Towards Continuous Dialogue Corpus Creation: writing to corpus and generating from it

Andrei Malchanau[1], **Volha Petukhova**[1] **and Harry Bunt**[2]
[1]Spoken Language Systems Group, Saarland University, Germany
[2]Tilburg Center for Communication and Cognition, Tilburg University, The Netherlands
{andrei.malchanau,v.petukhova}@lsv.uni-saarland.de; harry.bunt@uvt.nl

Abstract

This paper describes a method to create dialogue corpora annotated with interoperable semantic information. The corpus development is performed following the ISO linguistic annotation framework and primary data encoding initiatives. The Continuous Dialogue Corpus Creation (D3C) methodology is proposed, where a corpus is used as a shared repository for analysis and modelling of interactive dialogue behaviour, and for implementation, integration and evaluation of dialogue system components. All these activities are supported by the use of ISO standard data models including annotation schemes, encoding formats, tools, and architectures. Standards also facilitate practical work in dialogue system implementation, deployment, evaluation and re-training, and enabling automatic generation of adequate system behaviour from the data. The proposed methodology is applied to the data-driven design of two multimodal interactive applications - the Virtual Negotiation Coach, used for the training of metacognitive skills in a multi-issue bargaining setting, and the Virtual Debate Coach, used for the training of debate skills in political contexts.

Keywords: dialogue resources, interoperable semantic annotations, international standards, dialogue system design

1. Introduction

A steadily growing interest can be observed in data-driven modelling of phenomena related to natural language, vision, behavioural and organizational processes. Data have become essential to advance the state of the art in many areas including the development of spoken (multimodal) dialogue systems. Conversational applications such as Apple's Siri, Microsoft's Cortana and Google Now became successful and robust partly due to the amount of real user data available to their developers. The most recent trend in dialogue system design involves end-to-end dialogue systems using neural network models trained on previously collected dialogue data, without any detailed specification of dialogue states (Wen et al., 2017; Bayer et al., 2017). This requires large amounts of data to cover a reasonable number of possible dialogue states and participant actions.

Dialogue data have often been collected in Wizard-of-Oz experiments (Dahlbäck et al., 1993), where the dialogue system is replaced by a human Wizard who simulates the system's behaviour according to a pre-defined script.

An alternative is to use simulated users. With good user modelling, a dialogue system could be rapidly prototyped and evaluated. Simulated data sets are, however, rather scarce (Schatzmann et al., 2006).

Resources for data-driven learning of task-oriented systems are also collected with existing systems (Bennett and Rudnicky, 2002; Henderson et al., 2014). For example, the DialPort project addresses the need for dialogue resources by offering a portal connected to different existing dialogue systems (Lee et al., 2017).

Learning algorithms have also been proposed to train a dialogue system online. System behaviour is initially learned from a minimal number of dialogues and is then optimized as more data arrives (Daubigney et al., 2012). As a data collection strategy this approach may not be really successful, since the initial system performance can be rather poor.

Building an annotated dialogue corpus is an expensive activity, especially when it requires manual annotation. Over the years, many annotated dialogue corpora have been created, however annotations and their formats differ from resource to resource. The community has recognized this problem by addressing the interoperability of dialogue resources. ISO 24617-2 "Semantic annotation framework, Part 2: Dialogue acts" (ISO, 2012), in particular aims to contribute to the interoperability of annotated dialogue corpora. New corpora have been created (Petukhova et al., 2014a), existing corpora re-annotated (Bunt et al., 2013) using the standard annotation scheme, and existing annotations mapped to ISO 24617-2 (Petukhova et al., 2014b). The DialogBank is a new language resource that contains dialogues of various kind with gold standard annotations according to the ISO 24617-2 standard (Bunt et al., 2016).

This paper explores yet another way to create semantically annotated dialogue corpora: base corpus developments on the framework of ISO linguistic (i.e. semantic) annotation standards[1]. The approach follows a *continuous corpus creation* methodology where the corpus is used as a shared repository for analysis and modelling of interactive dialogue behaviour, and for implementation and evaluation of the dialogue system. Standard data models (i.e. annotation schemes, encoding and annotation formats) support the corpus development facilitating the creation of semantically rich and interoperable dialogue data for multiple domains, contributing to cost reduction in corpus creation. Standards also support practical work in dialogue system design, evaluation and re-training, and enables automatic generation of adequate system behaviour from the data.

The paper is structured as follows. Section 2 presents the overall methodology, discussing the main principles and key processes related to corpus development. Sec-

[1]We refer to (Ide and Pustejovsky, 2017) for an overview of existing standards.

tion 3 presents the ISO 26417-2 data model introducing the basic concepts and the Dialogue Act Markup Language (DiAML) as the main corpus annotation and exchange format between system components. The proposed approach is illustrated in Section 4, by applying it to recently performed corpus creation activities when designing two different applications - Virtual Debate and Negotiation Coaches (Petukhova et al., 2017b; Petukhova et al., 2017a). The paper is concluded by a summary of the main findings and a discussion of directions for future research.

