Investigating Proactivity in Multimodal Task-Guidance Dialogues

Sofia Brenna[1,2], Elisabetta Jezek[3], Matthias Kraus[4], Bernardo Magnini[1]
[1]Fondazione Bruno Kessler, [2]Free University of Bolzano, [3]University of Pavia, [4] Augsburg University
sbrenna@fbk.eu, elisabetta.jezek@unipv.it, matthias.kraus@uni-a.de, magnini@fbk.eu

Abstract

While proactivity, i.e., the ability to take the initiative and anticipate requests in order to improve the effectiveness of a conversation, has been traditionally investigated in task-oriented dialogues (e.g., booking a restaurant), less work addresses proactive behaviours in task-guidance dialogues (e.g., guide to execute recipes), where the expert instructor is supposed to interact and supervise a user in a real-world setting. In the paper, we analyse a corpus of video-recorded task-guided dialogues and explore two key features of proactivity in this context: (i) the role of multimodal features, with respect to chat-based dialogues; (ii) the role of instructions and actions grounded in a real situation. Through a comparison between task-oriented and task-guidance annotated dialogues, we find that task-guided dialogues are highly proactive interactions, where preventing mistakes and maintaining the correct process order is essential for achieving the dialogue goal. In addition, the video information available in the task-guidance setting can be corrective for false positive proactive behaviours, although without introducing substantial differences. To support our analysis and to foster further research we provide a corpus of multimodal task-guidance dialogues annotated according to proactivity.

Keywords: proactivity, task-guidance dialogues, multimodality

1. Introduction

Recent performance of generative AI models raises interest in systems that are able to supervise humans during the execution of a wide variety of tasks (e.g., repair a device, execute a recipe). In the context of a *task-guidance dialogue*, an AI agent is expected not only to know and describe task instructions one step at a time, but also to answer intermediate questions, clarify requested details, provide feedback and be vigilant enough to check for any user errors and correct them proactively (see, for instance, (Aggarwal et al., 2025)). In order to design effective models of task-guidance conversations, proactive behaviours are crucial to ensure the collaboration between the expert supervisor and the user, as this is necessary for the success of the task to be carried out.

While proactivity is receiving increasing attention and there are several attempts to computationally model proactive behaviours (Balaraman and Magnini, 2020)(Deng et al., 2023) (Liao et al., 2023) (Brenna and Magnini, 2024), this interest is mostly limited to traditional task-oriented dialogues (e.g., booking a restaurant) and much less research explores the role of proactivity in task-guidance conversations. In the paper, we build on top of recent studies and linguistic resources on proactivity in dialogues (i.e., the D-Pro (Brenna et al., 2025) and the ProDial (Kraus et al., 2022a) annotated corpora) and extend previous analysis to the richer context of task-guidance dialogues.

By focusing on proactivity beyond strictly task-oriented dialogues, we aim to investigate two main research questions: (i) To what extent previous definitions of proactivity and previous annotation schemas can be adapted to the richer context of multimodal task-guidance dialogues? (ii) What is the role of video-recorded information in multimodal task-guidance dialogues as far as proactivity is concerned? More specifically, how does the annotation of proactivity change with access to multimodality (audio-video)?

In order to investigate the two research questions, we considered the Watch-Talk-and-Guide (WTaG) (Bao et al., 2023) corpus, a collection of video recorded situated task-guidance interactions between a human User and a human Instructor in the recipes domain. We performed a proactivity annotation on a subset of the WTaG data (both text-only and multimodal) and conducted a systematic comparison with annotated task-oriented dialogues in the D-Pro collection (Brenna et al., 2025). As a result of the comparison, we found that task-guidance dialogues include a significant higher proportion of proactive turns.

The innovative contributions of the paper are the following:

- We extend the D-Pro corpus's annotation schema for proactivity proposed in (Brenna et al., 2025) to account for a wider variety of multimodal task-guidance phenomena, beyond traditional task-specific settings, resulting in a more general and portable schema.
- We provide a new set of annotated multimodal task-guidance dialogues, which can be used to foster further research[1].

[1]Annotated dialogues available at the follow-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 64–76
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

- We find that, while the visual component (i.e., recorded videos) can help correct some false positive proactive behaviour annotations, it does not introduce substantial differences in the assessment of proactivity compared to the speech component of the dialogue.

The structure of the paper is the following. After defining proactivity and reporting previous work on its annotation in task-oriented dialogues, we report the results of both the text-only and the multimodal annotation experiments we performed on the task-guidance dialogues included in the WTaG dataset, illustrating the challenges they raise for the annotation of proactivity compared to the task-oriented ones of the D-PRO corpus. We conclude with a discussion of the findings and plans for further research.

2. Background

2.1. Defining proactivity

Collaborative behaviour in dialogue refers to participants' various actions and strategies to work together towards effective communication, shared understanding, and attainment of conversational goals.

Among the most prominent linguistic strategies, we find *proactivity*. Derived from the definition of proactivity in organisational behaviours (Grant and Ashford, 2008), the term has been used in NLP since at least (Li et al., 2016) to refer to conversational agents' ability to create or control the conversation by taking the initiative and anticipating the impacts on themselves or human users, rather than passively responding to the user's request (see (Deng et al., 2023) for an overview).

From a linguistic perspective, the first systematic attempt to define the rules governing human behaviour in communication can be traced back to Grice's *cooperative principle*, spelled out in the *maxims of conversation* (Grice, 1975). The work of Austin (Austin, 1962) and Searle (Searle, 1969, 1975) integrated Grice's contribution by identifying a typology of *speech acts* and by examining their application condition in detail. Their proposal has been taken up in NLP under the label of *dialogue act* (Stolcke et al., 2000), *conversation act* and *intent* (Bunt et al., 2010; Bunt and Girard, 2005; Bunt, 2006; Traum and Hinkelman, 1992).

