; etc. In our work, we set out to address this challenge. We develop an embodied multilingual conversation agent with social and cultural skills that serves for migrants with language and cultural barriers in the host country as a trusted information provision party and mediator in questions related to basic care and healthcare. The agent is targeted to possess the following characteristics: **(i)** be able to retrieve multimedia background content from the web in order to show itself informed and knowledgeable about themes relevant to the user (only then will the user trust the agent);[1] **(ii)** understand and interpret the concerns of the user expressed by a combination of facial, gestural and multilingual verbal signals, embedded into a specific cultural, social and emotional context; **(iii)** plan the dialogue using ontology-based reasoning techniques in order to be flexible enough and react


*MARMI, in conjunction with ICMR, 2016* New York City, New York, USA


[1]For instance, Pfeifer et al. [33] show that the capacity of the agent to play games known by the user, or, in other words, to share skills and knowledge with the user, helps establish a trustful attitude of the user towards the agent.

appropriately to unexpected turns of the user; **(iv)** communicate with the user using verbal and non-verbal (facial and gestural) signals in accordance with the given linguistic, cultural, social and emotional context; **(v)** communicate with the user using verbal and non-verbal (facial and gestural) signals in accordance with the given linguistic, cultural, social and emotional context.

In what follows, we present the global design of our agent (henceforth referred to as KRISTINA) and sketch the different modules of which it is composed.

# 2. GLOBAL DESIGN OF THE AGENT

Figure 1 shows the global design of the KRISTINA agent. The agent is composed of a collection of modules that ensure informed multimodal expressive conversation with a human user. The communication analysis modules are controlled by the state-of-the-art *Social Signal Interpretation* (SSI) framework [48], which furthermore carries out event-driven fusion of the data from different modi. The dialogue-oriented modules are embedded in the *Visual Scene Maker* (VSM) framework [14]. A dedicated search engine acquires background multimodal information from the web and relevant curated information sources. The integrated representation of acquired and user-transmitted multimodal information is stored in the Knowledge Base (KB) in terms of ontologies, which facilitates the realization of flexible reasoning-based dialogue strategies.

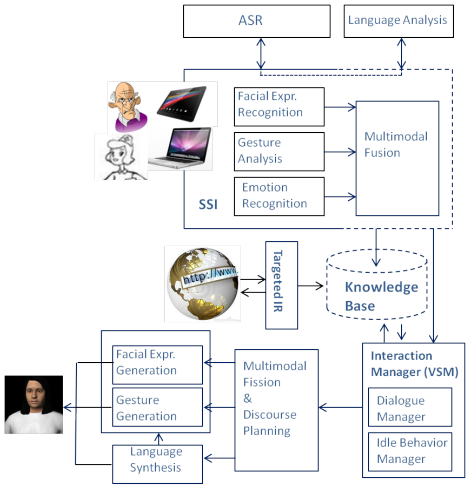


**Figure 1: Architecture of the KRISTINA agent**

A dedicated fission and discourse planning module shall ensure an adequate assignment of the content elements that are chosen for communication to the user to the individual modi and their coherent and coordinated presentation. The language generation module feeds its intermediate and final outcome also to the facial expression and gesture generation modules in order to ensure, e.g., accurate lip synchronization and beat gestures of the virtual character.

# 3. THE AGENT'S KNOWLEDGE

To ensure that the agent is "knowledgeable" about the topic of the conversation in order to be able to interpret the multimodal input (question, comment, request, etc.) of the user and come up with the appropriate reaction, the knowledge representation in the agent must be theoretically sound and scalable. Furthermore, the knowledge repositories must be dynamically extendable, i.e., the agent must be able "to learn" from both the input of the user and the external world.

## 3.1 Multimodal Knowledge Representation

Our multimodal knowledge representation framework in the Knowledge Base (KB) includes a number of ontologies designed to support the dialogue with the user and to represent the relevant basic care and healthcare background information from the web. The ontologies cover: (i) models for the representation, integration and interpretation of verbal and non-verbal aspects of user communication [34]; (ii) domain models that capture the various types of background knowledge, including user profile ontologies [17]; ontologies for modeling routines, habits and behavioural aspects [29], and ontologies for healthcare and medical information [36].