2. Corpus Creation Methodology

An important step in designing any multimodal dialogue system is to model natural human dialogue behaviour, as a basis for developing dialogue system components. Each module in a dialogue system performs a task such as dialogue act classification, event identification, co-reference resolution, or semantic role labelling, and is integrated according to the adopted architectural approach (e.g. pipeline, multi-agent or multi-threaded), which determines how the modules communicate and exchange their processing results. In such a data-inspired design approach, success will heavily depend on the *quality*, *costs* and *application range* of the underlying corpus data. These three aspects are influenced by multiple variables such as number, tasks and roles of dialogue participants involved in an interactive situation (real vs simulated humans vs artificial agents); dialogue setting, modalities and media available; granularity and nature of annotations and analysis (manual vs automatic vs no annotations); infrastructures, platforms, tools and formats accessible. All these variables impact the corpus creation design, the complexity of the set-up, and the processing steps.

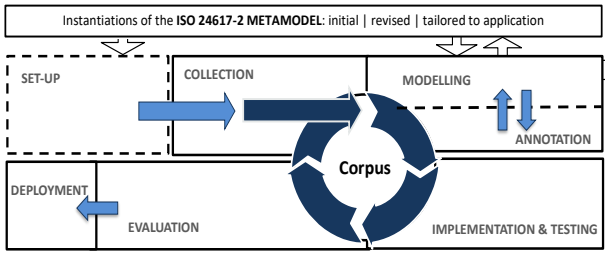

Figure 1: Continuous dialogue corpus creation (D3C).

We propose a continuous dialogue corpus creation (D3C) methodology consisting of the following steps (Fig. 1):

1. Set-up: based on the ISO 24617-2 metamodel define an interaction scenario and specify data collection requirements; provide details for participants roles and tasks, recording setting (equipment and environment) and description of data collection process;
2. Collect: record, encode and store human-human dialogue primary [2] data for the specified scenario;
3. Model: revise the standard data model with attributes derived from annotated data: apply the standard ISO 24617-2 metamodel, include other SemAF concepts on demand and tailor to the application domain;
4. Annotate: apply standard and domain-specific annotation scheme(-s) to classify a particular set of entities and their properties;
5. Implement & Test: build (train) and test dialogue system components based on the underlying annotations performed and resulting dialogue models; optional tests are possible experimenting with tuned and/or modified parameters;
6. Evaluate: perform objective (system performance) and user-based (user perception) evaluation with the integrated dialogue system prototype in the laboratory and close to operational environments; log evaluation sessions and analyse results;
7. Deploy (*optional after each iteration*): write to the corpus, document and prepare to be released including signals, primary data, annotations and corpus manual with schemes, guidelines and format specifications;
8. Repeat steps 1-7 for the full cycle for a refined set-up, or steps 3-6 to re-train system modules based on data obtained in user-based evaluation sessions.

The proposed methodology is in the line with principles of semantic annotation defined in the ISO standard 24617-6 which characterizes the ISO semantic annotation framework (ISO, 2016). The standard includes the CASCADES (**C**onceptual analysis, **A**bstract syntax, **S**emantics, and **C**oncrete syntax for **A**nnotation language **DES**ign) annotation schemes design model (Bunt, 2015). The model enables a systematic (re-)design process: from conceptual ('metamodel') and semantic choices ('abstract' syntax) to more superficial decisions such as the choice of particular XML attributes and values ('concrete' syntax). The method can be used to design of a new annotation scheme or provides support for improving an existing annotation scheme through feedback loops. The CASCADES is integrated with the MATTER method (Pustejovsky et al., 2017) for annotation and data modelling, conceptualized as the Model, Annotate, Train, Test, Evaluate and Revise cycle which inspired the presented methodology.

3. ISO 24617-2 Data Model

Well-established data models are the key enablers for corpus and system development. They are a prerequisite for the corpus to be of good quality, provides ways to systematically incorporate extensions, and ensures interoperability, enabling sharing, merging and comparison with other resources. Data models, formalized descriptions of data objects and relations between them, are designed to capture the structure and relations in diverse types of data and annotations. Well-specified standard resource formats and processes facilitate the exchange of information between dialogue system modules. Mappings between primary data and the data model are operationalized via schema-based data-binding processes (Ide and Romary, 2004).