In other works originating from the social psychology of language, the concept of *accommodation* has been put forth, which offers an additional theoretical framework for analysing *proactivity* in NLP. *Accommodation* is the process of modifying one's communication style, vocabulary, code, and tone (including politeness, (Brown and Levinson, 1987; Bargiela-Chiappini, 2003)) to better align with a conversation partner (cf. *speech accommodation theory* (Giles et al., 1973; Giles, 1979; Giles et al., 1991; Giles and Powesland, 1997; Burt, 1994; Scotton, 1988) and *communication accommodation theory* (Giles and Ogay, 2006)). This adaptation facilitates understanding, promotes effective collaboration, and fosters a positive interactional atmosphere.[2]

In NLP, a significant body of work has also been dedicated to the concept of *initiative* in a dialogue (Traum, 1997), This concept is closely connected to proactivity, since being proactive inherently requires taking the initiative to anticipate future needs, rather than responding reactively. Initiative has been studied in several dialogue types, for example in task-oriented and advisory dialogues (Whittaker and Stenton, 1988; Walker and Whittaker, 1990). and in negotiation dialogues (Nouri and Traum, 2014). (Walker and Whittaker, 1990) equate initiative to control, associate four utterance types with the allocation of control to participants, and identify types of control shift. (Chu-Carroll and Brown, 1999), followed by (Kersey et al., 2009) and others, distinguish between *dialogue initiative* and *task initiative*: dialogue initiative is held by the participant guiding the conversation, while task initiative belongs to the one leading goal planning. This distinction separates the two types of initiative, aligning with (Jordan and Di Eugenio, 1997), who refute (Walker and Whittaker, 1990) arguing that control pertains to the dialogue level, whereas initiative relates to the problem-solving level. Finally, focusing on the turn level, (Nouri and Traum, 2014) distinguish between two aspects of initiative: establishing new discourse obligations and providing unsolicited material.

In light of the current understanding of the notion of proactivity in the NLP community, in the context of task-oriented dialogues, (Brenna et al., 2025) proposed to consider an utterance as proactive when one of the participants does not act merely in response to the requests the other participant has made, so the behaviour is self-prompted and not simply reactive. Moreover, the participant has a long-term, goal-directed behaviour that predicts future states and needs, so his/her proactive behaviour is somehow effective for the achievement of the dialogue goal.

ing address: https://github.com/sofiabrenna/proactive-WTaG

[2] Accommodation has been investigated in NLP in relation to the design of spoken dialogue systems (cf. vocal accommodation in (Raveh, 2021) and prosodic accommodation in (De Looze et al., 2014)).

2.2. Task-Guidance Dialogues

Dialogue research commonly distinguishes between open-ended "chit-chat" and task-oriented interaction (Mctear, 2020). Chit-chat systems are designed for unbounded social exchange, whereas task-oriented interaction are optimized to achieve a concrete user goal under domain constraints. Interactions with task-oriented systems are typically transactional, either delivering question answering (Guo et al., 2018) or form/slot-filling interactions for information access (Williams et al., 2016). In slot-filling, the system incrementally elicits values for predefined dialogue slots, i.e., attribute-value pairs, such as food type, price range, or location in the restaurant search domain. These slots are then used to identify the user's goal and query a database to provide the required information.

Within task-oriented dialogues, task-guidance is a special case. Here, the system (or human instructor) actively helps the user carry out a multi-step procedure in real-time, e.g. assembling furniture (Bohus et al., 2024), cooking a recipe (Bao et al., 2023), fitness coaching (Panchal et al., 2024). In contrast to task-oriented dialogue, they thus target execution-centric rather than transaction-centric interaction. In this regard, task-guidance dialogues are procedural (they unfold as sequences of interdependent steps), temporal (timing and durations matter, e.g. "stir for 20 seconds" in the cooking domain), and most importantly situated (grounded in the spatial and temporal context of the user's environment and actions (Brooks, 2018; Ammanabrolu et al., 2022)). Particular challenges for the design and implementation of systems for task-guidance, are to model where the user currently is in the plan, what could go wrong, and how to intervene proactively. Such systems must also handle unstructured, multi-modal inputs (audio, video) and produce instructions that are interpretable, well-timed, and corrective when needed.

To achieve this, data-driven methods based on large dialogue corpora are applied. However, popular large-scale data sets for building typical task-oriented systems, such as MultiWoz (Budzianowski et al., 2018), and open-domain systems, such as Twitter (Ritter et al., 2010) or the Ubuntu dataset (Lowe et al., 2015a), are insufficient, either because of their unboundedness (open-domain, chit-chat) or their transactional nature (task-oriented). For this reason, several resources focusing on task-guidance dialogues have been published. Examples for this are the WTaG (Bao et al., 2023), HoloAssist (Wang et al., 2023), and QVED (Panchal et al., 2024) corpora. These focus on interactive assistants for providing support and guidance in domains such as cooking, object manipulation or fitness. Due to their situated and often instructional nature, these corpora are particularly interesting for investigating proactive dialogue as they naturally elicit anticipatory, goal-directed behaviour, aligning well with current definitions of proactive action. For this reason, in this work, we focus our investigations on the WTaG dataset.

2.3. Annotation of Proactivity in Task-Oriented Dialogues

Two existing annotation schemes designed for annotating proactivity in task-oriented dialogues are particularly relevant for our research on proactivity in the setting of task-guidance dialogues.

The first is the D-Pro annotation schema proposed in (Brenna et al., 2025), where proactivity is categorized on the basis of the following utterance's dialogue act:

- INFORM — the participant is providing some information.
- OFFER — the participant is proposing to do something or to provide some further information.
- SUGGEST — the participant is suggesting that the addressee should do something.
- REQUEST — the participant is demanding that the addressee do something or that the addressee provide some information.
- INSTRUCT — the participant is providing the addressee with instructions to follow.