To provide knowledge-driven interpretation schemata for the determination of the appropriate reaction to the input of the user, the framework combines, integrates and interprets the user's input in the context of background knowledge, coupling the interpretation with intelligent and context-aware decision making and reasoning mechanisms. The realization of the interpretation framework capitalizes on the association of contextual concepts with inference rules that, based on the recognized context of the conversation, perform additional inferences and either update the KB or generate a response to be sent back to Dialogue Management. So far, rules relevant to biographical information and user behaviour are realized. Conversations are supported in which the user informs or asks for biographical information, as, e.g., the age, or basic daily routines (e.g., daily water intake or hygiene).

## 3.2 Multimedia retrieval from the web

This module extracts content from web resources (including social media) to enhance the background knowledge of KRISTINA. The available types of web resources that contain basic care- and healthcare-related information vary significantly. They include web pages with information or instructions as well as blogs and online discussion areas such as forums and social networking sites, where interested parties can discuss and share opinions and experiences. In addition, web resources may encode their information in multiple languages and in various modalities such as text, images, videos, audio recordings, and slideshows. Therefore, versatile mechanisms are needed to ensure high quality information search and extraction.

The module includes the following submodules: (i) crawling and scraping to extract relevant content from web resources, for which we use Apache Nutch[2] and boilerpipe[3], and social media; (ii) content extraction that distills content elements from web resources and social media (using MetaMap,[4] Babelfy,[5] and Unitex[6]); (iii) social media topic detection that detects health-related topics based on clustering; (iv) indexing, search, and retrieval of multimodal information from web [42]; (v) domain-specific query formulation

[2]http://nutch.apache.org/
[3]https://boilerpipe-web.appspot.com/
[4]https://metamap.nlm.nih.gov/
[5]http://babelfy.org/
[6]http://www-igm.univ-mlv.fr/ unitex/

for querying the indexed repository and triggering the ontology population mechanism; and (vi) user interaction query formulation that copes with the derivation of knowledge-driven interpretations from users' input in order to compile the appropriate reactions.

# 4. MULTIMODAL INTERACTION

The multimodal interaction of the KRISTINA agent involves dialogue management and multimodal communication analysis and generation.

## 4.1 Knowledge-based Dialogue Management

Any flexible informed multimodal interaction requires a reasoning-driven dialogue management (DM) module—in particular, if it takes into account the social and cultural idiosyncrasies of the user.

### *4.1.1 Culture-Aware and Emotion-Sensitive Dialogue*

DM in a socially and culturally aware intelligent conversation agent differs from generic DM in that, on the one hand, the agent must be able to react to social and cultural input and generate culturally appropriate emotional output, and, on the other hand, the domain of the conversation is huge [35].

In most of the state-of-the-art DM models, maintenance of the dialogue state and selection of the next system action are the two main tasks of the dialogue manager [44]. To be able to cope with the large domain of conversation, we differentiate between two parts of the dialogue state: (i) a part that models the current state of the conversation with respect to the domain information (the *domain state*), which is transferred to the KB; (ii) a part that models other interaction phenomena or the user state, which is maintained by the dialogue manager. Consequently, new user input is not only used by the dialogue manager, but also processed by the KB to produce an updated domain state. Based on this state, the dialogue manager makes its decision on how to carry on the conversation.

In order to make the DM socially and culturally aware, we need to also take into account the emotional state of the user, combined with information about their cultural background. Thus, KRISTINA must be aware of cultural interaction idiosyncrasies and produce appropriate and emotional reactions. To handle emotions, emotion states must be related to plain system actions. That is, the DM is not only responsible for *what* the agent says, but also *how* it says it. For the realization of such a flexible DM, we use as basis OwlSpeak [44], which has already been used to integrate the user state [43].

### *4.1.2 Regulating Turn-Taking and Idle Behavior*

Besides the maintainance of the dialogue state and selection of the next system action, the control of the agent's turn-taking behavior and the control of a variety of non-verbal idle behavior patterns are essential aspects to be dealt with during the dialogue [24]. To manage these two tasks, we use the *Visual SceneMaker* (VSM) [14, 23], which has been designed to control the interactive behavior of virtual characters. VSM determines the agent's participant role changes during the dialogue, based on the observed user input and the agent's own contributions planned by the DM. The turn-taking decisions are made on the basis of a policy which determines whether the agent is allowed to interrupt the user's utterance and how it reacts to the user's attempts to barge in its own turn. VSM is also responsible for planning appropriate and vivid non-verbal behavior patterns while the agent is listening to the user or whenever the speaker and listener roles are not yet clearly negotiated. In this latter case, the agent fulfills the role of a bystander by displaying an idle behavior that is supposed to create an impression of engagement and attentiveness while waiting for the user's next dialogue move or before actively starting a contribution itself, for example, mimicking the user's affective state by mirroring their facial expressions, gestures or body postures or displaying different eye gazes [25].

