Educational Dialogue Systems for Visually Impaired Students: Introducing a Task-Oriented User-Agent Corpus

Elisa Di Nuovo♡, Manuela Sanguinetti♢, Pier Felice Balestrucci♡, Luca Anselma♡, Cristian Bernareggi♣ and Alessandro Mazzei♡

♡ Computer Science Department, University of Turin, Italy
♢ Department of Mathematics and Computer Science, University of Cagliari, Italy
♣ Department of Mathematics, Laboratory S. Polin, University of Turin, Italy
{elisa.dinuovo | pierfelice.balestrucci | luca.anselma | alessandro.mazzei}@unito.it
manuela.sanguinetti@unica.it
cristian.bernareggi@gmail.com

Abstract

This paper describes a corpus consisting of real-world dialogues in English between users and a task-oriented conversational agent, with interactions revolving around the description of finite state automata. The creation of this corpus is part of a larger research project aimed at developing tools for an easier access to educational content, especially in STEM fields, for users with visual impairments. The development of this corpus was precisely motivated by the aim of providing a useful resource to support the design of such tools. The core feature of this corpus is that its creation involved both sighted and visually impaired participants, thus allowing for a greater diversity of perspectives and giving the opportunity to identify possible differences in the way the two groups of participants interacted with the agent. The paper introduces this corpus, giving an account of the process that led to its creation, i.e. the methodology followed to obtain the data, the annotation scheme adopted, and the analysis of the results. Finally, the paper reports the results of a classification experiment on the annotated corpus, and an additional experiment to assess the annotation capabilities of three large language models, in view of a further expansion of the corpus. The corpus is released under the Creative Commons Attribution Non Commercial 4.0 International license and available, only for research purposes, at: `https://zenodo.org/records/10822733`.

Keywords: conversational agents, education, accessibility, visual impairment

1. Introduction

Visually impaired people (VIP) encounter numerous difficulties in accessing university education. This is particularly true for STEM subjects, where concepts and data are often visually conveyed, making it difficult for visually impaired students to engage with them.[1]

Graphical structures (e.g. tables, circuits, diagrams) have information organized in an internal structure that requires visual inspection. For this reason, they are not fully accessible to VIP as they are not screen-readable. In fact, VIP usually have to rely on fragmented, incomplete, textual descriptions, or in the best cases, to alt-text image descriptions. To cope with these gaps, we believe that exploiting Natural Language Processing and Generation (NLP and NLG) could be a viable solution, as presented in previous studies (Chockthanyawat et al., 2017; Mascetti et al., 2017; Mazzei et al., 2019).

Dialogue Systems (DSs) in particular can be a valuable tool for the implementation of more effective educational practices for VIP, helping to overcome a gap in the access to knowledge. Building on this notion, early research on tutoring DSs, for example, focused on identifying the semantic and pragmatic phenomena in student-tutor dialogues (Benzmüller et al., 2003). This led to the development of a corpus of this specific type of dialogues in the specific domain of mathematical theorem proving (Benzmüller et al., 2006). The corpus was then enriched with Dialogue Act (DA) annotation (Buckley and Wolska, 2008).

The work presented here follows a similar path, in that the corpus that we introduce features dialogues in an educational setting. However, it is different for an important aspect. Instead of using the Wizard-of-Oz technique to collect the corpus, we collected actual human-DS dialogues. The DS was built with the purpose of verbally explaining Finite State Automata (FSA), being able to interact in conversations about, e.g., states and transitions, and is part of a broader initiative on the development of tools aimed to reduce accessibility barriers in edu-

Elisa Di Nuovo, as of 1st October 2023 is employed by the Joint Research Centre, European Commission, Ispra (VA), Italy. Her research work for this paper has been carried out prior to joining the Joint Research Centre.

[1] A recent initiative to help visually impaired students visualize a future in STEM: `https://tinyurl.com/initiativeSTEMforBlind`.

LREC-COLING 2024, pages 5507–5519
20-25 May, 2024.

cational contexts. In particular, the main objective is to make graphical structures accessible to VIP, and FSA were chosen as they are paradigmatic for other graphical structures that can be represented through tables. With this goal in mind, a rule-based DS was preliminarily designed to describe the functioning of two pre-defined FSA that were chosen as use cases. To evaluate the main capabilities and limitations of the agent, an experimentation phase was carried out with human users, both sighted and visually impaired. The resulting interactions thus form the dataset for this annotated corpus.

The annotation of the dataset is aimed at extracting valuable information on several key aspects. First, it aims to delineate how users interact with the agent, highlighting their expectations with respect to the agent's ability to provide appropriate responses. Thus, the primary purpose is to collect a corpus which can aid in the creation of a context-aware dialogue manager for the DS, enhancing its applicability within an educational environment. Second, it assesses the robustness of the DS in handling requests formulated in different ways by different users. Finally, it aims to address the possible disparities in the modes of interaction between sighted users and VIP in order to guide further refinement of the system to better meet the needs of the end user. For this purpose, the corpus was subjected to two different dimensions of annotation: one concerning DAs, to identify the basic dynamics of user-agent interactions, and the other dedicated to identifying possible errors in the conversation, aiming to precisely identify areas that need improvement in the agent.

