Dialogue & Discourse 16(3) (2025) 96–130 doi: 10.5210/dad.2025.305

Prior Lessons of Incremental Dialogue and Robot Action Management for the Age of Language Models

Casey Kennington CASEYKENNINGTON@BOISESTATE.EDU
DEPARTMENT OF COMPUTER SCIENCE
BOISE STATE UNIVERSITY

Pierre Lison PLISON@NR.NO
NORWEGIAN COMPUTING CENTER

David Schlangen DAVID.SCHLANGEN@UNI-POTSDAM.DE
COMPUTATIONAL LINGUISTICS
UNIVERSITY OF POTSDAM

Editor: Hendrik Buschmeier

Submitted 10/2023; Accepted 11/2025; Published online 12/2025

Abstract

Efforts towards endowing robots with the ability to speak have benefited from recent advancements in natural language processing, in particular large language models. However, current language models are not fully incremental, as their processing is inherently monotonic and thus lack the ability to revise their interpretations or output in light of newer observations. This monotonicity has important implications for the development of dialogue systems for human–robot interaction. In this paper, we review the literature on interactive systems that operate incrementally (i.e., at the word level or below it). We motivate the need for incremental systems, survey incremental modeling of important aspects of dialogue like speech recognition and language generation. Primary focus is on the part of the system that makes decisions, known as the dialogue manager. We find that there is very little research on incremental dialogue management, offer some requirements for practical incremental dialogue management, and implications of incremental dialogue for embodied, robotic platforms in the age of large language models.

Keywords: spoken dialogue systems, incremental, human-robot interaction, dialogue management

1. Introduction

Large Language Models (LLMs) have become more prominent in robotics, and for good reason. Williams et al. (2024) explain that LLMs can offer "quick-enabling of full-pipeline solutions" for many aspects of robots ranging from enabling robots to engage humans in spoken interaction to generating action plans (Singh et al., 2023; Cohen et al., 2024; Singh et al., 2024; Mahadevan et al., 2024) and emotional behaviours (Mishra et al., 2023). While promising, some recent work has identified important qualities that LLMs lack which, if part of the model, would make interaction with robots seem more natural. A recent survey of spoken interaction on robots by Reimann et al. (2024) showcases a long history of research that *spoken dialogue systems* (SDSs) are key to endowing robots with handling common artifacts in spoken interaction which are not commonly found in text or written interaction, including (*inter alia*) turn-taking, requests for clarification and building

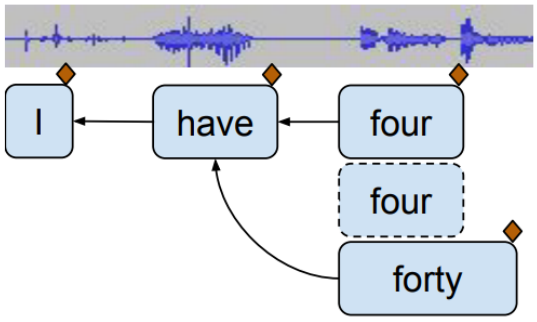

Figure 1: Example of incremental processing for speech recognition: *I*, *have*, and *four* are recognized, then four is revoked and replaced with *forty*. Diamonds denote the point in time when the information is passed to the next module. Figure adapted from Kennington et al. (2017).

common ground. At the heart of their focus is the *dialogue manager* because both SDSs and robots must *make decisions* about which actions they will take at any given moment, either by uttering a response, moving a robotic arm, or any other potential action within the capabilities of the robot. The DM (or corresponding robot action manager) not only decides which action to take, but also *when* to take that action; both are critical for natural interaction between robots and humans (Lison and Kennington, 2023). Both studies look at DM in HRI tasks and settings, comparing how different systems divide the decision-making responsibilities, concluding that DM on robots is still rather a new field; more data, tasks, benchmarks, and discussions are needed.

Fortunately, there exists a body of literature that spans over 25 years (Allen et al., 2001) of success in developing and improving systems that enable humans to talk naturally to machines: *incremental dialogue*, meaning that processing happens at a fine-grained, word-by-word level (see example in Figure 1). Comparisons between incremental and non-incremental systems have shown that incremental systems significantly improve system performance (Ghigi et al., 2014a), are perceived by humans as being more natural (Aist et al., 2007; Asri et al., 2014) and human-like (Edlund et al., 2008), which suggests that the most appropriate systems for robots should be incremental, echoing the requirements of "robot-ready" SDS for use in *human-robot interaction* (HRI) settings (Kennington et al., 2020). In this paper, we review literature relating to incremental SDS and explain how the *Incremental Unit* framework has influenced incremental research (Section 2.2.1).

LLM research can greatly benefit from this knowledge on incremental processing. Inoue et al. (2024b) points out that, for example, when humans engage in real-time dialogue with robotic agents, humans expect the robot to take seamless turns (i.e., without a gap in conversation, as happens within human-human dialogue) and they expect backchannels (e.g., nodding, or utterances like *yeah*, or *uh huh*), neither of which are handled with common LLMs. Their work applied *Voice Activity Projection* (VAP) to enable LLMs to predict when a person might stop speaking so the LLM can respond at an appropriate time. Chiba and Higashinaka (2025)'s recent VAP method also used an LLM to predict when to start speaking, and their work directly relied on incremental processing. According to the authors: "[...]even if systems are equipped with a natural turn-taking model, such a model will be ineffective if response generation cannot begin immediately once a turn-

shift is detected. Incremental response generation is an approach that addresses this issue." Using LLMs in incremental settings is a positive step, but more work is needed. Furthermore, other recent work (Hudeček and Dusek, 2023) asked if LLMs are all that is necessary for task-oriented dialogue (albeit outside of dialogue with robots) with some negatives (e.g., "LLMs underperform[...]" in important aspects of dialogue) and positives (e.g., "LLMs show the ability to guide the dialogue to a successful ending"). Finally, Wagner and Ultes (2024) investigated the usefulness of LLMs in dialogue interaction and found that they are effective, but need control and guidance to ensure that dialogue responses are coherent— two critical aspects in scenarios where robots are involved.

Taken together, while LLMs can be employed for a wide range of dialogue processing tasks, they still suffer from a number of limitations when it comes to incrementality. In particular, while a LLM decoder can process any kind of input including word-level input, they are trained to act upon complete, sentence-level input, rendering them unable to produce behaviour where input and output happen concurrently – although recent work on full-duplex models shows promising results (Zhang et al., 2025). In contrast, humans must process individual words while reading text, and psycholinguistic research has shown that speech comprehension happens at a word or even sub-word level (Tanenhaus and Spivey-Knowlton, 1995). Moreover, another requirement of incremental processing is *non-monotonicity*; i.e., that a model can react to change in input, for example when information coming from a speech recognizer is incorrect, the model needs to be able to revoke the erroneous input and change its internal state–causal language modeling is strictly monotonic, but incremental processing should allow for non-monotonic input.

With chatbots, the text-in, text-out nature of the interaction is well-suited for LLMs, but the expectation of human-like conversation becomes more challenging when people interact with robots due to the anthropomorphic characteristics of many robotic platforms. If, for example, a robot has what appear to be eyes, people expect that the robot can see them, or if the robot has an arm they expect the robot to be able to point or grasp objects. Furthermore, it has been shown that people anthropomorphize robots for gender (Reich-Stiebert and Eyssel, 2017; Eyssel and Hegel, 2012), intelligence (Novikova et al., 2015), and even age (Plane et al., 2018) depending on the robot's morphology, size, and movements, which affects the expectations of how robots behave: the more anthropomorphic a robot appears, the more human-like people tend to expect the robot to act.

In this paper, we review the literature for incremental SDS with a particular focus on the decision-making component known as *dialogue management* (DM; explained further below) for the sake of guiding ongoing and future work related to decision making on robots that interact with humans. We find in our review that that although other elements of SDS – such as automatic speech recognition and natural language generation – have seen substantial work on incrementalization, there is a noticeable lack of focus on incremental decision making. We identify some of the challenges and requirements to help guide future research on incremental decision making. The next section begins with background on incremental SDSs, focusing first on common modules then fully implemented and evaluated systems. The section that follows then focuses on DM, giving first a brief overview of DM research, then focuses on incremental DM. We then end this review with some concluding remarks and suggested paths for future work.

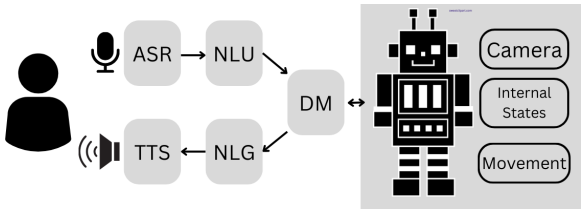

Figure 2: Traditional architecture for spoken dialogue systems composed of Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialogue Management (DM), Natural Language Generation (NLG), and Text-to-Speech Synthesis (TTS). The system is extended to be multimodal, where the added modalities are robot sensors and control.

2. Background: Incremental Spoken Dialogue Systems

In this section, we review literature on common incremental spoken dialogue system modules except DM, which we save for the following section. We explain incremental frameworks that have been adopted, and explain different paradigms of modeling incremental processing.

2.1 Spoken Dialogue Systems: Overview

Equally important to the distinction between incremental (word-level) and non-incremental (utterance or sentence-level) SDS is the distinction between end-to-end and modular SDSs. An end-to-end system is modeled using a single model that takes in input and produces an expected output directly, such as a question-answering system that produces an answer given a question, or social chatbot that produces responses given text input. End-to-end systems often focus on the capability of producing a written or spoken response no matter what the input is. End-to-end architectures now constitute the dominant approach for developing open-domain dialogue systems where the main focus is the social aspect of the interaction (Roller et al., 2020; Ni et al., 2023). The social aspects of interaction are, of course, important in a natural dialogue, but in task-oriented dialogue there is often something that is required outside of the dialogue itself for the dialogue to be considered successful; e.g., look up information in a database, perform some kind of robotic action, or complete a payment. Modular SDSs are often *task-based* in that they help the user achieve such a goal such as booking a flight; they do not usually focus on social aspects beyond what helps to accomplish the task (Budzianowski et al., 2018; Zhang et al., 2020b). This traditional distinction between open-ended end-to-end systems and task-based modular architectures is, however, increasingly blurry, as recent years have seen the emergence of end-to-end models specifically designed for task completion (Liu et al., 2018; Zhang et al., 2020a; Hosseini-Asl et al., 2020; Young et al., 2022) as well as newer agentic architectures. Interestingly, end-to-end models for task-oriented systems often operate by augmenting the generative model with implicit "modules" in the form of retrieval mechanisms (Qin et al., 2019), knowledge bases (Yang et al., 2020) or domain-specific ontologies (Chen et al., 2023), or by pre-training the response generation model in a modular fashion (Qin et al., 2023).