## 4.2 Multimodal communication analysis

The objective of the multimodal communication analysis is to convert the verbal and non-verbal input of the user into abstract representations that are projected onto ontologies.

### *4.2.1 Analysis of verbal communication*

The analysis of verbal communication is viewed as a cascade of three stages: (i) speech recognition, (ii) prosody recognition, and (iii) language analysis.

**Speech recognition.** For speech recognition, we use Vocapia's ASR[7], which use statistical speech models [20] for both acoustic and language modeling. Unfortunately, so far no large speech corpora in the basic and healthcare domains are available to train the models. Some studies cover selected aspects of the medical domain[8], others have very limited amounts of data (less than 1000 sentences from 27 speakers for Japanese, English, and Chinese [38, 30]). But such corpora are needed for optimal performance of ASR since it is well known that spoken language differs from written language (such that written language corpora cannot be used for the task). Written language (usually) has a correct grammatical structure. In contrast, when we speak, we frequently make hesitations, false-starts, repairs or ungrammatical constructions. Both disfluencies and ungrammatical structures are known to influence the performance of the speech recognition technologies [1], as do the speaker's accent, acoustic conditions, speaking style, topic, interactivity, limited system vocabulary, etc. Due to the lack of speech corpora in our domain, we trained our language models so far on large quantities of written data of which only a small portion corresponds to transcribed speech. Measures are taken to increase the share of spoken data, and the experiments show that already a minor increment of the share of spoken data let the performance improve by 10–30% for German, Spanish and Turkish,compared to the baseline that uses written data only.

**Speech prosody.** Accurate detection of prosody in the speech of the user is essential for an agent with social and emotional competence. For this purpose, we investigate the correlation between prosodic elements and the Information Structure [9], the prediction capability achieved when combining word level acoustic features and several linguistic features [10], and the analysis of the most consistent prosodic cues in longer speech passages [13].

**Language analysis.** The language analysis consists in itself of four substages: (i) surface-syntactic parsing, (ii)

[7]http://www.vocapia.com/

[8]E.g., the MedSLTproject (http://www.let.rug.nl/ tiedeman/) covers only headache, chest pain and abdominal pain questions.

deep-syntactic parsing, (iii) frame-semantics parsing, and (iv) projection to ontological representations. For surface-syntactic parsing, we adapt Ballesteros et al.'s [6] LSTM parser to spoken material, combining written language treebanks [27] with targeted dialogue recordings. Currently, the parser shows an improvement of over 4 points of the Labeled Attachment Score (from 69,57% to 73,86%) on spoken material. In order to abstract from syntactic idiosyncracies and get closer to a semantics-oriented structure, we map surface-syntactic structures onto deep-syntactic structures (DSyntSs) in the sense of [26]; cf. [4] for more details. To obtain syntax-agnostic representations and to generalize the meaning to a certain extent, DSyntSs are mapped onto FrameNet-based structures (FNSs).[9] To complete the analysis pipeline, the FNSs are projected to an ontological representation. However, FrameNet has not been created with ontological considerations in mind such that this projection is not straightforward. Our preliminary methodology projects FNSs to a DOLCE+DnS UltraLite (DUL) compliant representation, where Frames and Frame Elements are mapped to respective DUL constructs based on their type.

#### 4.2.2 Analysis of facial and gestural communication

Facial expressions and gestures communicate semantic and/or affective information. Until now, we focused on the analysis of the affective states of users, which is also central in state-of-the-art research [31, 45].

Through the application of sensor technologies, signal processing and recognition techniques, multi-modal cues that point to certain affective states can be measured and recognized. The cues can be used for automatic classification of human emotions communicated via different modi [50]. The face is one of these modi. Gestures constitute another, even if less prominent, and audio another one. So far, we addressed mainly affective face analysis. Traditionally, affective face analysis was done in terms of the recognition of diverse sets of (static) facial expressions [11, 39]. However, nowadays there is a consensus about the need for a dynamic analysis of expressions and the use of action units (AUs) from the *Facial Action Coding System* [12] as a standardized representation [40]. Currently, we address the estimation of facial AUs in a fully automatic manner by firstly extracting SIFT-based features from sets of automatically detected facial landmarks and then applying a set of independent linear classifiers to associate a probability to each of the targeted AUs. These classifiers are trained following [37], which allows training AU classifiers using datasets with a reduced amount of ground truth (only prototypical facial expressions are needed, which is more efficient than annotation of AUs).