The second is the ProDial schema proposed in (Kraus et al., 2022a), which interestingly defines four levels of proactivity, ordered by *proactivity intensity*—i.e., the degree of initiative taken by the agent. These levels can be understood as a dialogue-oriented mapping from classic *levels of autonomy* (Sheridan and Verplank, 1978):

- NONE — No unsolicited actions or information; the agent is purely reactive.
- NOTIFICATION — Unprompted surfacing of relevant facts, events, or status updates.
- SUGGESTION — Proposing next steps, alternatives, or plans for user consideration.
- INTERVENTION — Directly performing or enforcing actions on the user's behalf.

The last category presupposes the ability to act on the environment and is therefore *not applicable* to remote, purely conversational settings (both information-seeking and instructional), where the agent cannot manipulate the physical world.

3. Annotating Proactivity in Multimodal Task-Guidance Dialogues

To investigate proactivity in task-guidance dialogue, we select the WTaG dataset. The Watch, Talk,

and Guide (WTaG) dataset (Bao et al., 2023) is a multimodal dialogue collection for situated task-guidance based on natural interactions between a human User and a human Instructor in the recipes domain. The dataset is released as audio-visual recordings, combined with textual transcriptions of the dialogues, enriched with dialogue intents and mistakes annotations and with metadata, such as timestamps and information detected about the current recipe step, called Step Detection (e.g. *Roll tortilla*.)

As referenced above (Section 2.2), task-oriented dialogues and task-guidance dialogues differ in many respects, above all in that task-guidance dialogue is procedural, temporal, and situational. In addition to this, the instructional nature of task-guidance dialogues often grants a certain degree of autonomy to the user: once the instructions are clear and understood, a skilled user can proceed independently without further (or with very limited) dialogical interaction. In task-oriented dialogues, instead, progress towards the goal is impossible without interaction. In such dialogues, the interaction is not merely a support, but it is constitutive of the task performance—each conversational exchange provides the necessary input for the next procedural step. This reflects on: (i) the amount of details and information shared between participants; (ii) the lexical variety and richness of linguistic features; (iii) the type of dialogue acts that are realised in the interactions.

Based on these observations, in order to annotate proactivity in the new domain of WTaG dialogues, we find the need to introduce some refinements to the schemes introduced in the previous section. Particularly, although there is a certain degree of correspondence between INFORM and NOTIFICATION on the one side, and SUGGEST and SUGGESTION on the other side, we find the need to introduce REQUESTs in the action space of task-guidance dialogues, and especially the need to further differentiate between an INFORMATION-REQUEST and an ACTION-REQUEST.

Based on Bunt (2009) and Bunt et al. (2020)'s ISO standard taxonomy, developed for annotating dialogue with semantic information, we select and adapt some *General-Purpose Communicative Functions* to categorise proactivity based on utterance's dialogue acts. The following are the labels selected for the annotation of proactivity in task-guidance dialogue. The definitions here reported, for the most part inspired by Bunt (2009) DIT++ Taxonomy[3], are the same as those included in the annotation guidelines provided to the annotators.

- NONE — participant A does not produce an utterance that is simultaneously unsolicited and helpful: the utterance is not proactive.
- NOTIFICATION — participant A gives some information to addressee B, A makes B aware of some information.
- SUGGESTION — A suggests an action or an option that they believe to be potentially promising for achieving a certain goal.
- OFFER — A offers to do something in order to help B or to achieve the dialogue goal.
- INFO-REQUEST — A is asking for some information or instruction from B.
- ACTION-REQUEST — A wants B to perform the requested action, conditional on B's consent.
- INSTRUCTION — A wants B to perform an action and/or is instructing B on how to do it.
- INTERVENTION — A directly performs an action in the environment to help achieve the dialogue goal.

Here is an example of proactive utterance taken from WTaG dialogues for each label:

NOTIFICATION:
IN: oh you should use a butter knife to scoop nut butter from the jar
IN: it's on the right with the red lid
US: oh oh yes

SUGGESTION:
US: do I put the wet ingredients in the same bowl?
IN: um yes
IN: try to whisk it before adding the wet ingredients so that they're combined

OFFER:
IN: so bubbles should start to form
US: no bubbles umm
US: should I boil it a little more?

INFO-REQUEST:
IN: oh now that the water is boiled you wanna check the temperature first so you can use the thermometer
US: ohh
US: like sticking into the water?

ACTION-REQUEST:
US: ah there we go
IN: great
IN: ok now you can put it on a plate

INSTRUCTION:
US: with that it's the appropriate thickness
IN: yeah that's good
IN: and you should roll it tight enough so to prevent gaps but not so tight where the filling leaks.

[3]Resource available here: https://dit.uvt.nl/dit4/

As noted in Section 2.3, INTERVENTION acts are not applicable to remote settings. Nonetheless, we retain this label in our annotation schema to ensure generalisability across diverse human-computer interaction (HCI) scenarios.

With regard to the INFORMATION-REQUEST and ACTION-REQUEST, it is important to note that, although both are designated as "requests," they differ in their functional classification. In fact, INFORMATION-REQUEST does not constitute a request in the strict sense; rather, it belongs to the category of Information-Seeking Functions in DIT++ Taxonomy. By contrast, ACTION-REQUEST falls within the Directives category and is therefore more closely related to the INSTRUCTION label. Indeed, the primary distinction between ACTION-REQUEST and INSTRUCTION lies in the degree of directive force: ACTION-REQUEST presupposes the addressee's consent or willingness to carry out the requested action, whereas INSTRUCTION conveys a higher degree of directive force and authority.