[2] Data observed or collected directly from first-hand experience such as representation of written (e.g. text), spoken (e.g. orthographic transcriptions of audio) and multimodal (e.g. images or videos) behaviour. Typically, primary data objects are represented by "locations" in an electronic file, e.g. the span of characters comprising a sentence or word, or a point at which a given temporal event begins or ends. More complex data objects may consist of a list or set of contiguous or non-contiguous locations in primary data, see (Ide and Romary, 2004)

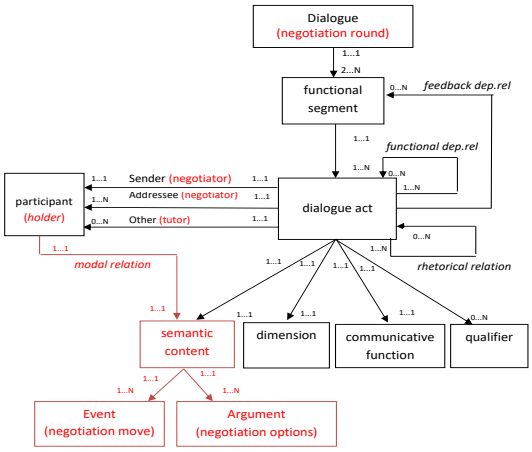

Figure 2: ISO 24617-2 metamodel for dialogue act annotation. Domain-specific extensions marked red, see Sections 4.1 and 4.3.

3.1. Basic concepts

The ISO 24617-2 data model (or 'metamodel', see Fig. 2) represents the fundamental upper-level concepts that are involved in dialogue act annotation. A dialogue consists of two or more functional segments. Each segment is related to one or more dialogue acts, reflecting the possible multifunctionality of functional segments. Each dialogue act has exactly one sender, one or more addressees, and possibly other participants. It has a semantic content of a certain type ('dimension'), and a communicative function, which may have any number of qualifiers. Dialogue acts are possibly related to other dialogue acts through functional dependence and rhetorical relations, and to dialogue segments through feedback dependence relations.

3.2. ISO Dialogue Act Markup Language

The Dialogue MarkUp Language (DiAML) (ISO, 2012) is used as the representation and exchange format in dialogue corpus and system development; DiAML is also used for communication among all system modules, and for representing intermediate and end results.

The representation of annotations in ISO DiAML makes use of the XML element `<dialogueAct>`, which has the following attributes: `@target`, whose value is a functional segment: `@sender`, `@addressee`, `@otherParticipant`; `@dimension`, `@communicativeFunction`; `@functionalDependence` and `@feedbackDependence`, and the three attributes `@certainty`, `@conditionality`, and `@sentiment`, with qualifiers as values. Additionally, rhetorical relations among dialogue acts are represented by means of `<rhetoLink>` elements. DiAML annotations can be extended with a semantic content, also shown in (Bunt et al., 2017), by introducing a `<semanticContent>` element. Consider the following ISO DiAML representation as an example:

```
<dialogueAct xml:id="dap1" sender="#p1"
         addressee="#p2" dimension="task"
         communicativeFunction="inform"
         target="#fsp1">
  <SemanticContent>
    <event xml:id="e1"  type="offer"/>
      <Arg>10_percent</Arg>
    <modalLink holder="#p1" target="#e1"
     modalRel="preference"/>
  </SemanticContent>
</dialogueAct>
```

The `<event>` element, which specifies information about the semantic content of a dialogue act, could be the same as the element with the same name that is used in the ISO annotation schemes for time and events (ISO 24617-1), for semantic roles (ISO 24617-4), and for spatial information (ISO 24617-7), and that has also been proposed for the annotation of modality (Lapina and Petukhova, 2017) and quantification (Bunt, 2017). This opens the possibility to specify quite detailed information about the semantic content of dialogue acts, including domain-specific semantics as shown in (Petukhova et al., 2017a) for negotiations.

4. Use cases

We illustrate the proposed approach by discussing corpus and system development architectures for two applications - the Virtual Debate Coach (VDC, (Petukhova et al., 2017b)) and the Virtual Negotiation Coach (VNC, (Petukhova et al., 2017a)).

4.1. Set-up

The design of any system requires a clear understanding of the users, their goals and the usage situation. This helps to determine the system's functionality, reduces design mistakes and often provides good inspiration and orients. The data collection set-up includes first of all the specification of the intended users and system requirements. A users analysis is conducted to define key user groups (age, gender, cultural and educational backgrounds, etc.) and identify their interest areas, known attitudes, values and priorities. Context of use, settings and users' needs have a direct impact on the role the system will play in an interactive situation, and subsequently on the system functionality. Apart from the communicative tasks that a dialogue system has, namely to understand and adequately react to users' dialogue contributions, a dialogue system has tasks dependent on the application domain in relation to the role(-s) it plays, e.g. as an assistant, adviser or mediator, as a passive observer, as a tutor or as a coach. Users, context and system requirements are used not only to make important design decisions but also to define appropriate verification and evaluation strategies. The evaluation tasks should be representative for most users such that results can be generalized beyond the specific sample.