Meaningful cues extracted from available affective modi are combined through fusion strategies in order to generate a final prediction. Present studies have shown varying degrees of success or even failure of multimodal fusion [50, 21]. KRISTINA aims at recognizing natural and spontaneous affective states in realtime, which requires sophisticated fusion schemes. Our work on fusion draws on Lingenfelser's [22] "event-driven" fusion, whose algorithm is based on [15]. The algorithm does not force decisions throughout considered modalities for every time frame, but instead asynchronously fuses time-sensitive events from any given number of modi. This has the advantage of incorporating temporal alignments between modi and being very flexible with respect to the type and mode of used events. In [22], this algorithm was used to combine the recognition of short-timed laugh (audio) and smile (video) events for a continuous assessment of a user's level of positive valence. We are extending it to cover the whole valence arousal space, spanned by positive and negative valence and arousal axes.

[9]At this stage, it is still unclear whether FNS, PropBank or VerbNet structures are best suited as intermediate representations between DSyntSs and ontologies.

### 4.3 Multimodal communication generation

#### 4.3.1 Verbal communication generation

Verbal communication in our agent starts from ontological representations, following the inverse cascade of processing stages presented in Section 4.2.1: (i) projection of ontological representations to FNSs; (ii) generation of DSyntSs from FNSs, (iii) generation of linearized and inflected SSyntSs; (iv) speech synthesis.

The projection from FN to DSyntS involves the introduction of the lexical units of the desired output language, and the establishment of the syntactic structure of the sentence. For this task, we use multilingual rule-based graph-transduction grammars and dictionaries, as proposed in [49]. To map DSyntSs onto surface, we use a state-of-the-art statistical graph transduction model [5], which is further adapted to the idiosyncrasies of spoken language.

In parallel to the cascaded proposition realization model, a hierarchical prosodic model is deployed, which captures prosodic events as a complex interaction of acoustic features occuring at different phonological levels in the utterance (i.e., prosodic phrases, prosodic words and syllables) [9]. Such a prosody module makes use of linguistic features from the previous text generation stages (departing from the content level) to predict the location of prosodic labels in terms of prominence and phrasing events which are relevant for communication purposes. Finally, these prosodic events are realized as a combination of acoustic features, based upon training using domain specific voice samples.

#### 4.3.2 Non-verbal communication generation

The conversational agent is realized as an embodied conversational agent (ECA), which enables us to use verbal and non-verbal modi for communication. The embodiment of the agent is realized through a credible virtual character. The research on social embodiment still faces many challenges [3]. Just ensuring credibility (as opposed to realism) implies the believability of the rendering of the agent, avoidance of the trap of the uncanny valley [28], and animation through facial expressions and gestures, when appropriate. Gestures and facial expressions must be generated according to the semantics of the message that is to be communicated. Since the generation of facial expressions using tags (smile, surprise, etc.) would limit the possible facial expressions and require a manual design of all possible expressions for each character, we use the increasingly popular valence-arousal representation of emotions [16] for this purpose; cf., also [18, 19]. Our model can generate and animate facial expressions in the continuous 2D valence-arousal space by linearly interpolating only five extreme facial poses. Because of its parametric nature, the valence-arousal space can be easily applied to a variety of faces. Using the semantics and other features, gestures are generated, keeping in mind the cul-

tural context of the conversation.

In addition, we deal with web-based 3D rendering of ECAs and speech–lip synchronization. Interactive 3D graphics on the web has advanced considerably in recent years, to the extent that fully featured 3D scene editors [2] permit the creation of advanced interfaces capable of supporting ECAs, although still some challenges with respect to the limitations of bandwith for data transfer must be resolved. The work on speech–lip synchronization is based on [8] and [32]. To increase its efficiency and quality, a reduced set of features is used to compute the model. The model is based on the use of vowels as key points, and consonants and silences as transitions between them. A neural network trained model generates transitions, where prosodic features and emotional state are considered to define the duration, the peak position, the attack and decay time and slope of the transitions.