In task-oriented dialogues, one participant assumes the role of Instructor, responsible for guiding the interaction and directing the user toward the successful completion of the dialogue goal. When this role is embodied by an AI agent, as is often the case in human-computer interaction contexts, the way requests are formulated acquires particular importance. From the perspective of an agent that must select an appropriate interaction strategy, it is crucial to distinguish between soliciting additional information from the user, and requesting that the user perform an action. The latter may be acknowledged as a command and, consequently, may be perceived negatively by human users. Accordingly, how requests are framed plays a central role in maintaining a collaborative and user-aligned interaction.

3.1. Annotation Phases

The annotation experiment on the WTaG dataset consists in the manual labelling of eleven task-guidance dialogues, randomly chosen from the WTaG dataset, performed by two expert annotators. The experiment follows a 2-steps annotation process and is performed at the utterance level in the dialogue. First, the annotation is performed with access to dialogue transcriptions only, then the process is repeated with full access to the multimodal context, so with audio-video recordings in addition to textual dialogues. Step Detection turns are kept in the dialogues in both textual and multimodal annotation, in order to give annotators a general idea of the events occurring in the recipe process (textual) and to keep track of them (multimodal), but are discarded when computing annotation results. Expert annotators did not rely on prior familiarity with WTaG recipes, which could have biased text-only annotations. Since the recipes required a precise sequence of steps, the focus in WTaG dialogues was on following the procedure rather than accomplishing the final dish, which, although simple, reflected a cultural context (American) different from that of the annotators.

Inter-Annotator Agreement To validate the reliability of the annotations manually conducted on part of WTaG, two expert annotators independently labelled the data with full multimodal access (audio–video recordings and dialogue transcriptions). After a training phase and a pilot round to refine the annotation guidelines, inter-annotator agreement was computed using Cohen's *kappa* (Pustejovsky and Stubbs, 2012) —which accounts for agreement occurring by chance —on 25% of all annotated dialogue turns. The Cohen's *kappa* score for the annotation of proactive utterances is k = 0.74 and for the categorisation of dialogue acts is k = 0.71[4], indicating substantial agreement between annotators (Landis and Koch, 1977). With respect to dialogue act labels, there were no disagreements between INFORMATION-REQUEST and ACTION-REQUEST, nor, perhaps surprisingly, between ACTION-REQUEST and INSTRUCTION labels. On the other hand, a few instances of disagreement involved the SUGGESTION label, as in the example below:

U60 IN: there is also a glass bowl next to you, you can use that
U61 IN: to help you
U62 US: oh
U63 IN: you might want to zero out the weight of that glass bowl though

One annotator categorized underlined utterance U63 as a SUGGESTION, while the other classified it as an ACTION-REQUEST.

Exploratory Experiment on Non-Expert Multimodal Annotation To investigate the effect of multimodality on dialogue proactivity annotation, we conducted an early-stage experimental trial with a non-expert annotator on a small subset of two dialogues. The annotator received minimal training and was asked to label the dialogues first in a text-only setting, and subsequently in a full multimodal setting. Text-only annotation resulted in an inter-annotator agreement of k = 0.606, whereas multimodal annotation achieved k = 0.663. These scores indicate moderate agreement, with slightly higher agreement when both audio-visual and textual context were available.

[4]Both values computed after removal of the Step Detection turns.

4. Data and Experiments

In this section, we describe the annotation of proactivity in the WTaG dataset and analyse its distribution across task-guidance dialogues. The annotated data are then compared with existing statistics from the D-Pro Corpus, for a cross-domain investigation of proactive behaviour in task-oriented interactions.

Annotated WTaG Dataset To ensure an adequate and directly comparable amount of data to the datasets already annotated for proactivity in the D-Pro Corpus, we annotated a similar number of dialogue turns to the average D-Pro sub-corpus, enabling turn-level comparability. Specifically, the annotated WTaG dataset contains 599 turns, slightly more than the D-Pro sub-corpus average of 571 turns.

Metric	Annotated WTaG	D-Pro Micro Avg.
# Dialogues	11	30.2
# Turns	599	571
# StepDet.	123	–
# Utt.	943	1205.6
# Instructor Utt.	387	944.3*
# User Utt.	556	735.7*
# Tokens	4823	8262
# Types	551	1427.8
# TTR	11.4	15.3
Avg. Turns per Dial.	54.45	18.90
St. Dev. Turns per Dial.	16.31	14.79
Avg. Utt. per Dial.	85.73	39.92
Avg. Utt. per Turn	1.57	2.11

Table 1: Dataset composition and dialogue statistics in annotated WTaG dialogues. Starred values are excluding one sub-corpus from D-Pro, since D-Pro WhatsApp does not have User/Instructor role distinction. All WTaG dialogue data are computed after removal of Step Detection turns.

Information on the composition of the annotated WTaG dialogues are reported in Table 1, with a comparison to micro averaged data in D-Pro. WTaG dialogues contain, on average, a higher number of turns and utterances; on the other hand, the length of each utterance and turn is shorter, since overall tokens and types in WTaG annotated dialogues are about half of D-Pro's average token and types. The Type-Token Ratio (TTR) score indicates that language variety in WTaG is limited: this reflects the dataset's multimodal, task-guidance nature, where much of the dialogue is devoted to grounding acts and to giving confirmation of events happening in the scene, thus leaving little room for more varied or elaborated language use.

4.1. Experiments on WTaG

In this section, we present the annotation process and experimental analysis of proactivity in WTaG dialogues. We first describe the textual annotation phase and the distribution of proactive utterances and dialogue acts, followed by the multimodal annotation phase and a comparison of results between the two settings.