The 24617-2 ISO data model forms the basis for a domain-specific data collection set-up specifying the type of interaction, participants roles, tasks and actions performed. For example, in our negotiation training scenario, we have a negotiation *session* consisting of one or multiple training *rounds* featuring different goals assigned to trainees by a *Tutor*. Tutors (humans or simulated agents) attend the session and provide feedback to *Trainees* performing a negotiation task. Tutoring interventions are expected to inform trainees of mistakes, propose corrections, provide instructions, initiate 'try again' rounds, or highlight trainees' successes. This involves immediate real-time 'in-action' and summative 'about-action' feedback (Schön, 1983). The

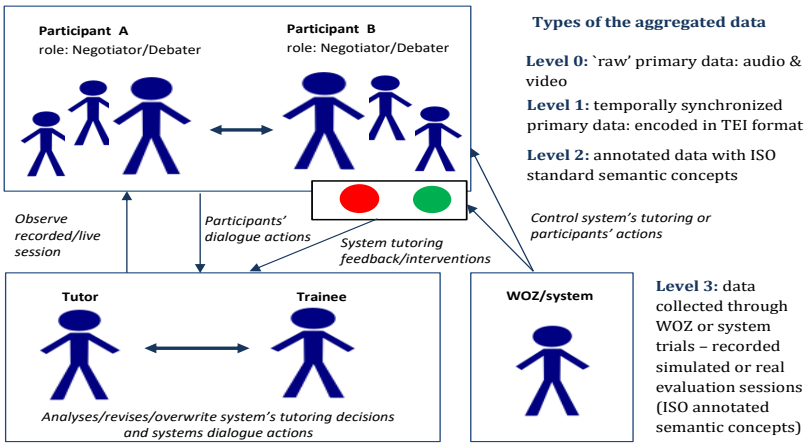

Figure 3: Example of the system and data collection set-up.

task of trainees as *Negotiators* is to propose offers and react to offers of the partner. An extended ISO 24617-2 metamodel (see concepts marked red in Fig. 2) underlies all system and corpus development. A general framework for data collection is set up as shown in Figure 3. We specify participants roles and tasks, as well as data types collected at each recording, processing and evaluation stages including simulated and real dialogue system behaviour in the role of tutor and participant.

The technical set-up specifies recording conditions, equipment, instructions for technical personnel, as well as details on type and granularity of data that should be recorded, and how and where it should be stored, see (Haider et al., 2017).

4.2. Collection and Processing

In multimodal dialogue applications, speech is the main modality. Speech recordings should be of sufficient quality to be used for further processing. Our experience is that recorded 96KHz/24bits audio signals allow a very good tracking of prosodic variations and can be down-sampled to train an Automatic Speech Recognition (ASR) system.

Body movements and facial expressions were tracked using a Kinect 3D sensor. The Kinect video streams and tracking data were temporally synchronised with audio signals with frames of equal 33ms size. The resulting media were converted to view, browse and annotate using the Anvil tool[3].

The speech of a dialogue participant was transcribed semi-automatically by (1) running the ASR system and (2) correcting transcriptions manually. All transcription were stored per participant and dialogue session in TEI compliant format (ISO, 2006).

Prosodic properties related to voice quality, fluency, stress and intonation were computed using PRAAT (Boersma and Weenink, 2009). Kinect body and face tracking data were stored in an `XML` format with elements for frames, faces, joint orientation and bone rotation.

4.3. Annotating and Modelling

The ISO 24617-2 dialogue act taxonomy is designed to capture the meaning of dialogue contributions in multiple dimensions, resulting in multi-layered annotations. Nine dimensions are distinguished, addressing information about a certain *Task*; the processing of utterances by the speaker (*Auto-feedback*) or by the addressee (*Allo-feedback*); the management of difficulties in the speaker's contributions (*Own-Communication Management*) or that of the addressee (*Partner Communication Management*); the speaker's need for time to continue the dialogue (*Time Management*); the allocation of the speaker role (*Turn Management*); the structuring of the dialogue (*Dialogue Structuring*); and the management of social obligations (*Social Obligations Management*).

The semantic content of a dialogue act can be specified in terms of predicate-argument structures, named entities, semantic roles, etc., applying other available standards of the ISO Semantic Annotation Framework. An example of domain-specific semantics is provided for negotiation dialogues in terms of negotiation events such as offer, counter-offer, concession, etc., and their arguments.

In negotiation dialogues, the majority of utterances (59%) is modalized. Participants introducing their options provide information about preferences and abilities. They also request the preferences of their opponents. Parties tend to mention the least desirable events. Apart from preferences and dislikes, a negotiator has certain goals to achieve, which are signalled by teleological modal expressions. Thus, the use of prioritizing modality is frequent. Modality corresponds to the speaker's evaluation of the probability of events; it concerns what the speaker believes to be possible, necessary or desirable. Thus, the classified modality related to the speaker's preferences, priorities, needs and abilities is defined (Lapina and Petukhova, 2017). The metamodel is extended accordingly.

Relations between dialogue acts were annotated, such as the question-answer functional dependence relation, the relation between a feedback act and the dialogue part that the feedback is about, and rhetorical relations, see also (Petukhova et al., 2011). The recognition of dependence and rhetorical relations allows context-dependent interpretation of speaker intentions as well as processing of inter-sentential phenomena, e.g. co-reference resolution.