# 5. CONCLUSIONS

We presented work in progress on KRISTINA, an embodied intelligent agent. Its communication modules are designed to facilitate the conduct of socially competent emotive multilingual conversations with individuals in need of advice and support in the context of basic care and healthcare. The technologies that we develop will be validated in prolonged trials of each prototype that marks the termination of a software development cycle, with a representative number of migrants recruited as users from the migration circles in two European countries: elderly Turkish migrants and their relatives and short term Polish care giving personnel in Germany and North African migrants in Spain.

# 6. ACKNOWLEDGMENTS

The presented work is funded by the European Commission under the contract number H2020–645012–RIA.

# 7. REFERENCES

[1] M. L. Adda-Decker. *Pronunciation variants across systems, languages and speaking style*. Modeling Pronunciation Variation for Automatic Speech Recognition, Netherlands, 1998.

[2] J. Agenjo, A. Evans, and J. Blat. Webglstudio: A pipeline for webgl scene creation. In *Proceedings of the 18th International Conference on 3D Web Technology*, pages 79–82, New York, NY, USA, 2013. ACM.

[3] E. André. Challenges for social embodiment. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pages 35–37. ACM, 2014.

[4] M. Ballesteros, B. Bohnet, S. Mille, and L. Wanner. Data-driven deep-syntactic dependency parsing. *Natural Language Engineering*, pages 1–36, 2015.

[5] M. Ballesteros, B. Bohnet, S. Mille, and L. Wanner. Data-driven sentence generation with non-isomorphic trees. In *Proceedings of the 2015 Conference of the NAACL: Human Language Technologies*, pages 387–397, Denver, Colorado, May–June 2015. ACL.

[6] M. Ballesteros, C. Dyer, and N. A. Smith. Improved transition-based parsing by modeling characters instead of words with LSTMs. *CoRR*, 2015.

[7] T. Bickmore, D. Schulman, and C. Sidner. Automated interventions for multiple health behaviors using conversational agents. *Journal of Patient Education and Counseling*, 92(2):142–148, 2013.

[8] M. Cohen and D. Massaro. Modeling Coarticulation in Synthetic Visual Speech, 1993.

[9] M. Domínguez, M. Farrús, A. Burga, and L. Wanner. Using hierarchical information structure for prosody prediction in content-to-speech application. In *Proceedings of the 8thInternational Conference on Speech Prosody (SP 2016)*, Boston, MA, 2016.

[10] M. Domínguez, M. Farrús, and L. Wanner. Combining acoustic and linguistic features in phrase-oriented prosody prediction. In *Proceedings of the 8thInternational Conference on Speech Prosody (SP 2016)*, Boston, MA, 2016.

[11] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[12] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[13] M. Farrús, G. Lai, and J. Moore. Paragraph-based prosodic cues for speech synthesis applications. In *Proceedings of the 8thInternational Conference on Speech Prosody (SP 2016)*, Boston, MA, 2016.

[14] P. Gebhard, G. U. Mehlmann, and M. Kipp. Visual SceneMaker: A Tool for Authoring Interactive Virtual Characters. *Journal of Multimodal User Interfaces: Interacting with Embodied Conversational Agents, Springer-Verlag*, 6(1-2):3–11, 2012.

[15] S. W. Gilroy, M. Cavazza, M. Niranen, E. André, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, and M. Billinghurst. Pad-based multimodal affective fusion. In *Affective Computing and Intelligent Interaction and Workshops*, 2009.

[16] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.

[17] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, and M. von Wilamowitz-Moellendorff. Gumo–the general user model ontology. In *User modeling 2005*. Springer, Berlin / Heidelberg, 2005.

[18] J. Hyde, E. J. Carter, S. Kiesler, and J. K. Hodgins. Assessing naturalness and emotional intensity: a perceptual study of animated facial motion. In *Proceedings of the ACM Symposium on Applied Perception*, pages 15–22. ACM, 2014.

[19] J. Hyde, E. J. Carter, S. Kiesler, and J. K. Hodgins. Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1719–1728. ACM, 2015.

[20] L. G. Lamel. Speech recognition. In *R*, pages 305–322. 2003.

[21] F. Lingenfelser, J. Wagner, and E. André. A systematic discussion of fusion techniques for multi-modal affect recognition tasks. In *ICMI*, pages 19–26, 2011.

[22] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran. An event driven fusion approach for enjoyment recognition in real-time. In *MM*, pages 377–386, 2014.

[23] G. Mehlmann and E. André. Modeling Multimodal Integration with Event Logic Charts. In *Proceedings of the 14th International Conference on Multimodal Interaction*, pages 125–132. ACM, New York, NY, USA, 2012.