Textual Annotation The first annotation phase is conducted on WTaG dialogue transcriptions provided with the original dataset, after minimal preprocessing to enable a human-friendly visualization of the dialogues. The annotation is made at the utterance level in two steps: (i) the annotator decides whether to mark an utterance U_n as proactive or not based on the two criteria of initiative and helpfulness; (ii) the annotator categorises every utterance marked as proactive into one of the 7 categories from the annotation schema (see Section 3). As shown in Table 2, a total of 227 utterances—approximately 24% of all utterances in the annotated dialogues—are marked as proactive. Out of these, 26 instances of proactive action are employed to prevent a wrong step or action from slowing down the task progress. Moving to turn-level results, 196 out of 599 annotated turns (over 32%) contain at least one proactive utterance. It is common for a turn to include only a single proactive utterance, as indicated by the Pro Utterance per Pro Turn value of 1.18. This suggests that while proactive behaviour is regularly employed in the dialogues, it usually occurs in single, isolated utterance, rather than in extended proactive exchanges.

As far as categorisation of proactive dialogue acts is concerned, statistics are reported in the bar chart: Figure 1 shows that INSTRUCTION acts are the most frequent among proactive utterances, with 77 instances, followed by INFO-REQUEST acts with 71 occurrences. This suggests that proactivity is a recurrent feature of instructional dialogue,

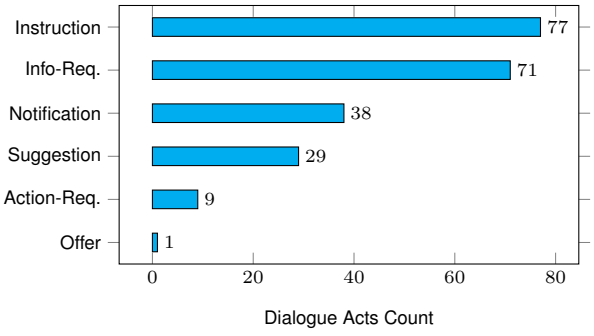

Figure 1: Frequency of dialogue acts in proactive utterances in annotated WTaG dialogues.

Metric	Textual WTaG	Multimodal WTaG	D-Pro Micro Average
#Pro Turns	196	194	-
#Pro Utt.	227	223	-
#Instructor Pro Utt.	142	143	-
#User Pro Utt.	85	80	-
Pro Utt. %	24.07%	23.65%	15.31%
Pro Turns %	32.72%	32.39%	19.54%
Pro Utt. per Pro Turn	1.16	1.15	1.65

Table 2: Comparison of proactivity statistics between textual and multimodal WTaG annotations, with D-Pro micro average values.

primarily oriented toward the task itself — manifesting either in directly guiding the addressee through task execution or in requesting instructions from the interlocutor. Less direct guidance interventions, such as NOTIFICATION and SUGGESTION acts, occur 38 and 29 times respectively, and are also well represented. Forms of guidance that allow greater freedom in choosing subsequent actions, such as ACTION-REQUEST acts, are comparatively rare (9 occurrences), as the recipe domain demands precision in both actions and timing — a requirement more effectively fulfilled through INSTRUCTION acts. The sole OFFER instance in the dialogues takes the form of the User's indirect and polite request for confirmation of the instructions for the current step of the recipe, anticipating the Instructor's next guidance move. Figure 1 excludes the null label — NONE — corresponding to non-proactive utterances (839 instances), as well as the INTERVENTION label, which does not apply to the remote interaction setting of WTaG dialogues.

Multimodal Annotation The second annotation phase is carried out with access to the complete multimodal context, comprising the dialogue transcriptions and the associated audio-visual recordings. As shown in Table 2, only minimal changes occurred in the multimodal annotation process compared to the text-only setting. The percentage of proactive turns over total dialogue turns remains almost unchanged (two occurrences), and the number of proactive utterances as well shows a very slight decrease (four occurrences). Overall, access to the extended, multimodal context enables more accurate annotation, and it allows corrections for errors or omissions in the textual transcriptions that had previously misled judgments about the novelty of utterance content in the text-only annotation phase — resulting in a couple of utterances being incorrectly labelled for proactivity. An example of this can be found in Figure 2, where the User's utterance U17 is erroneously marked as proactive, since it is actually preceded by the Instructor's proactive utterance U16, and since it eventually consists in a reactive confirmation request. At the same time, while annotating with access to text only, an expert annotator who has previously seen videos of the same recipes dataset may be able to infer what is happening in the multimodal context, suggesting that additional context is not always strictly necessary to understand the events and annotate accordingly. Nevertheless, the multimodal context serves as a valuable resource rather than a distraction. Early annotation experiments (see 3.1) demonstrated the usefulness of multimodal context in helping non-expert annotators better understand dialogue dynamics.

4.2. Comparison with D-Pro Corpus

The dataset for the experimental comparison is D-Pro, a corpus enriched with manually curated proactivity-oriented annotations. D-Pro comprises 151 dialogues from 5 task-oriented dialogue sub-corpora, namely, Italian Whatsapp Corpus (Hewett, 2017), the Italian Nespole! Corpus (Burger et al., 2001; Mana et al., 2003), Jilda (Sucameli et al., 2020, 2021), the Italian Ubuntu Chat Corpus (Lowe et al., 2015b), and MultiWOZ 2.2 (Zang et al., 2020). Most of the dialogues are in Italian, with the only exception of the MultiWOZ 2.2 dialogues and some dialogues from the Italian Whatsapp Corpus, due to code mixing and code switching employed by the speakers. With WTaG annotated to a size comparable to each D-Pro sub-corpus (600 turns), it becomes interesting to compare proactivity in task-guidance recipe dialogues with proactivity in a task-oriented setting across different domains (Table 2).