As the name suggests, modular systems are made up of modules that have well-defined roles in the system, and which can be made to communicate with each other. Figure 2 depicts visually a modular SDS. For example, a prototypical SDS is often made up five modules including automatic

speech recognition (ASR) that transcribes speech to a text representation of the human utterances, natural language understanding (NLU) that takes the text and yields a computable semantic abstraction, dialogue management (including dialogue state tracking) that makes a high-level decision about the next action to take (e.g., look up information in a database and respond to the user), natural language generation (NLG) that takes the dialogue manager's decision and determines which words to use and in what order, and text-to-speech (TTS) which actually speaks the words. These modules are further explained below.

Modular SDSs that process incrementally have an added complexity in that all of the modules must operate at granularities that downstream modules can make use of, such as at the word level from ASR to NLU. For example, given a system on a robot that is made up of the standard five modules as explained above (along with connections to robotic modules), and someone utters *Hand me the green book on the left*, an incremental SDS begins to process as soon as speech is detected. The ASR outputs each word, one at a time, and the NLU updates its interpretation each time a word is outputted by the ASR and the NLU likewise produces outputs as it gathers information about the utterance, for example, tagging *hand* as the action as the first word is uttered, and a specific book as the target once *the green book* it has been uttered. The DM is tasked with querying a module that takes in visual information and instructing an arm to reach for the book in question. An incremental DM might already extend its arm in no particular direction as the first word is uttered to signal understanding, then towards any green book once *green book* is uttered, then narrow the target down further as *on the left* is uttered. The NLG could then start uttering a response like *green book* as it begins to move its arm then *ah, here we go* once it determines a unique referent.

The above example highlights some things that differentiate an incremental DM from a more traditional DM. First, the incremental DM receives installments of information over time, whereas a traditional DM receives all of the information at once after everything has been uttered and the NLU has finished processing the ASR's transcription. The incremental DM has therefore an important role that is lacking in non-incremental DM it not only must decide which action to take, but it also must decide *when* to take that action given the information that it has so far, and—perhaps a bigger challenge—perform concurrent actions as it is still receiving input (i.e., "full-duplex" models). Traditional SDS has often relied on endpointing; i.e., waiting for silence after a speaker begins to speak, which burdens the ASR with determining *when* to act. However, pauses in speech are not always signals that someone is done speaking, and incremental SDS that relies on a DM to determine when to take an action can potentially use speech, silence, as well as information from the content of the utterance (i.e., via the NLU) to make decisions about *when* to act.

2.2 Frameworks & Architectures

2.2.1 THE INCREMENTAL UNIT FRAMEWORK

The incremental unit (IU) framework, a well-established approach to incremental processing, will be a recurring reference in this paper, following the works of Schlangen and Skantze (2009, 2011). The IU framework views each bit of information created by the modules (e.g., words produced by ASR and slots produced by NLU) as part of a global network of interconnected IUs no matter which module produced them. The framework defines functions for changing the network including how nodes of the network are added and how the nodes are interconnected. Newly created IUs by a module (e.g., words by ASR) can be *added* to the IU network, *revoked* from the network if the module determines that an IU was erroneously added in light of new information (e.g., the ASR first

added the word IU *four* but later revoked and added *forty*), and IUs can be *committed*, meaning they have already been added to the IU network, and are guaranteed to not be revoked.

To be added to the IU network, an IU must be connected to other IUs that already exist in the network through two relations:

- *Same-level links* which are relations between IUs created by the same module – e.g., if the ASR recognizes *the* and *dog* as two IUs, the later word *dog* has a same level link to *the*.
- *Grounded-in links* where a relation is created between and IU an the IU(s) that gave rise to that IUḞor instance, the IUs *the* and *dog* from the ASR might give rise to a *subject* tag in the NLU, leading to two grounded-in links, one to each word IU.

When IUs are operated on (i.e., added, revoked, or committed) the modules that triggered the operation signal downstream modules that consume their output about the change. For example, as the ASR module recognizes words from a microphone, it adds each of them to the IU network and signals to the NLU module that a new word has been added.

A module's ability to *revoke* an IU is important in a natural dialogue interaction. The above ASR example of revoking is fairly straight-forward and happens internally between the modules, but there are cases where modules that produce output need to revoke information, for example a robot is planning on uttering something that, given new information, should be changed. In a setting where a robot and a person are working together to move around colored boxes, if the person says "move the green box to the left" but the ASR system mistakenly recognises "gray" instead of "green", the robot will initially approach and attempt to move a gray box. The robot generates a plan to move towards the box and produce an utterance to signal understanding such as *okay, I'm on my way to move the gray box*. But when the revoke from *gray* to *green* happens, the robot must revise its movement and verbal responses accordingly. If the robot hasn't completed its utterance, it still has a chance to change the utterance to *okay, I'm on my way to move the green box* and change its direction to the box the person referred to, known in robotics as *replanning* (Cashmore et al., 2019). This is an illustration of *non-monotonicity*: modules can update the information that they pass to each other, and often updates need to happen after the robot has already taken some kind of action. More on this in Section 3.

The IU framework has been implemented in several software packages, notably in Java as InproTK (Baumann and Schlangen, 2012) and more recently in Python as Retico (Michael and Möller, 2019) and Remdis (Chiba et al., 2024). Other conceptual frameworks such as the Information State Approach (Traum and Larsson, 2003) and Cohen's belief-desire-intent model (Cohen, 2017) remain valid in incremental SDS, including within the IU framework, though they are not strictly incremental dialogue frameworks. Later versions of InproTK and, more recently, Retico has been extended to include common robot capabilities such as object detection and control of multiple robot platforms (Kennington et al., 2020; Manaseryan et al., 2025), towards bridging the gap between SDS and HRI research. While not strictly incremental, OpenDial (Java and Python) has been used in spoken HRI studies, and can serve as a decision making module in both InproTK and Retico (Lison and Kennington, 2016; Jang et al., 2020).

2.2.2 RESTART VS. UPDATE INCREMENTAL MODELS

Khouzaimi et al. (2014) points out that not all methods and models are inherently incremental, though many can be made to work incrementally under certain constraints. While their proposed

method is an important step in improving the incrementality of current systems, it should be highlighted that there are two distinct approaches to modeling incremental systems: restart incremental and update incremental, which we explain below.

Restart Incrementality Restart incremental models take in inputs and produce incremental outputs (e.g., at the word level), but the input is repeated as the prefix grows, and models themselves are agnostic to the incremental updates. Any model (e.g., a language model using zero-shot classification) could be used restart incrementally. For example, a NLU module that is restart incremental would take in the following input (time moves from top to bottom; each line represents input to a NLU model):

the
the dog
the dog barks

Update Incrementality In contrast to restart incremental models, update incremental models do not need repeated input and the model is designed to maintain a state that updates for each incremental input. An NLU model that works in an update incremental way would not need to repeat a growing prefix from the ASR:

the
dog
barks

The model explained in Kennington and Schlangen (2017), for example, is a Bayesian update-incremental model that produced a distribution over possible slot values that updated the distribution at each word increment. An open question that we explore below is if a DM model should be either restart or update incremental.

2.3 Common Modules in Incremental, Interactive Systems

2.3.1 AUTOMATIC SPEECH RECOGNITION

Current ASR systems receive streaming input and produce partial transcriptions, and can often work at word-level increments. Early incremental ASR were implemented in Sphinx (Baumann et al., 2009), and newer neural ASR systems could be made to operate at the character level (Hwang and Sung, 2016). The most common evaluation metric for ASR is *word error rate*, and recent neural models have shown very low error rates in common ASR benchmark datasets. However, evaluation of incremental ASR requires a closer look at how often a model alters its output and latency of results (Baumann et al., 2016; Whetten et al., 2023). Because of the nature of the IU processed by ASR, evaluating its performance both independently as well as within larger systems is crucial, as certain errors may affect downstream components.

Because ASR requires streaming input, they are inherently incremental in terms of input, though not all recognizers produce incremental output; they often wait until a pause in the speech (i.e., end-pointing). However, recent ASR models have become very effective at accurate transcription in multiple languages and they produce incremental output. The Whisper (Radford et al., 2022), wav2vec (Baevski et al., 2020), and Deep Speech 2 Amodei et al. (2016) all produce incremental output, the former 2 being incorporated into the Retico framework. Imai et al. (2025) recently eval-

uated conversational speech recognition on several state-of-the-art ASR models (including Whisper and wacv2vec), taking gender into account in their evaluation. The results were mixed; ASR has come a long way in the past two decades, but more work is needed to accommodate different demographics of speakers, and correcting errors.

Incremental ASR is more challenging on robots in spoken HRI settings because robot voice can be picked up by the ASR, thereby 'confusing' the robot. This can be at least partially addressed using diarization (i.e., tracking the voice of particular individuals within a speech signal, including a robot) or by the robot tracking its own speech signal and filtering it out of the ASR input.