The full DiAML representation of utterance *P1: I prefer all outdoor smoking allowed* produced by the sender P1 addressed to P2 is a task-related Inform act with the semantic content $\Box offer(ISSUE = 1; VALUE = A)$ is as follows:

```
<dialogueAct xml:id="da1" sender="#p1"
   addressee="#p2" dimension="task"
```

[3] http://www.anvil-software.org/

```
    communicativeFunction="inform"
    target="#fsp1TSK38" qualifier="certain">
     <NegotiationSemantics>
     <NegotiationMove xml:id="nm1"
        type="offer"/>
        <Arg>issue-1; option-A</Arg>
        <modalLink holder="#p1" target="#nm1"
           modalRel="preference"/>
     </NegotiationSemantics>
</dialogueAct>
```

Note that we introduced a <NegotiationSemantics> element into DiAML to represent the domain-specific semantic content of a dialogue act for negotiations. This gives certain flexibility allowing to plug in other domain-specific semantics into DiAML. For instance, for debates, a <DebateSemantics> DiAML element was specified.

```
<dialogueAct xml:id="da1" sender="#p1"
  addressee="#p2" dimension="task"
  communicativeFunction="inform"
  target="#fs38"qualifier="certain">
    <DebateSemantics>
     <Argument type="for"/>
        <Topic>tax\_increase</Topic>
    </DebateSemantics>
</dialogueAct>
```

4.4. Implementation and Testing

The Virtual Negotiation and Debate Coaches "hear" and "see" a wide range of signals, interpret them and act as a negotiation partner or debate opponent, and/or as a tutor.

The speech signals and tracking data serve as input for further processing. The Kaldi-based ASR (Povey, 2011) was trained based on 759 hours of data [4] achieving performance of 34.4% Word Error Rate (WER), see (Singh et al., 2017). For semantic interpretation, the ASR output was used for the event, arguments and modality classification, and communicative function recognition. Conditional Random Fields models (Lafferty et al., 2001) were trained to predict negotiation moves which specify events and their arguments, as well as their boundaries in ASR 1st-best string. The classifier predicts three types of classes: negotiation move (event), issue and preference value (event participants, i.e. semantic roles). A 10-fold cross-validation using 5000 words of transcribed speech from the negotiation domain yielded an F-score of 0.7 on average. The obtained interpretation is of type $offer(ISSUE = X; VALUE = Y)$. The Support Vector Machine (Vapnik, 2013) modality classifiers show accuracies in the range between 73.3 and 82.6% (Petukhova et al., 2017a). The obtained interpretation of a modalized negotiation move stating preference is represented as $\Box offer(ISSUE = X; VALUE = Y)$.

The manually ISO 24617-2 annotated Debate Trainee Corpus (Petukhova et al., 2017b) and Multi-issue Bargaining Corpus (Petukhova et al., 2016) were used to train various communicative function classifiers. Additionally, the in-domain data was enriched with those from the Map-Task (Anderson et al., 1991), AMI(Carletta, 2006), and

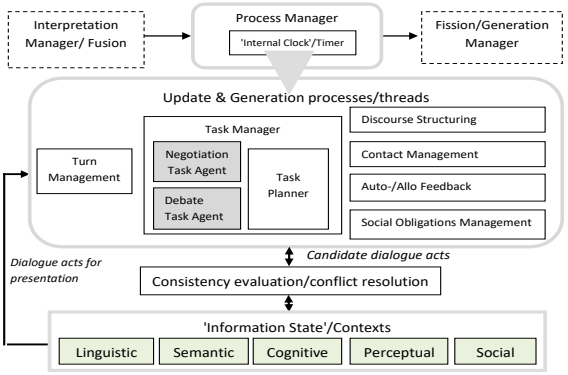

Figure 4: Negotiation and Debate Task Agents (gray boxes) incorporated into the Dialogue Manager architecture.

Switchboard-DAMSL (Jurafsky et al., 1997) corpora. F-scores ranging between 0.83 and 0.86 were obtained in SVM-based clasification experiments, which corresponds to state-of-the-art performance, see (Amanova et al., 2016). Kinect tracked data is used to detect hand/arm co-speech gestures[5] and their types, e.g. beats, adaptors, iconic, deictics and emblems. SVM and Gradient Boosting (Friedman, 2002) classifiers were trained and achieved F-scores of 0.72 (Petukhova et al., 2017c). The motion interpretation component related to hand/arms position detection of the designed Presentation Trainer ((Van Rosmalen et al., 2015; Schneider et al., 2015)) is integrated into the VDC system. Annotations of dependence relations and discourse relations were used to obtain context dependent interpretation. Dependence relations were straightforwardly computed from the dialogue history stored in the linguistic context of the Dialogue Manager (DM), see below. The discourse relations recognition is important for discourse-based argument structure recognition (Petukhova et al., 2017b). The SVM-based classifier yielded F-scores of 0.54 on a coarse 3-class task (Contingency, Evidence, No-Relation) and 0.46 on a fine-grained 7-class task (Justification, Reason, Motivation, Exemplification, Explanation, Exception and No-Relation).