[24] G. Mehlmann, K. Janowski, and E. André. Modeling Grounding for Interactive Social Companions. *Journal of Artificial Intelligence: Social Companion Technologies, Springer-Verlag*, 30(1):45–52, 2016.

[25] G. Mehlmann, K. Janowski, T. Baur, M. Häring, E. André, and P. Gebhard. Exploring a Model of Gaze for Grounding in HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 247–254. ACM, New York, NY, USA, 2014.

[26] I. Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany, 1988.

[27] S. Mille, A. Burga, and L. Wanner. AnCora-UPF: A multi-level annotation of Spanish. In *Proceedings of DepLing 2013*, pages 217–226, Prague, Czech Republic, 2013.

[28] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE*, 19(2):98–100, 2012.

[29] B. Motik, B. Cuenca Grau, and U. Sattler. Structured objects in owl: Representation and reasoning. In *Proceedings of the 17th international conference on World Wide Web*, pages 555–564. ACM, 2008.

[30] G. S. Neubig. *Towards High-Reliability Speech Translation in the Medical Domain*. CNLP, 2013.

[31] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. *Human Computing and Machine Understanding of Human Behavior: A Survey*, volume 4451, pages 47–71. 2007.

[32] S. Pasquariello and C. Pelachaud. Greta: A simple facial animation engine. *Soft Computing and Industry - Recent Applications*, pages 511–525, 2002.

[33] L. Pfeifer Vardoulakis, L. Ring, B. Barry, C. Sidner, and T. Bickmore. Designing relational agents as long term social companions for older adults. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, 2012.

[34] J. Posner, J. Russell, and B. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development and psychopathology. *Development and psychopathology*, 17(3), 2005.

[35] L. Pragst, S. Ultes, M. Kraus, and W. Minker. Adaptive dialogue management in the kristina project for multicultural health care applications. In *Proceedings of the 19thWorkshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, pages 202–203, Aug. 2015.

[36] D. Riaño, F. Real, F. Campana, S. Ercolani, and R. Annicchiarico. An ontology for the care of the elder at home. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine: Artificial Intelligence in Medicine*, AIME '09, pages 235–239, Berlin, Heidelberg, 2009. Springer-Verlag.

[37] A. Ruiz, J. Van de Weijer, and X. Binefa. From emotions to action units with hidden and semi-hidden-task learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3703–3711, 2015.

[38] S. K. Sakti. *Towards Multilingual Conversations in the Medical Domain: Development of Multilingual Medical Data and a Network-based ASR System*. LREC. Iceland, 2014.

[39] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[40] A. Savran, B. Sankur, and M. T. Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774–784, 2012.

[41] W. Shen, Q. Hao, H. Yoon, and D. Norrie. Applications of agent-based systems in intelligent manufactoring: An updated review. *Advanced Engineering Informatics*, 20(4):415–431, 2013.

[42] T. Tsikrika, K. Andreadou, A. Moumtzidou, E. Schinas, S. Papadopoulos, S. Vrochidis, and Y. Kompatsiaris. A unified model for socially interconnected multimedia-enriched objects. In *Proceedings of the 21st MultiMedia Modelling Conference (MMM2015),*, 2015.

[43] S. Ultes, M. Kraus, A. Schmitt, and W. Minker. Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 374–383. ACL, Sept. 2015.

[44] S. Ultes and W. Minker. Managing adaptive spoken dialogue for intelligent environments. *Journal of Ambient Intelligence and Smart Environments*, 6(5):523–539, Aug. 2014.

[45] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'ericco, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3:69–87, 2012.

[46] O. Vinyals and Q. Le. A neural conversation model. In *Proceedings of the 31st International Conference on Machine Learning*, 2015.

[47] A. Vlachantoni, R. Shaw, R. Willis, M. Evandrou, J. Falkingham, and R. Luff. Measuring unmet need for social care amongst older people. *Population Trends*, (145):1–17, 2011.

[48] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The social signal interpretation (SSI) framework–multimodal signal processing and recognition in real-time. In *Proceedings of ACM International Conference on Multimedia*, 2013.

[49] L. Wanner, B. Bohnet, N. Bouayad-Agha, F. Lareau, and D. Nicklaß. MARQUIS: Generation of user-tailored multilingual air quality bulletins. *Applied Artificial Intelligence*, 24(10):914–952, 2010.

[50] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31:39–58, 2009.