2.3.2 NATURAL LANGUAGE UNDERSTANDING

Understanding natural language in SDSs also has a long history. In most NLU models, the input corresponds to text. It the case of SDS, the input to NLU is transcribed speech. The output of NLU is important to consider here, because it is often what serves as the input to the DM. The output needs to abstracted sufficiently over the input text to form a computable meaning representation that the DM can use for making a decision on how to act. That meaning representation in incremental NLU has been represented in various ways in the literature including tagged words, logical forms, or frames (i.e., a set of key-value pairs known as *slots*), recent models tend to use tags to produce slots and frames as output, or latent representations (e.g., embeddings). Below is an example frame for an utterance made to command a specific robot action: *Move the red ball into the box on the left* made up of four slots:

`intent`	command
`object`	red ball
`target`	left box
`action`	move `object` to `target`

Like incremental ASR, incremental NLU produces output as early as possible (for example, individual filled slots), but unlike ASR the input is discrete words instead of a continuous speech signal, so the intervals of when output is produced can vary depending on the input and the domain. Early incremental NLU focused on inferring semantic frames. Each input word produced a partially complete frame as output (Devault et al., 2011; DeVault and Traum, 2012, 2013; Yamauchi et al., 2013; Kennington and Schlangen, 2014; Kennington et al., 2014b, 2015). Part of the frame is also the dialogue act; i.e., the overarching type of utterance produced by the interlocutor (e.g., a question or an assertion), which has also a history of incremental models (Petukhova and Bunt, 2011). Early actionable output from NLU is particularly important in HRI settings, where robots can already be moving towards referred objects before the person finishes their request, which is an important physical backchannel: a robot beginning to move towards an object is a signal to the user that the robot is understanding the unfolding utterance, as done in Hough and Schlangen (2016)

Similar to their non-incremental counterparts, incremental NLU can benefit from syntactic parsing to guide language understanding, but in the case of incremental NLU, the parsers must also work incrementally (i.e., produce a partial syntactic parse such as a tree for each word input). There has been ample research in incremental parsing for different syntactic formalisms, including dependency parsing (Nivre, 2008), combinatory categorical grammar parsing (Hassan et al., 2008; Beuck and Menzel, 2013), as well as frameworks with a more semantic focus, such as robust minimal recursion semantics (Copestake, 2007; Peldszus et al., 2012), dynamic syntax (Eshghi et al.,

2013) (see Hough et al. (2015) for a comparison of robust minimal recursion semantics and dynamic syntax for incremental dialogue) and abstract meaning representation (Damonte et al., 2017).

In multimodal SDS and interaction with robots, the NLU component must sometimes resolve references to objects that exist in the shared space with the robotic system and the human interlocutor. Incremental reference resolution can be viewed as the ability to narrow down possible referents in a shared visual space to an individual object. An incremental reference resolution model might, for example, understand the word *red* to refer to objects that have a red color, and *book* to then further narrow down from all red objects to only red books. Incremental reference resolution is sometimes an integral part of NLU (Kennington et al., 2014b), but have also been designed for modules that only resolve references (Schlangen et al., 2009; Paetzel et al., 2015; Kennington and Schlangen, 2015; Schlangen et al., 2016; Kennington and Schlangen, 2017), information that the DM may need to use for making a decision.

Other works have explored how deep learning architectures can be used for incremental NLU, including recurrent architectures (Shivakumar et al., 2019) and to what degree architectures that are not inherently incremental (e.g., self-attention transformers which are designed to process multi-word input in parallel) can be used for incremental NLU (Madureira and Schlangen, 2020), with mixed results. It is important to explore further how neural models can work incrementally because many dialogue phenomena are incremental in nature. For example, Shalyminov et al. (2017) showed that that deep neural dialogue models failed on common spoken phenomena like restarts and self-corrections.

More recent work has explored how transformer LMs can be successfully used for incremental NLU. Madureira and Schlangen (2020) showed that both bidirectional long short-term memory (LSTM) models and transformer-based encoders assume that an input sequence to be encoded is available a-prior in its entirety, to be processed either forwards and backwards (in the case of bidirectional LSTMs) or as a full sequence (in the case of transformer-based encoders). The results of their work support the possibility of using bidirectional encoders in their developed *incremental* mode while training their non-incremental qualities (i.e., parallel processing). Kahardipraja et al. (2021) explored using *linear* transformers with a recurrence mechanism to examine the feasibility of linear transformers for incremental NLU. They found that linear transformers have better performance and faster inference than standard transformers when used in a restart-incremental fashion.

HRI settings make the requirements of NLU more challenging due to the fast-paced, multimodal nature of the interaction. LLMs trained only on text are limited, but the recent proliferation of multimodal LLMs have real potential for being used on robot platforms. Modalities include images, speech, and video, for example the ONE-PEACE (Wang et al., 2023) and PALM-E models (Driess et al., 2023). A recent survey explains the modeling trends (e.g., one vs. two-tower; different methods of representing images) for vision LMs (Fields and Kennington, 2023); improvements in visual LMs will benefit HRI research because robots need to see and talk about objects in a shared space.

Other recent work focuses on how transformers handle incremental NLU revisions. Most transformer models are *causal* in that they are forced to produce a single output once an ambiguity is resolved, but Madureira et al. (2024) proposed an interpretable way to analyse incremental states to show how transformers handle ambiguity. They showed that transformer sequential structures encode information on the garden path effect, as well as the resolution of garden paths. Another model, TAPIR, a two-pass method that modeled the revision process itself showed better performance on incremental metrics compared to transformers used restart-incrementally (Kahardipraja et al., 2023).

2.3.3 NATURAL LANGUAGE GENERATION AND SPEECH SYNTHESIS

Early work in incremental NLG focused on resolving references in situated dialog. Kelleher and Kruijff (2006) presented an approach to generating locative expressions using a basic incremental algorithm that considered *visual salience* as a computation of an object's perceivable size and centrality relative to the viewer, choosing words that distinguish between the target object and distractor objects. While the algorithm the authors present is "incremental", it is not evaluated as a word-by-word incremental model, but given the co-location and potential of being used at the word level, we include it here. More recent work has shown that incremental installments of words that refer to a visual object using a model trained on visual object/word pairings that uses a beam search to determine the best possible word to utter can use a model of vision/word that isn't trained specifically for NLG (Zarrieß and Schlangen, 2016).

Incremental NLG that builds on the IU framework include work that used a buffer of words to be uttered, and three operations ADD, REVOKE, and PURGE were used for operating on the buffer (Dethlefs et al., 2012a). The ADD operation, of course, means a word is added to the buffer and eventually uttered, unless it was REVOKEd (removed from the buffer) or PURGEd (all words currently in the buffer are removed in favor of a new hypothesis/goal). The NLG module often produced words faster than they could be articulated by a TTS, giving an incremental NLG time to determine which words should be uttered, and in which order. The authors also carried out experiments to explore how NLG interacts with output generation of other modalities, such as information on a screen (Dethlefs et al., 2012b). In general, the research has shown how incremental generation produces systems that are more reactive and perceived as more natural to human dialogue partners. In HRI settings, incremental NLG is important, for example, when referring to real-world objects (Zarrieß and Schlangen, 2016).

Others also looked at how incremental multimodal generation affects the interaction qualities when the SDS is part of a virtual agent; including not only NLG but also hand gestures and eye gaze by the agent (Van Welbergen et al., 2012). Instead of planning all articulations before they were realized, the model generated behaviours incrementally and linked increments in the multiple output modalities to each other, so what happened corresponded temporally to other modalities (e.g., saying *that* in conjunction with a pointing gesture). Such articulation requires that the speech synthesis also be incremental, as a system utterance currently being generated by the NLG and transferred to the TTS might change before the TTS actually articulates a word in the utterance, thereby changing prosody or duration; e.g., the system may want to hold the floor longer so will need to take longer to speak or insert artifacts such as *ummm* (Buschmeier et al., 2012; Baumann, 2014).

Improvements in ASR and NLU have enabled systems to be far more capable than even a few years ago, but the biggest gains in NLG have been due to LMs. Generative large LMs are flexible in that inputs can be structured text and models can be made to produce useful structured output that is useful for SDS and HRI. For example, the DM can output a request to look up information in a database, then take that structured information and input it into a LM, which produces a surface utterance that has the necessary information. Fortunately, while LM *input* processing is not incremental as defined, LM *output* is naturally incremental due to inherent modeling (i.e., autoregression). However, while LM output can be paused (Goyal et al., 2023) or interrupted, the output itself cannot be changed given new input.

2.3.4 INCREMENTAL SYSTEMS & EVALUATION

Beyond individual modules, full systems are more complex and difficult to evaluate, though some prior works have shown how incremental systems are better in some domains than their non-incremental counterparts. For example, a virtual in-car dialogue that presented information incrementally was shown to be safer and more effective at helping users remember information (Kousidis et al., 2014). The system was able to detect changes in the car's control (e.g., changing lanes or speed) and if any change was detected, the system would pause its output and resume after the driving was constant. This allowed drivers to focus on driving instead of non co-located interlocutors.

In another system, Fischer et al. (2021) used incremental speech adaptation to initiate human-robot interactions in noisy (in-the-wild) scenarios. The robot incrementally adjusted the loudness of its voice depending on the circumstances, and was perceived positively by human users. Finally, Ghigi et al. (2014b) showed that that an incremental dialogue strategy significantly improved system performance by eliminating long and often off-task utterances that generally produce poor speech recognition results. User behaviour is also affected; the user tends to shorten utterances after being interrupted by the system.

In both of these examples of full system evaluation, human participants were recruited to interact with the systems. Following common human-agent, human-robot, or human-interface research, human participants are often presented with one of two different versions of the system; a baseline/control version and a test version that focuses on a specific system component. For example, the in-car dialogue system was evaluating incremental information presentation and pausing vs. a system that kept talking; in both cases participants were asked a true/false question about the information they just heard which they answered by pressing a button on the steering wheel. Metrics include objective measures and subjective measures. In the case of the in-car system, objective measures included successful lane changes, the true/false answer, and driving speed, all which could be measured at any given time. Subjective measures include surveys about the participants' experience with the system. Objective and subjective measures are often combined to give a more holistic picture of the evaluation. For example, participants may have felt that they drove the same in both the control and test versions of the system (subjective), but failed to answer questions correctly, or failed to drive at the prescribed speed.