At the semantic fusion level, verbal, prosodic and motion tracking information is combined to obtain complete multimodal dialogue act interpretations, consumed by the Dialogue Manager (DM). The DM, designed as a set of processes (threads), receives data, updates the information state and generates the system next action(-s), see also (Malchanau et al., 2015). The DMs in the VNC and in the VDC applications differ, since the two systems have different roles and tasks. As the Debate Coach, the systems observes debaters' behaviour, evaluates it on criteria related to (1) how convincing is a debater's argumentation; (2) how well are debate arguments structured; and (3) how well is an argument delivered, and generates real-time 'in-action' feedback, see (Petukhova et al., 2017b). As the Negotiation Coach, the system performs as a negotiation partner and also provides feedback on a trainee's negotiation behaviour. Here, the DM incorporates an Negotiation Task Agent (NTA), which interprets and produces negoti-

[4] The following resources were used: the Wall Street Journal WSJ0 corpus, HUB4 News Broadcast data, the VoxForge, the LibriSpeech and AMI corpora.

[5] Co-speech gestures are visible hand/arm movements produced alongside speech and are interpretable only through their semantic relation to the synchronous speech content.

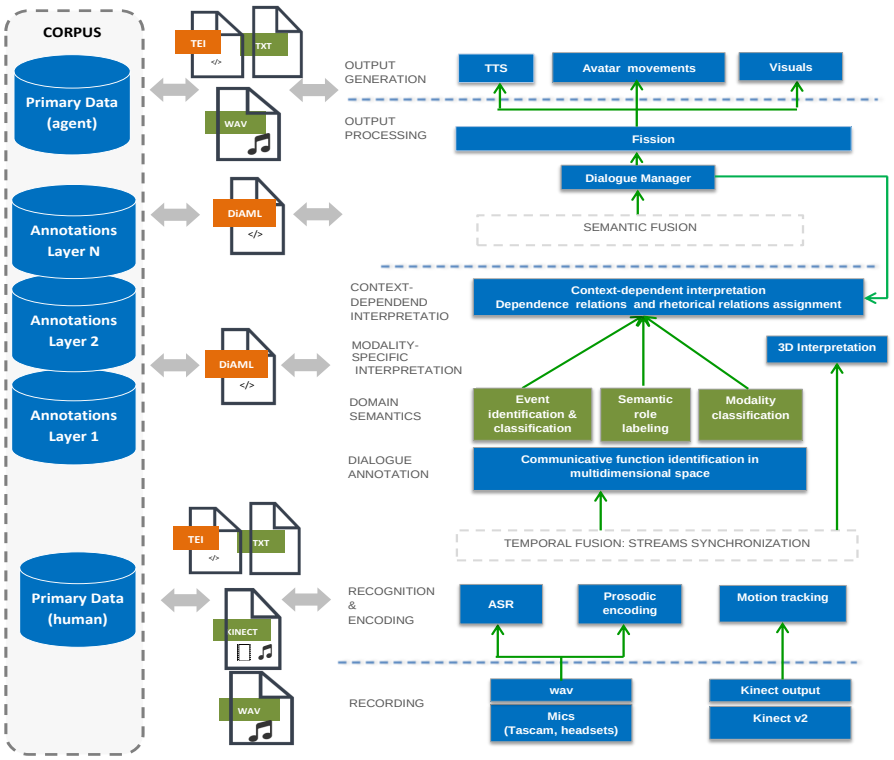

Figure 5: Overall corpus creation and system architecture. From bottom to top, signals are received through input devices, and processed by tailored modules. After annotation concerned with communicative function classification, domain-specific semantic and modality-specific interpretations, context-depended representations are fused and passed to the Dialogue Manager for context model update and system response generation. The system output is rendered in different output modalities. The generated behaviour is written back to the corpus as primary data and representations as annotations, and proposed for editing by human annotators and system module re-training.

ation actions based on the estimation of partner's preferences and goals, and adjusts its strategy according to the perceived level of the opponent's cooperativeness. It reasons about the overall state of the negotiation task, and attempts to identify the best negotiation move for the next action. The DM takes care of feedback and dialogue control actions concerning contact and social obligations management, as well as recovery and error handling actions.

While task-related dialogue acts are application- and user-specific, in a shared cultural and linguistic context, the choices concerning the frequency of dialogue control actions and the variety of expressions are rather limited, notably for feedback and turn management. Models of dialogue control behaviour once designed can therefore be applied in a wide range of communicative situations. This was one of the main motivations behind the multi-layered, multi-threaded DM architecture (Figure 4) where task-related and dialogue control agents/managers are separated. When integrated into different dialogue systems mostly parts of Task Managers are replaced, while other parts were largely re-used without sever changes.