The first interesting finding is that instructional WTaG dialogues exhibit higher levels of proactivity compared to D-Pro dialogues: 24% of utterances in WTaG are proactive, compared to an average of 15% in D-Pro, and 32% of turns are proactive, versus 20% in D-Pro. Among the D-Pro sub-corpora, only the WhatsApp subset holds higher values, with 25% of utterances and nearly 36% of turns marked as proactive. Additionally,

T7	U12	PRO	INSTRUCT	In: right so you should be using the peanut butter	PRO	INSTRUCT	In: right so you should be using the peanut butter
T8	U13			Us: peanut butter			Us: peanut butter
T9	U14			In: yeah			In: yeah
T10	U15			Us: ok			Us: ok
T11	U16			In: with a	PRO	INSTRUCT	In: with a knife
T12	U17	PRO	INFO-REQ	Us: with a knife?			Us: with a knife?
T13	U18			In: yeah			In: yeah

Figure 2: Example of differences in annotation due to omissions in the dialogue transcripts. Textual annotation on the left, multimodal annotation on the right.

this is consistent with the observation that proactive turns in D-Pro more frequently encompass multiple proactive utterances compared to those in WTaG: D-Pro dialogues have 1.65 proactive utterances per proactive turn, compared to only 1.16 in WTaG. This indicates that in D-Pro dialogues proactive contributions are often expressed through more extended sequences of utterances, while in WTaG proactivity is more likely to occur as isolated, single-utterance interventions within a turn. This is consistent also with the fact that WTaG turns are quite short, in terms of tokens, as can be inferred from Table 1.

Furthermore, we analyse WTaG dialogues for the presence of a conversational pattern that has emerged in D-Pro dialogues, where proactive utterances are build on top of same-turn reactive ones:

`[trigger turn + triggered turn]`

where the triggered turn's composition is:

`[reactive utt. + proactive utt.]`

Given the short length of turns, this kind of pattern is observed in fewer cases in WTaG dialogues, where the vast majority of proactive content is self-contained and is not bounded to a previous reactive utterance. Some occurrences of this pattern can be found where one proactive utterance immediately follows a confirmation, a backchanneling, or a grounding act (e.g., *great, yeah, ohh, ok, umm*).

We also examine the distribution of proactive turns across the dialogue span by dividing each dialogue into five segments, in order to determine whether proactivity is concentrated at the beginning or ending, in the middle, or distributed more evenly throughout the whole dialogue. Based on D-Pro Ubuntu dialogues, which are also task-guidance dialogues, we expect a fairly uniform distribution. Results, visualised in Figure 3 as a line chart, show the proactivity distribution of WTaG dialogues in comparison with D-Pro dialogues. WTaG distribution proves the hypothesis to be true, that proactive turns follow a very similar pattern to D-Pro Ubuntu, with a flatter line than the other sub-corpora and evenly distributed proactivity, even in the initial stages of the interaction. A slight downward tilt towards the end of the dialogue is expected, because the dialogue goal has already been reached and fewer proactive contributions are needed. The Chi-square test comparing the distributions across the six dialogue corpora at five observation points yielded a significant result ($\chi^2(20) = 38.88, p = 0.0069$), indicating that proactivity distributions differ across groups. Pairwise comparisons show that WTaG differs significantly from MultiWoZ ($\chi^2(4) = 13.61, p = 0.0087$) and Jilda ($\chi^2(4) = 10.25, p = 0.0364$). The comparison between WTaG and WhatsApp approached significance ($\chi^2(4) = 9.32, p = 0.0536$), whereas no significant differences were observed between WTaG and Nespole ($\chi^2(4) = 5.23, p = 0.264$), Nespole and Ubuntu ($\chi^2(4) = 7.76, p = 0.1009$), or between WTaG and Ubuntu ($\chi^2(4) = 0.55, p = 0.9688$), the latter showing an almost identical distribution to

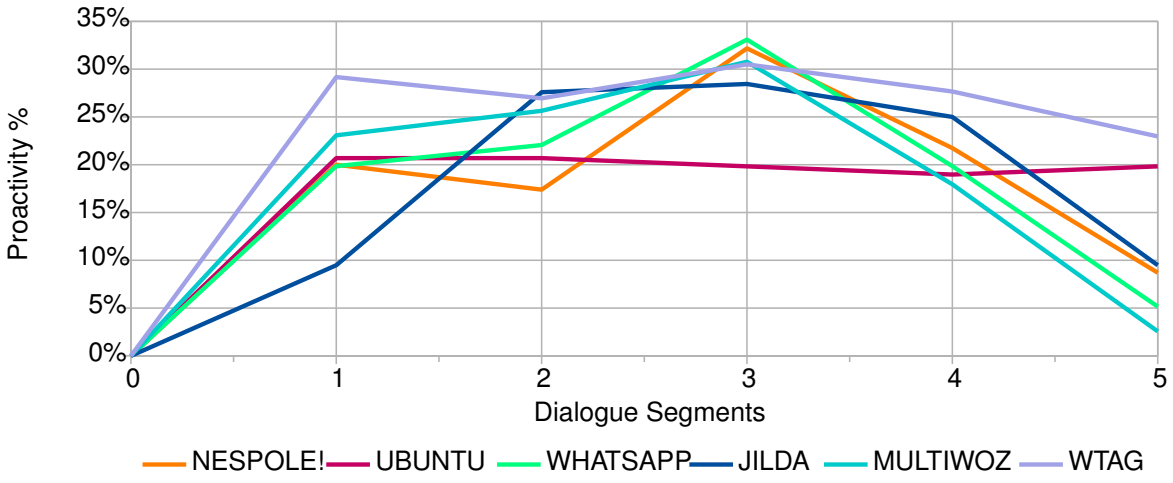

Figure 3: Distribution of proactive turns over five dialogue segments.

WTaG. In contrast, WhatsApp and Ubuntu differ significantly ($\chi^2(4) = 12.03, p = 0.0171$). After excluding Ubuntu and WTaG, the Chi-square test was no longer significant ($\chi^2(12) = 14.92, p = 0.2459$, suggesting that these two task-guidance corpora account for most of the global differences in proactivity distribution across groups.