Köhn (2018) reviewed incremental processing in the field of natural language processing (including parsing, machine translation, among others which are beyond our scope), pointing out that granularity, grounding, monotonicity, and timeliness are all aspects of incremental processing that play into how incremental systems are perceived. Most incremental SDS research is performed with the level of granularity set at the word level, but it might be better in certain cases to work at sub-word or phrase levels, or on speech directly (see Kebe et al. (2022) for a non-incremental model grounded in raw speech). Grounding, moreover, is how a system aligns its output (in the case of SDS, generated speech) to what is happening in the dialogue state including physical context and the conversation up until that point. Grounding is particularly important (and challenging) for HRI settings and tasks, as the human and the robot need to track objects, dialogue history (including entity tracking; i.e., objects that have been discussed before). Task-completion is a common metric (i.e., did the human and robot pair complete the assigned task, like put together a puzzle?), but in many cases the task is more social and more focused on human impressions of their interactions with the robot.

Another challenge in incremental evaluation that is more specific to incremental processing is monotonicity. Monotonicity is an open question in SDS research; an ideal incremental ASR for example, would only output the correct word as early as possible as they are spoken without the need for revoking and replacing words. Thus while monotonicity is an ideal to strive for, system modules make mistakes and need to be able to repair those mistakes (hence the need for the IU framework), but knowing how monotonic a system or an individual module is can be a useful metric for measuring stability. Finally, timeliness is important: the system needs to respond quickly, but the system should reach a level of confidence that the response is the proper one. The challenge lies in striking a balance between these two opposing optimisations: ensuring timeliness while maintaining accuracy, which underscores the importance of non-monotonic operations.

LLMs are increasingly being used to evaluate dialogue systems, notably through LLM-as-a-judge setups (Gu et al., 2025), although trust and reliability remain important concerns compared to actual human evaluation studies (Pan et al., 2024). Lab settings, however, often do not generalize to real-world settings due to constrained systems and low number of participants. Crowd-sourcing is a way to increase the number of humans who can evaluate a full system, but quality control is often difficult.

3. Review of Incremental Dialogue Management

In this section, we review literature relating to incremental DM. We give an overview of DM, dialogue state tracking, and attempts at modeling incremental DM.

3.1 A Brief Overview of Dialogue Management

DM lies at the crossroads between NLU and NLG and is responsible for controlling the general flow of the interaction, often in relation with the task(s) that should be fulfilled by the dialogue agent. In their seminal work on the *Information State* approach to DM, Traum and Larsson (2003) mention four objectives:

1. updating a representation of the dialogue context on the basis of interpreted communication (from all dialogue participants) ;
2. providing context-dependent expectations for interpretation of observed signals as communicative behaviour ;
3. interfacing with task/domain processing (e.g., database, planner, execution module, other back-end system), to coordinate dialogue and non-dialogue behaviour and reasoning ;
4. deciding what content to express next and when to express it.

Current DM approaches distinguish between two central (and consecutive) components, respectively called *dialogue state tracking* and *action/response selection*.

3.1.1 DIALOGUE STATE TRACKING

The task of maintaining a representation of the current dialogue state over the course of the interaction is called *dialogue state tracking* (Williams et al., 2016; Ren et al., 2018; Heck et al., 2020). The dialogue state aims to reflect the system knowledge of the current conversational situation, and

often includes multiple variables related to the dialogue history, common ground, external context (including the physical context, in the case of human–robot interaction), and the task(s) to perform.

This update of this dialogue state should occur upon the reception of any new observation that may potentially impact the system's understanding of the current conversational situation, such as new user utterances, but also changes in the physical context of the interaction (for instance, new entities perceived in the visual scene, or updates on the current location of the robot). For incremental systems, those observations will typically correspond to incremental units produced by the NLU module.

In task-oriented systems, the dialogue state is often represented as a list of slots to fill (Williams et al., 2016; Mrkšić et al., 2017), where a slot typically represents a required or optional attribute whose value should be derived from the user inputs to complete the task. For instance, a restaurant booking system might have slots for the date, time and number of people. Although such slot-filling representation can be applied to many domains, it remains restricted to a fixed list of predefined slots, and may therefore be difficult to apply to conversational domains with varying numbers of entities and relations between them. This is notably the case in human–robot interaction, where the number of persons in a room, or the number of objects detected in the current visual scene is not fixed in advance and may change over the course of the interaction. In such settings, representing the dialogue state as a *graph* of entities connected through various relations is a preferred alternative (Ultes et al., 2018; Walker et al., 2022).

Approaches to dialogue state tracking also differ in whether they explicitly represent uncertainty related the current dialogue state using probability distributions. Many DM approaches represent the current dialogue state as a mere collection of key-value pairs (slots and their values). Although this representation does simplify both dialogue state tracking and action selection (in particular when this selection is optimized using reinforcement learning), it makes it harder to capture uncertain, ambiguous or untrustworthy information, which may arise from e.g. error-prone sensory inputs (e.g., imperfect object recognition or ASR) or non-deterministic inference (e.g. linguistic ambiguities). An alternative is to explicitly represent the dialogue state as partially observable and define a probability distribution over possible state values (Young et al., 2013; Mrkšić et al., 2017), often called the *belief state*. This belief state can notably be expressed as a Bayesian network over state variables (Thomson and Young, 2010).

3.1.2 ACTION/RESPONSE SELECTION

The second core DM task is *action selection*, whose role is to determine the next (verbal or non-verbal) action(s) that the system should undertake, based on the dialogue state updated through dialogue state tracking. Although those actions frequently correspond to verbal system responses, they may also express other types of actions, such as API calls or high-level physical actions in the case of robotic platforms. A given dialogue state may lead to the selection of several actions to execute in parallel or in sequence (for instance, a robot may simultaneously move to a new location and utter a sentence to describe his movement to the user) or to no action at all.

The selection of the next action/response may take several forms, from handcrafted flowcharts and logical rules to data-driven techniques. Early work includes Larsson (2002), which surveyed existing approaches to DM including logic-based, finite state, form-based, and plan-based approaches, but the author regarded those approaches as limited in their practicality—most were theoretical models without a concrete implementation. To remedy this situation, Larsson (2002) introduced

Issue-based Dialogue Management. Issues are modeled semantically as questions, which can be implemented in multiple theories (e.g., plan-based or form-based). This kind of dialogue is system-driven in that the system has a specific task that it must perform and it drives the dialogue by asking questions to the user, for example an automated travel agency would ask questions about price ranges, travel dates, origin and destination airports, and airlines if it is going to help a user find an appropriate flight. As is the case with most dialogues, a kind of "information exchange" takes place, the system is not requiring anything of the user beyond responding verbally with requested information.

Also seminal is the early work of Cohen and Levesque (1990) on plan-based approaches to DM building on earlier work by Allen (1979). More recently, Cohen and Galescu (2023) showcases a fully working multimodal conversational system that infers users' intentions and plans to achieve those goals. The system can infer obstacles to goals and actions and find ways to address them collaboratively. The DM is broken down int four parts: plan recognition, obstacle detection and goal adoption, planning, then execution. Planning here is an important aspect of the DM; it does not just identify an action to take now, it identifies a plan (i.e., a series of actions) that must be taken to achieve a higher-level user goal, making it potentially more amenable to multimodal (including IVA and robotic) control.

The mapping from dialogue state to action(s) is called a *dialogue policy*, and various methods have been developed to automatically learn such policies from real or simulated dialogue data. Supervised learning techniques can be employed to imitate the conversational strategies followed by human experts in a corpus of dialogue (Griol et al., 2008). However, the behaviour of human experts may be hard to imitate, especially as those experts often base their decisions on a different and richer understanding of the conversational context that what can be captured in a dialogue state. Those supervised learning techniques also suffer from data sparsity problems, as only a small fraction of the state space can realistically be covered by the dialogue examples.

To this end, a range of reinforcement learning methods have been proposed to automatically optimize dialogue policies based on a reward function (Rieser and Lemon, 2011; Young et al., 2013; Williams et al., 2017; Peng et al., 2018). Although the reward function is often defined manually based on the system objectives, it can also be learned from data (Su et al., 2018; Takanobu et al., 2019). The dialogues can be generated automatically using a user simulator (Schatzmann et al., 2006; Chandramohan et al., 2011; Ultes et al., 2017) or from actual dialogues with human users (Su et al., 2016; Shah et al., 2018).

The underlying process to optimize may be either framed as a Markov Decision Process (MDP), or, in case the dialogue state itself is consider to be uncertain, a Partially Observable Markov Decision Process (POMDP). While framing action selection as a POMDP makes it possible to explicitly account for uncertainties about the current dialogue state, it also complicates the dialogue policy optimization, due to the need to derive a policy in a continuous and high-dimensional belief state space. Dialogue policies can also be expressed in terms of probabilistic rules with a skeleton provided by the system designer while the rule parameters are estimated from dialogue data, as shown by Lison (2015a,b). Recent work in reinforcement learning goes well beyond the POMDP model, including reinforcement learning with human feedback (Ouyang et al., 2022), and proximal policy optimization (Shao et al., 2024), each with potential effectiveness for DM. Some recent work shows how LMs can be used for DM (Niu et al., 2024; Zhang et al., 2025).

Little work has been done, however, on the problem of *revising* current action plans of the dialogue manager in light of new (incremental) observations. For instance, a robot may start executing

a particular action plan, but suddenly hear a human user say "stop!", in which case the robot ought to interrupt its current sequence of actions and devise an alternative response. This ability to revise or regenerate plans is related to the problem of replanning in robotics and automation (Garrett et al., 2020; Zhou et al., 2023).

The output of the action selection should convey what the system should say or do next, and is often structured as a logical form (Traum and Larsson, 2003; Lison, 2015b). In the case of a verbal response, the NLG module is then responsible for converting this representation into an actual utterance. Alternatively, the dialogue manager may generate a prompt containing natural language instructions on how to respond, and use this prompt as input to a LLM in charge of producing the response.