Given the dialogue acts provided by the DM, the Fission module generates system responses, splitting content into different modalities: Avatar[6] and Voice (TTS[7]) actions are generated for the system in partner mode, and visual feedback as tutoring actions. The latter include feedback on presentational aspects and cooperativesness level, visualized by happy and sad face emoticons. At the end of each negotiation and debate session, summative feedback is generated about several aspects of the trainee performance and learning progress.

5. Corpus Evaluation and Deployment

Full session recordings, system recognition and processing results, and the generated dialogue system responses were logged and converted to `.anvil` format for post-processing with the Anvil tool. This tool allows user-defined coding schemes, offering various tier relationships and controlled vocabularies. The tiered format is convenient for transcriptions and annotations in multiple modalities and dimensions. Stretches of communicative multimodal behaviour are marked up with multiple tags, especially when the various tags provide functional information relating to a particular dimension of interaction, such as feedback, turn taking, or time management, see (Petukhova and Bunt, 2010; Bunt et al., 2012; Petukhova, 2011). Annotations are stand-alone and performed using the Anvil specification designed for ISO 24617-2[8].

The Anvil functionality was extended to allow experimenting with variations in system behaviour by tuning, replaying and repairing it. Corrected transcriptions and annotations served: (1) evaluation, measuring inter-annotator agreement to assess corpus data usability, and module-

[6]Commercial software of Charamel GmbH has been used, see (Reinecke, 2003)

[7]Vocalizer of Nuance, `http://www.nuance.com/for-business/text-to-speech/vocalizer/index.htm`, was integrated.

[8]An example specification is available at `http://www.anvil-software.org/data/diaml-spec-v0.5.xml`

based evaluation contrasting system and human performance on all processing tasks; (2) revision of scenario, requirements and data models; and (3) re-training modules on more and better data in order to improve the system performance.

Two resulted corpora are evaluated and deployed when designing the Virtual Negotiation Coach and Virtual Debate Coach applications. They are documented and either released or are in preparation to be released to the research community - Multi-Issue Bargaining Corpus[9] and Debate Trainee Corpus[10].

Figure 5 summarizes the overall corpus and system development framework.

6. Conclusions and future work

Given the importance for a wide range of linguistic applications of data annotated with the interoperable semantic concepts there is a need for cost-effective and accountable solutions to acquire and create such resources on a large scale. This paper proposed the continuous corpus creation methology, supported by ISO semantic annotation standards. On the one hand, the application of the methodology leads to the creation of new interoperable dialogue resources, and on the other hand it enables the design, evaluation and improvement of dialogue system components using these resources. In this approach a corpus is used as a common shared repository which is continuously updated with new recorded and processed data and which is used to generate and tune the system behaviour from it where all system modules exchange messages in standard commonly accepted formats. Well-defined standard data models enable these processes.

Further dialogue resources and tools are in preparation for release. Future work will be also concerned with the integration of new and recently updated ISO standard data models such as those for multimodal events, space and quantification.

7. Bibliographical References

Amanova, D., Petukhova, V., and Klakow, D. (2016). Creating annotated dialogue resources: Cross-domain dialogue act classification. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Slovenia.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC Map Task corpus. *Language and speech*, 34(4):351–366.

Bayer, A. O., Stepanov, E. A., and Riccardi, G. (2017). Towards end-to-end spoken dialogue systems with turn embeddings. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2516–2520, Sweden.

Bennett, C. and Rudnicky, A. (2002). The Carnegie Mellon Communicator corpus. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado.

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer program. Available at `http://www.praat.org/`.

Bunt, H., Kipp, M., and Petukhova, V. (2012). Using DiAML and ANVIL for multimodal dialogue annotation. In *Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2012)*, Turkey.

Bunt, H., Fang, A. C., Liu, X., Cao, J., and Petukhova, V. (2013). Issues in the addition of ISO standard annotations to the Switchboard corpus. In *Proceedings of the 9th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-9)*, pages 59–70, Germany.

Bunt, H., Petukhova, V., Malchanau, A., A., F., and Wijnhoven, K. (2016). The DialogBank. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Slovenia.

Bunt, H., Petukhova, V., and Fang, A. (2017). Revisiting the ISO standard for dialogue act annotation. In *Proceedings of the 13th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 37–51, France.

Bunt, H. (2015). On the principles of semantic annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, UK.

Bunt, H. (2017). Towards interoperable annotation of quantification. In *Proceedings of the 13th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 84–95, France.

Carletta, J. (2006). Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of Oz studies - why and how. *Knowledge-based systems*, 6(4):258–266.

Daubigney, L., Geist, M., Chandramohan, S., and Pietquin, O. (2012). A comprehensive reinforcement learning framework for dialogue management optimization. *IEEE Journal of Selected Topics in Signal Processing*, 6(8):891–902.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Haider, F., Luz, S., and Campbell, N. (2017). Data collection and synchronisation: Towards a multiperspective multimodal dialogue system with metacognitive abilities. In *Dialogues with Social Robots*, pages 245–256. Springer.