5. Discussion

As far as our first research question (i.e., to what extent previous definitions of proactivity and previous annotation schemas can be adapted to the richer context of multimodal task-guidance dialogues), we showed that task-guidance dialogues required an extension of the D-Pro annotation schema. Specifically, the extension was necessary to account for the situations in which the user is asking (INFORMATION-REQUEST) for some instruction from the instructor, and in which the instructor (ACTION-REQUEST) wants the user to perform the requested action.

In addition, our analyses, based on the outcome of the experiments reported in Section 4, support the hypothesis that proactivity is more abundantly employed in less restricted interactions, suggesting a link to participants' agency: the more freedom they are given in a conversation, the more likely they are to take an active role, rather than responding passively. In fact, similar results are observed in D-Pro sub-corpora comparison, where spontaneous, less-structured dialogues hold the highest amount of proactive utterances: and that is valid for both chat-based dialogues (WhatsApp sub-corpus) and telephone conversations (Nespole sub-corpus).

Furthermore, task-guidance dialogues are highly collaborative interactions, where preventing mistakes and maintaining the correct sequence of actions is essential for achieving the dialogue goal. They are inherently procedural, temporally structured, and situated, which fosters opportunities and also the need to frequently initiate proactive acts.

Our findings from the comparison between textual and multimodal annotation (Table 2) indicate that detecting proactive behaviours in dialogue does not necessarily require multimodal information, even when the dataset is natively audio-visual, as long as transcriptions are provided. While video or audio elements may sometimes provide corrective or supporting information, the core elements of proactivity — novelty and helpfulness — can typically be detected with reliable results through expert manual annotation of texts alone. This seems to suggests that text-based analysis is sufficient, which can reduce the complexity and time associated with processing multimodal data. While this may hold true for expert, human annotators, this may not necessarily be the case for automated annotation systems, e.g. in an LLM-as-a-judge framework, and it may not be the case with non expert annotators as well (as argued on the basis of preliminary experiments reported in Section 4.1). Multimodal context may indeed serve as a supplementary resource in ambiguous cases, providing additional context when the textual element alone is insufficient.

A key strength of our annotation framework is its portability across datasets and task types, including non-linear and branching workflows, because it operationalizes proactivity through a domain-independent inventory of dialogue acts that remain interpretable regardless of the underlying procedure. Consistent with this, dialogue-act–based proactivity has been successfully applied in other interaction settings, including human–robot interaction (Kraus et al., 2020, 2022b) and decision-making under uncertainty with intelligent agents (Kraus et al., 2023).

6. Conclusions

We have investigated proactive behaviours in the context of multimodal task-guidance dialogues, conducting a systematic comparison with proactivity in textual chat-based task-oriented dialogues. Our findings show that (i) annotating proactivity in task-guidance dialogue require an extension of previous annotation schema; (ii) task-guidance dialogue require high collaborative interaction, resulting in a significant higher proportion of proactive turns; (iii) the visual component (the recorded videos) does not play a significant role in terms of additional proactive information with respect to the textual information.

As an additional outcome of the investigation, we make available both the annotation schema for proactivity and a subset of the Watch, Talk, and Guide dataset fully annotated with proactivity.

As for future research, we plan to use the annotated dialogues to instruct (through few-shot in-context learning) a large multimodal model to generate instructions imitating human-like proactive behaviours in task-guidance interactions.

7. Acknowledgements

Bernardo Magnini has been partially funded by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

Elisabetta Jezek has been partially funded by the University of Pavia with the ENHANCE SH project.

Matthias Kraus has been partially funded by the Bavarian Research Foundation within the project FORSocialRobots (AZ-1594-23).

8. Bibliographical References

Lavisha Aggarwal, Vikas Bahirwani, Lin Li, and Andrea Colaco. 2025. Generating dialogues from egocentric instructional videos for task assistance: Dataset, method and benchmark.

Prithviraj Ammanabrolu, Renee Jia, and Mark Riedl. 2022. Situated dialogue learning through procedural environment generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8099–8116.

John Langshaw Austin. 1962. *How to do things with words*. William James Lectures. Oxford University Press.

Vevake Balaraman and Bernardo Magnini. 2020. Proactive systems and influenceable users: Simulating proactivity in task-oriented dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue-Full Papers, Virually at Brandeis, Waltham, New Jersey, July. SEMDIAL*.

Yuwei Bao, Keunwoo Peter Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alexander De La Iglesia, Megan Su, Xiao Lin Zheng, and Joyce Chai. 2023. Can foundation models watch, talk and guide you step by step to make a cake? *arXiv preprint arXiv:2311.00738*.

Francesca Bargiela-Chiappini. 2003. Face and politeness: New (insights) for old (concepts). *Journal of Pragmatics*, 35(10-11):1453–1469.

Dan Bohus, Sean Andrist, Yuwei Bao, Eric Horvitz, and Ann Paradiso. 2024. " is this it?": Towards ecologically valid benchmarks for situated collaboration. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pages 41–45.

Sofia Brenna, Elisabetta Jezek, and Bernardo Magnini. 2025. Investigating proactivity in task-oriented dialogues. *Dialogue & Discourse*, 16(1):31–67.

Sofia Brenna and Bernardo Magnini. 2024. Last utterance proactivity prediction in task-oriented dialogues. In *Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024)*. CEUR-WS.org.

Rodney A Brooks. 2018. Intelligence without reason. In *The artificial life route to artificial intelligence*, pages 25–81. Routledge.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge University Press.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, pages 5016–5026.

Harry Bunt. 2006. Dimensions in dialogue act annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Harry Bunt. 2009. The dit++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an iso standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Harry Bunt and Yann Girard. 2005. Designing an open, multidimensional dialogue act taxonomy. In *DIALOR'05, Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue, Nancy*, pages 37–44.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

Susanne Burger, Laurent Besacier, Paolo Coletti, Florian Metze, and Celine Morel. 2001. The nespole! voip dialogue database. In *Seventh European Conference on Speech Communication and Technology*.