3.1.3 TURN-TAKING AND END-OF-TURN PREDICTION

Dialogue management in spoken dialogue systems is not only about what to do, but also about when to do it. This question of timing has, unfortunately, not received as much attention as it should have. A common but sub-optimal approach is to wait until the current speaker has stopped speaking for a given period of time, and seek to predict whether they are likely to continue or not (Ferrer et al., 2002). Raux and Eskenazi (2009) presented a finite-state model for turn-taking in spoken dialogue systems, relying on a cost matrix and a decision-theoretic framework to determine whether to take the dialogue floor, release it, wait or keep the floor. Several machine learning models have also been developed to automatically predict when the utterance of the current speak is about to end (De Kok and Heylen, 2009; Maier et al., 2017b). Roddy et al. (2018) presented a data-driven approach to predict a range of turn-taking behaviours when encountering pauses or overlaps, based on on speech-related features. Skantze (2021) and Ohagi et al. (2024) provide a general survey of the various approaches to turn-taking in both embodied and non-embodied speech-based dialogue systems.

Early deep learning approaches to end-of-turn prediction include Maier et al. (2017a), which applied a long short-term memory model to the task using live acoustic features. Such recurrent models are inherently incremental. Using transformer LMs to predict turn-taking is well represented in the recent literature. While not inherently incremental as defined above, turn-taking requires models to handle continuous input. TurnGPT made early use of LMs to detect turn shifts in dialogue (Ekstedt and Skantze, 2020), with some discussion as to how the model could be used to predict end-of-turn. This work was extended in Inoue et al. (2024a), which uses VAP which includes contrastive predictive coding of a cross-attention transformer (as a plus, the model is effective on a CPU). More recently, Chiba and Higashinaka (2025) also applied VAP within an incremental framework (in their case, Remdis (Chiba et al., 2024)) and Roddy and Harte (2020) proposed a model of response timing that is designed for use in incremental systems; human evaluations indicated that they perceived the interaction qualities as more natural when the model was in use. Shukuri et al. (2023) were also concerned with timing and turn-taking and proposed a method for using LMs as meta-controllers of dialogue (where the dialogue system is made up of LMs).

3.2 Incrementalizing Dialogue Management

Buß et al. (2010) introduced an Information State Approach to incremental DM using the IU framework where the IUs themselves composed the information state. In their method, they focused on the collaborative nature of many dialogues in a micro domain of playing a puzzle game. All mod-

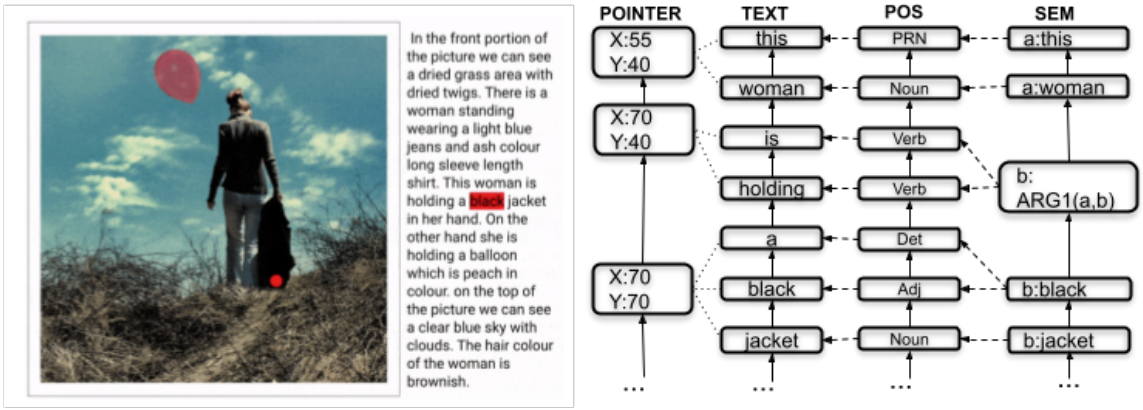

Figure 3: From Kennington and Schlangen (2021), an example of Pointer, Word, POS, and SEM IU annotations for a sample from the Localized Narrative dataset. Solid lines denote SLLs, dashed denote GRINs, and the dotted lines denote an alignment between two modalities. Image taken from https://google.github.io/localized-narratives.

ules, including ASR NLU, TTS, and a floor tracker were modeled at the incremental word level. The incremental DM reacted to information from the NLU, game board state change (i.e., non-linguistic relevant state actions), and the floor tracker. The central element of information was the iQUD (incremental QUD, following Ginzburg (2012)) and is rule-based. They evaluated using an incremental and a non-incremental version of their system and found that the incremental versions were rated higher human-likeness and reactivity by human observers of recorded dialogue of both incremental and non-incremental interactions. This is promising, but limited as a methodology for incremental dialogue.

The same authors followed up this work with Buß and Schlangen (2011) that introduced DIUM—dialogue incremental unit manager—that is also rule-based, but builds on their prior work by leveraging edits that can be made in a dialogue system that is built on the IU-framework. One positive aspect of incremental dialogue is that systems can respond appreciably faster than non-incremental counterparts, but a potential drawback of early response is that the response is based on information which has already been, or is currently being, updated in processing modules. For example, an ASR recognizes *I would like to book a train to Hamm* passes each word to a NLU module that informs the DM with information about which action to take and which object to take the action on. The DM begins to act by looking up train information in a database and informing the NLG about how to respond, and TTS begins to vocalize the response, but at that moment *Hamm* is revoked and replaced with *Hamburg*. This ASR update is propagated to the NLU and likewise to the DM. What action should the DM now take given that TTS is currently uttering something about the wrong city? This is a shortcoming of incremental systems that needs to be addressed according to the authors. Instead of reducing revisions (i.e., waiting for more information) which means waiting longer, and instead of ignoring the problem completely, DIUM offers a third alternative: acknowledge the problem and repair it explicitly. The IU information state is adaptable to addressing the problem directly because a revoke—an important part of the IU-framework—is an abrupt change to the information state that

can be addressed by triggering an explicit repair, for example *Oops, I thought you said Hamm, but it was actually Hamburg. Let me get that information for you.*

Unfortunately, this line of research has not been pursued since the 2011 DIUM paper. However, recently, Kennington and Schlangen (2021) proposed a sketch of using the IU-framework as a method of representing a multimodal, fine-grained information state for use in physical, co-located settings such as HRI. Like Buß and Schlangen (2011), their sketch explained how the information state can consist of the full IU network including connections between IUs, as well as all prior edits. Figure 3 shows an example of a fine-grained, incremental information state using an example from the Localized Narrative Dataset (Pont-Tuset et al., 2020).

Later work explored incremental DM using a *Time Board* where input, output, and decisions made by the DM are posted on the Time Board (Yaghoubzadeh et al., 2015; Yaghoubzadeh and Kopp, 2016). The Time Board is an important piece of incremental DM, according to the authors, because not only does it maintain a history of the ongoing dialogue, future events (e.g., decisions) are also posted and coordinated. Events that have been initiated, for example the system begins an utterance that the NLG is currently constructing and TTS is uttering, can clearly show that they are not yet complete, so a new event that needs to interrupt the ongoing event can produce natural behaviour (e.g., saying *um* or *oops*, or *sorry*).

Going beyond rule-based incremental DM, Selfridge and Arizmendi (2012) introduced a first step towards an incremental POMDP-based system. They proposed an incremental interaction manager (IIM) to mediate communication between an incremental ASR and a partially-observable DM. The IIM worked by evaluating potential DM decisions by applying incremental ASR output to temporary instances of the DM, allowing the system to maintain multiple DM s across time and prune away DM s that are unlikely to advance the dialogue. This enables the partially observable DM to work with incremental ASR n-best lists, but the work demonstrated in Selfridge et al. (2012) has regrettably not been pursued further.

A *barge-in* is when person A begins speaking, then person B attempts to take the floor while person A is still speaking. While often rude, this is common in interactive game scenarios, and it is important for a system that needs to have the ability to stop talking when a human barges in because timing is critical. Selfridge et al. (2013) modeled a simple method for detecting barge-ins, and Pincus and Traum (2017) brought together multiple aspects of incremental dialogue in a word-game task that required fast-paced dialogue where barge-in was required. See Figure 4 for an example. The system had an incremental ASR and learned a policy of when it should handle interruptions made while the system was speaking, and learning when to initiate barge ins. Though the focus was on barge-ins, there are some useful take-always from this work: first, that people often overlap in speech. Second, systems should be ready to yield the floor when they are barged-in on, and they should have the ability to barge in on a human's ongoing speech if there are appropriate stakes involved (e.g., a system needs to inform a human of an impending problem in a nuclear facility). None of these would be possible without incremental processing, and this work shows that timing is an important aspect of the kind of policy a DM needs to learn about.

Manuvinakurike et al. (2017) also looked at incremental dialogue policy learning in a fast-paced game scenario where the user was presented with multiple images and needed to produce a referring expression to that object; the system was tasked with identifying which object the user was referring to. The system could highlight the image that it determined was being referred and say *got it* or it could suggest that the system and user move onto the next set of images (e.g., "let's move onto the next one") because it is unlikely to be able to refer to the correct one given the user's utterance.

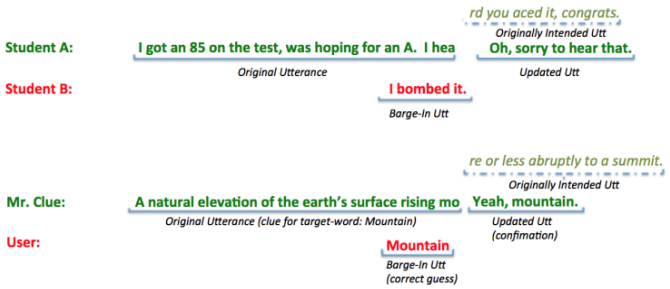

Figure 4: From Pincus and Traum (2017), an example of human-human and game intelligent update dialogues with barge-in.

The learned policy was to either `wait` (i.e., let the user continue speaking), `As-I` (i.e., the system selects what it thinks the referred object is), or `As-S` (i.e., skip to the next set of images). The policy was incremental in that it had to learn at each word increment which action to take. The system and user earned points for identifying images quickly, but it lost points if it referred to an image incorrectly. The system, therefore, had to learn when to wait, select the image, or determine that it was better to move on. The evaluation showed that the learned policy worked better than the hand-coded policy in that it enabled more correctly identified images within a shorter amount of time. Like Pincus and Traum (2017), the focus of this policy revolves around timing of simple actions rather than complex actions, indicating that the purpose of an incremental DM should include handling timing decisions.