Henderson, M., Thomson, B., and Williams, J. D. (2014). The second dialog state tracking challenge. In *Proceedings of the 15th Annual SIGdial Meeting on Discourse and Dialogue*, pages 263–272.

Ide, N. and Pustejovsky, J. (2017). *Handbook of Linguistic Annotation*. Springer.

Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Natural language engineering*, 10(3-4):211–225.

ISO. (2006). *TEI-ISO 24610-1:2006 Language resource management: Feature structures, Part 1: Feature structure representation*. ISO, Geneve.

ISO. (2012). *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO*

[9] `https://catalog.ldc.upenn.edu/LDC2017S11`

[10] See Petukhova et al., 2018, in this volume.

24617-2. ISO Central Secretariat, Geneva.

ISO. (2016). *Language resource management – Semantic annotation framework – Part 6: Principles of Semantic Annotation. ISO 24617-6*. ISO Central Secretariat, Geneva.

Jurafsky, D., Schriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation: Coders manual.

Lafferty, J., McCallum, A., Pereira, F., et al. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lapina, V. and Petukhova, V. (2017). Classification of modal meaning in negotiation dialogues. In *Proceedings of the 13th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 59–70, France.

Lee, K., Zhao, T., Du, Y., Cai, E., Lu, A., Pincus, E., Traum, D., Ultes, S., Rojas Barahona, L. M., Gasic, M., Young, S., and Eskenazi, M. (2017). Dialport, gone live: An update after a year of development. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 170–173, Germany.

Malchanau, A., Petukhova, V., Bunt, H., and Klakow, D. (2015). Multidimensional dialogue management for tutoring systems. In *Proceedings of the 7th Language and Technology Conference (LTC 2015)*, Poland.

Petukhova, V. and Bunt, H. (2010). Towards an integrated scheme for semantic annotation of multimodal dialogue data. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

Petukhova, V. et al. (2018). The Metalogue Debate Trainee Corpus: Data collection and annotations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation(LREC 2018)*, Japan.

Petukhova, V., Prévot, L., and Bunt, H. (2011). Multi-level discourse relations between dialogue units. In *Proceedings of the 6th Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-6)*, pages 18–28, Oxford: Oxford University.

Petukhova, V., Gropp, M., Klakow, D., Schmidt, A., Eigner, G., Topf, M., Srb, S., Motlicek, P., Potard, B., Dines, J., et al. (2014a). The DBOX corpus collection of spoken human-human and human-machine dialogues. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Iceland.

Petukhova, V., Malchanau, A., and Bunt, H. (2014b). Interoperability of dialogue corpora through ISO 24617-2-based querying. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Iceland.

Petukhova, V., Stevens, C., de Weerd, H., Taatgen, N., Cnossen, F., and Malchanau, A. (2016). Modelling multi-issue bargaining dialogues: Data collection, annotation design and corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Slovenia.

Petukhova, V., Bunt, H., and Malchanau, A. (2017a). Computing negotiation update semantics in multi-issue bargaining dialogues. In *Proceedings of the SemDial 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, Germany.

Petukhova, V., Mayer, T., Malchanau, A., and Bunt, H. (2017b). Virtual Debate Coach Design: Assessing multimodal argumentation performance. In *Proceedings of the 2017 ACM on International Conference on Multimodal Interaction (ICMI 2017)*, UK. ACM.

Petukhova, V., Raju, M., and Bunt, H. (2017c). Multimodal markers of persuasive speech : designing a Virtual Debate Coach. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 142–146, Sweden.

Petukhova, V. (2011). *Multidimensional Dialogue Modelling. PhD dissertation*. Tilburg University, The Netherlands.

Povey, D. (2011). The Kaldi speech recognition toolkit. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, Big Island, HI, US. IEEE Signal Processing Society.

Pustejovsky, J., Bunt, H., and Zaenen, A. (2017). Designing annotation schemes: From theory to model. In *Handbook of Linguistic Annotation*, pages 21–72. Springer.

Reinecke, A. (2003). Designing commercial applications with life-like characters. *Lecture notes in computer science*, pages 181–181.

Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

Schneider, J., Börner, D., Van Rosmalen, P., and Specht, M. (2015). Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI 2015)*, pages 539–546, Seattle, WA, USA. ACM.

Schön, D. A. (1983). The reflective practitioner: How professionals think in action. In T. Smith, editor, *Basic Books*. Temple Smith, London.

Singh, M., Oualil, Y., and Klakow, D. (2017). Approximated and domain-adapted LSTM language models for first-pass decoding in speech recognition. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Sweden.

Van Rosmalen, P., Börner, D., Schneider, J., Petukhova, V., and Van Helvert, J. (2015). Feedback design in multimodal dialogue systems. In *Proceedings of the 7th International Conference on Computer Supported Education*, pages 209–217, Portugal.

Vapnik, V. N. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th European Chapter of the Association for Computational Linguistics (EACL 2017)*, Spain.