Susan Meredith Burt. 1994. Code choice in intercultural conversation: Speech accommodation

theory and pragmatics. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 4(4):535–559.

Jennifer Chu-Carroll and Michael K. Brown. 1999. An evidential model for tracking initiative in collaborative dialogue interactions. In Susan Haller, Alfred Kobsa, and Susan McRoy, editors, *Computational Models of Mixed-Initiative Interaction*, pages 49–87. Springer Netherlands, Dordrecht.

Céline De Looze, Stefan Scherer, Brian Vaughan, and Nick Campbell. 2014. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Howard Giles. 1979. Accommodation theory: Optimal levels of convergence. In *Language and social psychology*, pages 45–65. Basil Blackwell.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1:1–68.

Howard Giles and Tania Ogay. 2006. Communication accommodation theory. In Bryan B. Whaley and Wendy Samter, editors, *Explaining communication: Contemporary theories and exemplars*, pages 293–310. Lawrence Erlbaum Associates Publishers.

Howard Giles and Peter Powesland. 1997. Accommodation theory. In Nikolas Coupland and Adam Jaworski, editors, *Sociolinguistics*, pages 232–239. Macmillan Education UK, London.

Howard Giles, Donald M. Taylor, and Richard Bourhis. 1973. Towards a theory of interpersonal accommodation through language: some Canadian data. *Language in Society*, 2(2):177–192.

Adam M. Grant and Susan J. Ashford. 2008. The dynamics of proactivity at work. *Research in Organizational Behavior*, 28:3–34.

Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech acts*, pages 41–58. Brill.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-action: Conversational question answering over a large-scale knowledge base. *Advances in neural information processing systems*, 31.

Freya Hewett. 2017. *Sequential Organisation in WhatsApp Conversations.* Unpublished Bachelor's Thesis, Free University of Berlin, Summer Semester.

Pamela W. Jordan and Barbara Di Eugenio. 1997. Control and initiative in collaborative problem solving dialogues. In *Working Notes of the AAAI Spring Symposium on Computational Models for Mixed-initiative Interaction*, pages 81–84.

Cynthia Kersey, Barbara Di Eugenio, Pamela Jordan, and Sandra Katz. 2009. Knowledge co-construction and initiative in peer learning interactions. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, page 325–332, NLD. IOS Press.

Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 107–116.

Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2022a. ProDial – an annotated proactive dialogue act corpus for conversational assistants using crowdsourcing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3164–3173, Marseille, France. European Language Resources Association.

Matthias Kraus, Nicolas Wagner, Ron Riekenbrauck, and Wolfgang Minker. 2023. Improving proactive dialog agents using socially-aware reinforcement learning. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 146–155.

Matthias Kraus, Nicolas Wagner, Nico Untereiner, and Wolfgang Minker. 2022b. Including social expectations for trustworthy proactive human-robot dialogue. In *Proceedings of the 30th ACM conference on user modeling, adaptation and personalization*, pages 23–33.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. 2016. Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. In *Proceedings of the*

Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pages 2845–2851. IJCAI/AAAI Press.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015a. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. 2015b. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference*, pages 285–294.

Nadia Mana, Susanne Burger, Roldano Cattoni, Laurent Besacier, Victoria MacLaren, John McDonough, and Florian Metze. 2003. The nespole! voip multilingual corpora in tourism and medical domains. In *Eighth European Conference on Speech Communication and Technology*.

Michael Mctear. 2020. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, United States.

Elnaz Nouri and David Traum. 2014. Initiative taking in negotiation. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 186–193, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Böhm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, et al. 2024. What to say and when to say it: Live fitness coaching as a testbed for situated interaction. *Advances in Neural Information Processing Systems*, 37:75853–75882.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Eran Raveh. 2021. *Vocal accommodation in human-computer interaction: modeling and integration into spoken dialogue systems*. Ph.D. thesis, Saarländische Universitäts- und Landesbibliothek.

Alan Ritter, Colin Cherry, and William B Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Carol Myers Scotton. 1988. odeswitching as indexical of social negotiations. In Monica Heller, editor, *Codeswitching: Anthropological and Sociolinguistic Perspectives*, pages 151–186. De Gruyter Mouton, Berlin, New York.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

John R. Searle. 1975. Indirect speech acts. In Peter Cole and Jerry L. Morgan, editors, *Speech acts*, pages 59–82. Brill.

Thomas B Sheridan and William L Verplank. 1978. Human and computer control of undersea teleoperators. Technical report, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Irene Sucameli, Alessandro Lenci, Bernardo Magnini, Maria Simi, and Manuela Speranza. 2020. Becoming JILDA. In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLIC-it 2020*, volume 2769 of *CEUR Workshop Proceedings*, Bologna. CEUR-WS.

Irene Sucameli, Alessandro Lenci, Bernardo Magnini, Manuela Speranza, and Maria Simi. 2021. Toward data-driven collaborative dialogue systems: The jilda dataset. *Italian Journal of Computational Linguistics*.

David R. Traum. 1997. Views on mixed-initiative interaction. In *AAAI97 Spring Symposium On Mixed-Initiative Interaction*, pages 169–171.

David R. Traum and Elizabeth A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. Technical report, University of Rochester, USA.

Marilyn Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *28th Annual Meeting*

of the Association for Computational Linguistics, pages 70–78, Pittsburgh, Pennsylvania, USA. Association for Computational Linguistics.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. 2023. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281.

Steve Whittaker and Phil Stenton. 1988. Cues and control in expert-client dialogues. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 123–130, Buffalo, New York, USA. Association for Computational Linguistics.

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

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*.