Incremental DM in a multi-party HRI setting was reported in Kennington et al. (2014a), that used the IU framework, used an independent OpenDial (Lison and Kennington, 2016) DM for every human that it detected in a game setting. Overall, the DM worked effectively, but it only controlled the dialogue interaction, not robot actions.

Approaches to incremental dialogue state tracking have also been developed. Žilka and Jurčíček (2015) introduced LecTrack, a word-level recurrent neural network state tracker model evaluated on DSTC2 data. The recurrent neural network they used was a LSTM because it is a kind of neural network that can be modeled to work at the word level and maintain its internal state (i.e., update-incremental) at each word increment (see Figure 5). Their evaluations on a subset DTSC2 dataset showed as being on-par with state-of-the-art non-incremental state trackers. There also exists various approaches to dialogue state tracking based on autoregressive LMs (Feng et al., 2023; Hudeček and Dusek, 2023), which rely on instruction-tuned LLMs to extract slot-value pairs from the dialogue history.

4. Discussion

One of the primary challenges of incremental SDS in HRI settings is handling uncertainty including sensory uncertainty and uncertainty that is inherent when communicating with humans. Certainly, all systems are required to handle uncertainty, but the problem is more acute with incremental, multimodal systems in HRI settings because they are tasked with acting on incomplete information that could be forthcoming at a later point.

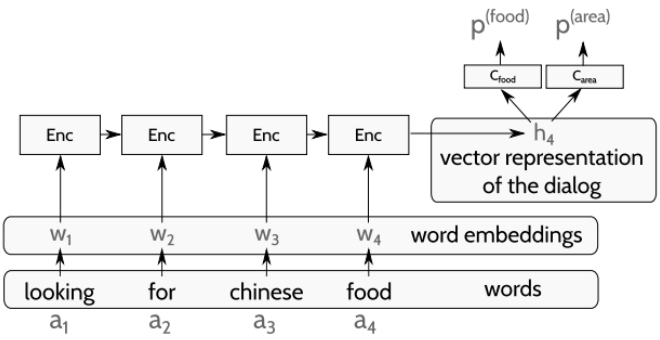

Figure 5: From Žilka and Jurčíček (2015), a schematic of a LSTM-based dialogue state tracker.

Evaluating DM is challenging in general because a proper evaluation usually amounts to a fully-working SDS with human evaluation, but the DM could be working perfectly while the ASR or the NLG modules are not working properly for the task, resulting in poor evaluations from the humans. Offline evaluation is difficult for two reasons: while other modules like ASR, NLU, and even NLG can be evaluated with offline benchmarks, there isn't a clear offline evaluation for DM, though the dialogue state tracking challenge is one attempt to address this. The second difficulty is that there is not a dataset that is annotated for DM at an incremental level. It is therefore unknown whether a DM *should* make a decision at a specific point while a user is speaking, or how to handle errors in decisions when they are in the process of being articulated either in speech or a robotic action. The work on incremental DM explained above (Buß and Schlangen, 2011; Yaghoubzadeh et al., 2015) show methods that attempt to address these challenges, but not with properly annotated incremental data.

4.1 Desiderata

To address these challenges, we offer here desiderata for incremental DM on robotic platforms:

- *Incremental* DM *is responsible for timing*: knowing not just *what* decision to make, but *when* to make that decision are both important in fast-paced, incremental settings, particularly on robotic, embodied platforms where additional modalities play a role in understanding and interaction between user and system. DeVault et al. (2009) explored how learning when to respond to incremental results affects task success, and Kennington and Schlangen (2016) used a rule-based DM to make timing decisions on when to settle on a final decision on which action to take. Recent work relevant to HRI tasks in this area include Zhang et al. (2025) that used a full-duplex DM (i.e., the ASR was always producing input) based on Voice Activity Detection to help determine if the system should wait or act (AudioPaLM can likewise "speak and listen", which is a step in the right direction (Rubenstein et al., 2023), and Yaghoubzadeh et al. (2015)'s model of DM that used a Time Board is a likely good place to start.

- *Incremental* DM *needs to act on incomplete information*: When humans interact with each other, there are often backchannels (e.g., nodding) that signal understanding, or as someone is speaking a listener can signal understanding by taking an action. For example, if a speaker

makes a request *can you hand me the green book on the left?* the listener can already be turning and reaching for a green book before the utterance is complete. A robot that interacts with a person where the task involves handling objects should act in a similar way; for example, reaching for an object or driving towards a destination. If indeed the system made the wrong decision about which object to pursue, then the robot can change its course, but it's important that the robot act as soon as it has enough information to act, even if that act might be incorrect; the movement signals to the user that the robot is in the process of understanding.

- *Incremental* DM *needs to make fast, small decisions concurrently*: Traditional DMs often take in all information from the user during their turn then make a high-level decision once which can then potentially inform multiple modules like NLG to speak and a robot arm to move. An incremental DM, in contrast, needs to make smaller decisions that may lead to a final outcome, but the outcome may not yet be known. This is similar in principle to acting on incomplete information, but the nature of the actions is more fine-grained. For example, the DM may know that the user wants a robot to fetch an object in the kitchen, and though the robot doesn't know which object, it makes a smaller decision to move to the kitchen, and by the time it arrives in the kitchen it knows more about the specifics of the object that it is requested to retrieve.

- *Incremental LLMs*: Transformer LLM architectures generate output incrementally and some aspects of input are incremental, but they are not completely update-incremental. Using them in a restart-incremental manner is computationally expensive, but recent work has shown that minor changes to the model can improve incremental metrics and reduce computational overhead (Kahardipraja et al., 2021). LLMs are being used in many ways in robotics (see, for example, Singh et al. (2023) that uses LLMs for robot action planning), but work needs to be done for incremental processing on LLMs in HRI settings.

Recommendations There is a lack of incremental datasets. Most datasets can be used for incremental training and evaluation for some modules (e.g., ASR or NLG), but NLU and DM modules that produce incremental output that is on a different level of granularity than the word level, so it is unclear from NLU datasets as to *when* a slot should be filled or *when* the DM should make a decision. Efforts towards a dataset that has incremental annotations would be very beneficial to research in the setting of dialogue with robots. Models may or may not need to be trained incrementally, but evaluation metrics should be on the incremental level.

5. Conclusion

In this article, we reviewed literature relating to incremental dialogue management motivated by the need for incremental dialogue management in robotic platforms. We showed that there is ample work in incremental processing, but very little in incremental dialogue management itself. The review resulted in several key desiderata for incremental dialogue management, particularly needed in spoken dialogue-enabled human robot interaction.

Clearly, a decision-making module is a critical component in a robot that can interact with people using spoken dialogue. Taken together, this review in conjunction with other recent review work from Reimann et al. (2024) and Lison and Kennington (2023) are useful for robotics researchers who are interested in designing effective dialogue strategies between robots and humans.

Acknowledgments

We would like to thank the reviewers and editors for their helpful feedback.

References

Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, and Mary Swift. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Pragmatics*, volume 1, pages 149–154, Trento, Italy, 2007.

James F Allen, D Byron, M Dzikovska, G Ferguson, Lucian Galescu, and Amanda Stent. Toward conversational human-computer interaction. *AI Mag.*, 22(4):27–38, December 2001.

James Frederick Allen. *A plan-based approach to speech act recognition*. PhD thesis, 1979.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Vaino Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. Deep speech 2: end-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 173–182, New York, NY, USA, 2016. JMLR.org.

Layla El Asri, Romain Laroche, Olivier Pietquin, and Hatim Khouzaimi. NASTIA: Negotiating appointment setting interface. In *Proceedings of LREC*, pages 266–271, 2014.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv [cs.CL]*, June 2020.

Timo Baumann. Partial representations improve the prosody of incremental speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2932–2936, 2014.

Timo Baumann and David Schlangen. The InproTK 2012 release. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, pages 29–32, 2012.

Timo Baumann, Okko Buß, Michaela Atterer, and David Schlangen. Evaluating the potential utility of ASR N-best lists for incremental spoken dialogue systems. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1031–1034, 2009.

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there. In *Proceedings of the International Workshop Series on Spoken Dialogue Systems Technology (IWSDS) 2016*, 2016.

Niels Beuck and Wolfgang Menzel. Structural prediction in incremental dependency parsing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7816 LNCS, pages 245–257, 2013.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David Schlangen. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 295–303, Seoul, South Korea, July 2012. Association for Computational Linguistics.

Okko Buß and David Schlangen. DIUM - an incremental dialogue manager that can produce self-corrections. In *Proceedings of semdial 2011 (Los Angelogue)*, Proceedings of semdial 2011 (Los Angelogue), 2011.

Okko Buß, Timo Baumann, and David Schlangen. Collaborating on utterances with a spoken dialogue system using an ISU-based approach to incremental dialogue management. In *Proceedings of SIGdial*, pages 233–236, Tokyo, Japan, September 2010.

Michael Cashmore, Andrew Coles, Bence Cserna, Erez Karpas, Daniele Magazzeni, and Wheeler Ruml. Replanning for situated robots. *Proceedings of the International Conference on Automated Planning and Scheduling*, 29(1):665–673, July 2019.

Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefevre, and Olivier Pietquin. User simulation in dialogue systems using inverse reinforcement learning. In *Interspeech 2011*, pages 1025–1028, 2011.

Zhi Chen, Yuncong Liu, Lu Chen, Su Zhu, Mengyue Wu, and Kai Yu. OPAL: Ontology-Aware Pretrained Language Model for End-to-End Task-Oriented Dialogue. *Transactions of the Association for Computational Linguistics*, 11:68–84, 01 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00534. URL https://doi.org/10.1162/tacl_a_00534.

Yuya Chiba and Ryuichiro Higashinaka. Investigating the impact of incremental processing and voice activity projection on spoken dialogue systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3687–3696, 2025.

Yuya Chiba, Koh Mitsuda, Akinobu Lee, and Ryuichiro Higashinaka. The remdis toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models. In *Proceedings of IWSDS*, pages 1–6, 2024.

Phil Cohen. Steps towards collaborative multimodal dialogue (sustained contribution award). In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 4, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450355438. doi: 10.1145/3136755.3154480. URL https://doi.org/10.1145/3136755.3154480.

Philip R Cohen and Lucian Galescu. A planning-based explainable collaborative dialogue system. *arXiv [cs.AI]*, February 2023.

Philip R Cohen and Hector J Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2):213–261, 1990.

Vanya Cohen, Jason Xinyu Liu, Raymond Mooney, Stefanie Tellex, and David Watkins. A survey of robotic language grounding: Tradeoffs between symbols and embeddings. *arXiv [cs.RO]*, May 2024.

Ann Copestake. Semantic composition with (robust) minimal recursion semantics. In *Proceedings of the Workshop on Deep Linguistic Processing - DeepLP '07*, page 73, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Marco Damonte, Shay B Cohen, and Giorgio Satta. An incremental parser for abstract meaning representation. In *European Chapter of the Association for Computational Linguistics (EACL)*, volume 1, pages 536–546, 2017.

Iwan De Kok and Dirk Heylen. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 91–98, 2009.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. Optimising incremental generation for spoken dialogue systems: Reducing the need for fillers. In Barbara Di Eugenio and Susan McRoy, editors, *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 49–58, Utica, IL, May 2012a. Association for Computational Linguistics.

Nina Dethlefs, Verena Rieser, Helen Hastie, and Oliver Lemon. Towards optimising modality allocation for multimodal output generation in incremental dialogue. In *Proceedings of the ECAI Workshop on Machine Learning for Interactive Syustems*, pages 31–36, Montepellier, France, 2012b.

David DeVault and David Traum. A demonstration of incremental speech understanding and confidence estimation in a virtual human dialogue system. In Gary Geunbae Lee, Jonathan Ginzburg, Claire Gardent, and Amanda Stent, editors, *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 131–133, Seoul, South Korea, July 2012. Association for Computational Linguistics.

David DeVault and David Traum. A method for the approximation of incremental understanding of explicit utterance meaning using predictive models in finite domains. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1092–1099, 2013.

David DeVault, Kenji Sagae, and David Traum. Can I finish?: Learning when to respond to incremental interpretation results in interactive dialogue. *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, (September):11–20, 2009.

David Devault, Kenji Sagae, and David Traum. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue and Discourse*, 2(1):143–170, 2011.

Danny Driess, Fei Xia, Mehdi S M Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. *arXiv [cs.LG]*, March 2023.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8):630–645, 2008.

Erik Ekstedt and Gabriel Skantze. TurnGPT: A transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Stroudsburg, PA, USA, November 2020. Association for Computational Linguistics.

Arash Eshghi, Matthew Purver, and Julian Hough. Probabilistic induction for an incremental semantic grammar. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 107–118, Potsdam, Germany, 2013.

Friederike Eyssel and Frank Hegel. (S)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology*, 42(9):2213–2230, 2012.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. Towards LLM-driven dialogue state tracking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755, Stroudsburg, PA, USA, 2023. Association for Computational Linguistics.

Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *Seventh international conference on spoken language processing*, 2002.

Clayton Fields and Casey Kennington. Vision language transformers: A survey. *arXiv [cs.CV]*, July 2023.

Kerstin Fischer, Lakshadeep Naik, Rosalyn M Langedijk, Timo Baumann, Matouš Jelínek, and Oskar Palinko. Initiating human-robot interactions using incremental speech adaptation. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, pages 421–425, New York, NY, USA, March 2021. Association for Computing Machinery.

Caelan Reed Garrett, Chris Paxton, Tomas Lozano-Perez, Leslie Pack Kaelbling, and Dieter Fox. Online replanning in belief space for partially observable task and motion problems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5678–5684. IEEE, 2020.

Fabrizio Ghigi, Maxine Eskenazi, M Ines Torres, and Sungjin Lee. Incremental dialog processing in a task-oriented dialog. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014a.

Fabrizio Ghigi, Maxine Eskenazi, M Ines Torres, and Sungjin Lee. Incremental dialog processing in a task-oriented dialog. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 308–312, Singapore, 2014b.

Jonathan Ginzburg. *The Interactive Stance*. Oxford University Press, 2012.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. *arXiv [cs.CL]*, October 2023.

David Griol, Lluís F Hurtado, Encarna Segarra, and Emilio Sanchis. A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, 50(8-9):666–682, 2008.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL `https://arxiv.org/abs/2411.15594`.

Hany Hassan, Khalil Simaan, and Andy Way. A syntactic language model based on incremental ccg parsing. In *2008 IEEE Workshop on Spoken Language Technology, SLT 2008 - Proceedings*, pages 205–208. Institute of Electrical and Electronics Engineers, 2008.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašic. Trippy: A triple copy strategy for value independent neural dialog state tracking. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 35, 2020.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191, 2020.

Julian Hough and David Schlangen. Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 288–298, Los Angeles, September 2016. Association for Computational Linguistics.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. Incremental semantics for dialogue processing : Requirements , and a comparison of two approaches. In *Proceedings of IWCS*, pages 206–216. Association for Computational Linguistics, 2015.

Vojtěch Hudeček and Ondrej Dusek. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Stroudsburg, PA, USA, 2023. Association for Computational Linguistics.

Kyuyeon Hwang and Wonyong Sung. Character-level incremental speech recognition with recurrent neural networks. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2016-May, pages 5335–5339, January 2016.

Saki Imai, Tahiya Chowdhury, and Amanda J Stent. Evaluating open-source ASR systems: Performance across diverse audio conditions and error correction methods. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5027–5039, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. Real-time and continuous turn-taking prediction using voice activity projection. *arXiv [cs.CL]*, January 2024a.

Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. *arXiv [cs.CL]*, October 2024b.

Youngsoo Jang, Jongmin Lee, Jaeyoung Park, Kyeng Hun Lee, Pierre Lison, and Kee Eung Kim. PyOpenDial: A python-based domain-independent toolkit for developing spoken dialogue systems with probabilistic rules. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations*, pages 187–192, 2020.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. Towards incremental transformers: An empirical analysis of transformer models for incremental NLU. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1178–1189, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. TAPIR: Learning adaptive revision for incremental natural language understanding with a two-pass model. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4173–4197, Stroudsburg, PA, USA, July 2023. Association for Computational Linguistics.

Gaoussou Youssouf Kebe, Luke E Richards, Edward Raff, Francis Ferraro, and Cynthia Matuszek. Bridging the gap: Using deep acoustic representations to learn grounded language from percepts and raw speech. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10884–10893, June 2022.

John D Kelleher and G-J Geert-Jan Kruijff. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL'06)*, pages 1041–1048, 2006.

Casey Kennington and David Schlangen. Situated incremental natural language understanding using markov logic networks. *Computer Speech & Language*, 2014.

Casey Kennington and David Schlangen. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd*

Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 292–301, Beijing, China, 2015. Association for Computational Linguistics.

Casey Kennington and David Schlangen. Supporting spoken assistant systems with a graphical user interface that signals incremental understanding and prediction state. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–251, Los Angeles, September 2016. Association for Computational Linguistics.

Casey Kennington and David Schlangen. A simple generative model of incremental reference resolution in situated dialogue. *Computer Speech & Language*, 2017.

Casey Kennington and David Schlangen. Incremental unit networks for multimodal, fine-grained information state representation. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 89–94, Groningen, Netherlands (Online), June 2021. Association for Computational Linguistics.

Casey Kennington, Kotaro Funakoshi, Yuki Takahashi, and Mikio Nakano. Probabilistic multiparty dialogue management for a game master robot. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*, pages 200–201, Bielefeld, Germany, 2014a. ACM.

Casey Kennington, Spyros Kousidis, and David Schlangen. Situated incremental natural language understanding using a multimodal, linguistically-driven update model. In *Proceedings of CoLing 2014*, pages 1803–1812, 2014b.

Casey Kennington, Ryu Iida, Takenobu Tokunaga, and David Schlangen. Incrementally tracking reference in human / human dialogue using linguistic and extra-linguistic information. In *HLT-NAACL 2015 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference*, pages 272–282, Denver, U.S.A., 2015. Association for Computational Linguistics.

Casey Kennington, Ting Han, and David Schlangen. Temporal alignment using the incremental unit framework. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 297–301, New York, NY, USA, 2017. ACM.

Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. rrSDS: Towards a robot-ready spoken dialogue system. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Virtual, 2020. Association for Computational Linguistics.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. An easy method to make dialogue systems incremental. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 98–107, Philadelphia, PA, U.S.A., 2014. Association for Computational Linguistics.

Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and Stefan Schlangen. Situationally aware in-car information presentation using incremental speech generation: Safer, and more effective. In *Proceedings of the Workshop on Dialogue in Motion (DM), EACL 2014*, pages 68–72, 2014.

Arne Köhn. Incremental natural language processing: Challenges, strategies, and evaluation. In Emily M Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2990–3003, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

Staffan Larsson. *Issue-based Dialogue Management*. PhD thesis, Göteborg University, 2002.

Pierre Lison. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & language*, 34(1):232–255, 2015a.

Pierre Lison. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232–255, 2015b.

Pierre Lison and Casey Kennington. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - System Demonstrations*, 2016.

Pierre Lison and Casey Kennington. Who's in charge? roles and responsibilities of decision-making components in conversational robots. In *Proceedings of the Workshop on Human-Robot Conversational Interaction*, March 2023.

Bing Liu, Gökhan Tür, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, 2018.

Brielen Madureira and David Schlangen. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental NLU. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 357–374, 2020.

Brielen Madureira, Patrick Kahardipraja, and David Schlangen. When only time will tell: Interpreting how transformers process local ambiguities through the lens of restart-incrementality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4722–4749, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. Generative expressive robot behaviors using large language models. *arXiv [cs.RO]*, January 2024.

Angelika Maier, Julian Hough, and David Schlangen. Towards deep end-of-turn prediction for situated spoken dialogue systems. In *Proceedings Interspeech 2017*, pages 1676–1680, 2017a.

Angelika Maier, Julian Hough, and David Schlangen. Towards deep end-of-turn prediction for situated spoken dialogue systems. *Proceedings of INTERSPEECH 2017*, 2017b.

Anna Manaseryan, Porter Rigby, Brooke Matthews, Catherine Henry, Josue Torres-Fonseca, Ryan Whetten, Enoch Levandovsky, and Casey Kennington. rrSDS 2.0: Incremental, modular, distributed, multimodal spoken dialogue with robotic platforms. In *Proceedings of the 26th Annual*

Meeting of the Special Interest Group on Discourse and Dialogue, Avignon, France, 2025. Association for Computational Linguistics.

Ramesh Manuvinakurike, David DeVault, and Kallirroi Georgila. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 331–341, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

Thilo Michael and Sebastian Möller. ReTiCo: An open-source framework for modeling real-time conversations in spoken dialogue systems. In *Tagungsband der 30. Konferenz Elektronische Sprachsignalverarbeitung 2019*, ESSV, pages 134–140, Dresden, March 2019. TUDpress, Dresden.

Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. Real-time emotion generation in human-robot dialogue using large language models. *Front Robot AI*, 10:1271610, December 2023.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, 2017.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155, 2023.

Cheng Niu, Xingguang Wang, Xuxin Cheng, Juntong Song, and Tong Zhang. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8724–8741, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

Joakim Nivre. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553, 2008.

Jekaterina Novikova, Gang Ren, and Leon Watts. It's not the way you look, it's how you move: Validating a general scheme for robot affective behaviour. In Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2015*, pages 239–258, Cham, 2015. Springer International Publishing.

Masaya Ohagi, Tomoya Mizumoto, and Katsumasa Yoshikawa. Investigation of look-ahead techniques to improve response time in spoken dialogue system. In *Interspeech 2024*, pages 3580–3584, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv [cs.CL]*, pages 27730–27744, 2022.

Maike Paetzel, Ramesh Manuvinakurike, and David DeVault. "so, which one is it?" the effect of alternative incremental architectures in a high-performance game-playing agent. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 77–86, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

Qian Pan, Zahra Ashktorab, Michael Desmond, Martín Santillán Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. Human-centered design recommendations for LLM-as-a-judge. In Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H. Andrew Schwartz, editors, *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 16–29, TBD, August 2024. ACL. doi: 10.18653/v1/2024.hucllm-1.2. URL https://aclanthology.org/2024.hucllm-1.2/.

Andreas Peldszus, Okko Buß, Timo Baumann, and David Schlangen. Joint satisfaction of syntactic and pragmatic constraints improves incremental spoken language understanding. In *Proceedings of the 13th EACL*, pages 514–523, Avignon, France, April 2012. Association for Computational Linguistics.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2192, 2018.

Volha Petukhova and Harry Bunt. Incremental dialogue act understanding. *Pragmatics*, pages 235–244, 2011.

Eli Pincus and David Traum. An incremental response policy in an automatic word-game. In *CEUR Workshop Proceedings*, volume 1943, pages 1–8, 2017.

Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. Predicting perceived age: Both language ability and appearance are important. In *Proceedings of SigDial*, 2018.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision – ECCV 2020*, volume 12350 LNCS, pages 647–664. Springer International Publishing, 2020.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. Entity-consistent end-to-end task-oriented dialogue system with kb retriever. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 133–142, 2019.

Libo Qin, Xiao Xu, Lehan Wang, Yue Zhang, and Wanxiang Che. Modularized pre-training for end-to-end task-oriented dialogue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1601–1610, 2023. doi: 10.1109/TASLP.2023.3244503.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

Antoine Raux and Maxine Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 629–637, 2009.

Natalia Reich-Stiebert and Friederike Eyssel. (ir)relevance of gender? on the influence of gender stereotypes on learning with a robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 166–176, New York, NY, USA, 2017. Association for Computing Machinery.

Merle M Reimann, Florian A Kunneman, Catharine Oertel, and Koen V Hindriks. A survey on dialogue management in human-robot interaction. *J. Hum.-Robot Interact.*, March 2024.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786, 2018.

Verena Rieser and Oliver Lemon. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media, 2011.

M Roddy, Gabriel Skantze, and N Harte. Investigating speech features for continuous turn-taking prediction using lstms. In *19th Annual Conference of the International Speech Communication, INTERSPEECH 2018; Hyderabad International Convention Centre (HICC) Hyderabad; India; 2 September 2018 through 6 September 2018*, pages 586–590. International Speech Communication Association, 2018.

Matthew Roddy and Naomi Harte. Neural generation of dialogue response timings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2442–2452, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*, 2020.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. AudioPaLM: A large language model that can speak and listen. *arXiv [cs.CL]*, June 2023.

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

David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. In *Proceedings of EACL*, 2009.

David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111, May 2011.

David Schlangen, Timo Baumann, and Michaela Atterer. Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the 10th SIGdial*, pages 30–37, London, UK, 2009. Association for Computational Linguistics.

David Schlangen, Sina Zarriess, and Casey Kennington. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223, 2016.

Ethan Selfridge and Iker Arizmendi. Integrating incremental speech recognition and pomdp-based dialogue systems. In *In Proceedings of Dialogue and Discourse*, page 275–279, Seoul, South Korea, July 2012. Association for Computational Linguistics.

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of the SIGDIAL 2013 Conference*, pages 384–393, Metz, France, August 2013. Association for Computational Linguistics.

Ethan O Selfridge, Peter A Heeman, Iker Arizmendi, and Jason D Williams. Demonstrating the incremental interaction manager in an end-toend "lets go!" dialogue system,". In *Proc. of IEEE Workshop on Spoken Language Technology*, 2012.

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gökhan Tür. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, 2018.

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. Challenging neural dialogue models with natural data: Memory networks fail on incremental phenomena. In *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, ISCA, August 2017. ISCA.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y K Li, Y Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv [cs.CL]*, February 2024.

Prashanth Gurunath Shivakumar, Naveen Kumar, Panayiotis Georgiou, and Shrikanth Narayanan. Incremental online spoken language understanding. October 2019.

Kotaro Shukuri, Ryoma Ishigaki, Jundai Suzuki, Tsubasa Naganuma, Takuma Fujimoto, Daisuke Kawakubo, Masaki Shuzo, and Eisaku Maeda. Meta-control of dialogue systems using large language models. *arXiv [cs.RO]*, December 2023.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, May 2023.

Ishika Singh, David Traum, and Jesse Thomason. TwoStep: Multi-agent task planning using classical planners and large language models. *arXiv [cs.AI]*, March 2024.

Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178, 2021.

Pei-Hao Su, Milica Gasic, Nikola Mrkšić, Lina M Rojas Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441, 2016.

Pei-Hao Su, Milica Gašić, and Steve Young. Reward estimation for dialogue policy optimisation. *Computer Speech & Language*, 51:24–43, 2018.

Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 100–110, 2019.

Michael K Tanenhaus and Michael J Spivey-Knowlton. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632, 1995.

Blaise Thomson and Steve Young. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588, 2010.

David Traum and Staffan Larsson. The information state approach to dialogue management. In *Current and New Directions in Discourse and Dialogue*, pages 325–353. Springer, 2003.

Stefan Ultes, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Inigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gasic, et al. Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, 2017.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yen-Chen Wu, Steve Young, and Milica Gašić. Addressing objects and their relations: The conversational entity dialogue model. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 273–283, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5032. URL `https://aclanthology.org/W18-5032`.

Herwin Van Welbergen, Dennis Reidsma, and Stefan Kopp. An incremental multimodal realizer for behavior co-articulation and coordination. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7502 LNAI, pages 175–188, 2012.

Nicolas Wagner and Stefan Ultes. On the controllability of large language models for dialogue interaction. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–221, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

Nicholas Thomas Walker, Stefan Ultes, and Pierre Lison. Graphwoz: Dialogue management with conversational knowledge graphs. *arXiv preprint arXiv:2211.12852*, 2022.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. ONE-PEACE: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305. 11172*, 2023.

Ryan Whetten, Enoch Levandovsky, Mir Tahsin Imtiaz, and Casey Kennington. Evaluating automatic speech recognition and natural language understanding in an incremental setting. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Maribor, Slovenia, August 2023. SEMDIAL.

Jason D Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016.

Jason D Williams, Kavosh Asadi Atui, and Geoffrey Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–677, 2017.

Tom Williams, Cynthia Matuszek, Ross Mead, and Nick Depalma. Scarecrows in oz: The use of large language models in HRI. *J. Hum.-Robot Interact.*, 13(1):1–11, January 2024.

Ramin Yaghoubzadeh and Stefan Kopp. flexdiam – flexible dialogue management for problem-aware, incremental spoken interaction for all user groups (demo paper). *Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2016)*, 2016.

Ramin Yaghoubzadeh, Karola Pitsch, and Stefan Kopp. Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users. In *Intelligent Virtual Agents*, Lecture notes in computer science, pages 28–38. Springer International Publishing, Cham, 2015.

Takashi Yamauchi, Mikio Nakano, and Kotaro Funakoshi. A robotic agent in a virtual environment that performs situated incremental understanding of navigational utterances. In *SIGdial 2013*, pages 369–371, Metz, France, August 2013. Association for Computational Linguistics.

Shiquan Yang, Rui Zhang, and Sarah Erfani. Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888, 2020.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. POMDP-based statistical spoken dialog systems: A review. *Proc. IEEE*, 101(5):1160–1179, 2013.

Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629, 2022.

Sina Zarrieß and David Schlangen. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

Hao Zhang, Weiwei Li, Rilin Chen, Vinay Kothapally, Meng Yu, and Dong Yu. LLM-enhanced dialogue management for full-duplex spoken dialogue systems. *arXiv [cs.CL]*, February 2025.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9207–9219, 2020a.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10): 2011–2027, 2020b.

Siyuan Zhou, Yilun Du, Shun Zhang, Mengdi Xu, Yikang Shen, Wei Xiao, Dit-Yan Yeung, and Chuang Gan. Adaptive Online Replanning with Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lukáš Žilka and Filip Jurčíček. LecTrack: Incremental dialog state tracking with long short-term memory networks. In *Text, Speech, and Dialogue*, pages 174–182. Springer International Publishing, 2015.