Although the paper focuses primarily on outlining the salient features of the corpus, rather than delving into a detailed description of the DS or the results obtained from the experimentation conducted with human users, we provide below an account of the methodology employed to acquire the data (Section 3.1). We then outline the adopted annotation scheme and annotation process, including an assessment of the inter-annotator agreement and an analysis of the interactions based precisely on the outcomes of the resulting annotation (Sections 3.2, 3.3 and 3.4 respectively). After a technical validation of the resource using the Dual Intent and Entity Transformer (DIET) classifier (Bunk et al., 2020), a multi-task model for intent classification and entity extraction, which is the default classifier provided in RASA (Bocklisch et al., 2017) (Section 4), we finally conclude the paper with an exploratory validation experiment with some popular Large Language Models (LLMs) to annotate a corpus sample using the scheme devised for the DAs (Section 5).

2. Background and Related Work

In this section, we will briefly outline three aspects of the annotation in the conversational context that are specifically relevant to the work that we propose here: these are the annotation of DAs in dialogue corpora, the identification of errors in human-DS interactions, and the use of LLMs to support the annotation task.

Dialogue Acts Dialogue corpora are usually enriched with DA annotation. DAs, also known in linguistics with the term *speech acts* (Searle, 1969), are the explicitation of the actions we perform when we communicate.

There are different DA annotation schemes available in literature, as the general tendency is to create *ad-hoc* taxonomies per each project (Anderson et al., 1991) (Core and Allen, 1997; Jurafsky et al., 1998; Alexandersson et al., 1998; Clark and Popescu-Belis, 2004; Petukhova and Bunt, 2007; Vail and Boyer, 2014; Cai et al., 2023). Although since 2010 effort has been put in defining a DA annotation standard (Bunt et al., 2010, 2012, 2020), there are still scarce amounts of data annotated with this taxonomy (Bunt et al., 2019) (Ostyakova et al., 2023).

Recent work on NLG has demonstrated that DA annotation can improve the state-of-the-art task-oriented DS response (Yang et al., 2021; He et al., 2022; Wu et al., 2023). NLP techniques have also been applied for automating DA classification (Ivanovic, 2005; Ezen-Can and Boyer, 2015; Lin et al., 2023).

Conversational errors Within the context of dialogue annotation, some work has also focused on the identification of errors in the interactions. Aberdeen and Ferro (2003), for example, found that a complete understanding of misunderstandings requires a more thorough analysis than one based on DAs alone. Identifying possible conversational errors can in fact prove crucial in establishing recovery strategies and preventing breakdowns (Martinovsky and Traum, 2003; See and Manning, 2021). Furthermore, the identification of possible recurring errors in the DS may also allow a more timely assessment of which specific errors have a greater impact on user satisfaction, as well as on their expectations with respect to the DS capabilities (Aneja et al., 2020). This in turn can represent a key factor in the DS design.

In an attempt to cover all relevant conversational errors made by chat-oriented DSs, Higashinaka et al. (2021) defined a comprehensive taxonomy that builds upon two previously proposed theory- and data-driven taxonomies (Higashinaka et al., 2015a,b), and that exhibits an improved reliability, compared to the previous two, in terms of annotation agreement. Earlier works, on the other hand, focused instead on possible errors made by users (as in Bernsen et al. (1996)), in order to allow a better distinction between what should be considered a design error in the DS and what should rather be attributed to the behavior of the human user. Finally, the scheme proposed in Sanguinetti et al. (2020) aims to harmonize the taxonomies and terminologies presented in the works just mentioned, creating a unified framework for error annotation of both user and DS turns, and organized around Grice's maxims.

LLMs for annotation tasks After the release of the OpenAI interface in November 2022,[2] which allowed the free access to the GPT-3.5 generative model, there have been several research studies that focused on how LLMs behave when given a certain task.

On the task of annotating text, Gilardi et al. (2023) explore the potential of ChatGPT highlighting its good performance and cost-effectiveness in comparison to employing human annotators through platforms such as Amazon Mechanical Turk. In addition, Veselovsky et al. (2023) highlighted the increasing use of the model in crowd annotation, reporting an estimated amount of 33-46% of crowd workers using LLMs when completing their task. The use of such models for annotation is also explored in conversational contexts; Ostyakova et al. (2023), for example, point out that in the task of annotating dialogues with a multi-dimensional and hierarchical taxonomy of speech functions (Eggins and Slade, 2004), ChatGPT is able to reach human-like performance. Thanks to this, they propose to use ChatGPT automatic annotation as silver annotations to save both time and costs.

3. Corpus Development

3.1. Data Collection

As mentioned above, the data that form this corpus were obtained as a result of an evaluation experiment of a rule-based DS carried out with human users. Data was mainly collected in Spring 2023 plus a few dialogues collected in August-September 2023. The rule-based approach was motivated primarily by the need to ensure more consistent and correct responses, as opposed to an approach based on more recent LLMs. This choice was further motivated by the awareness of the known limitations of such models (Kasneci et al., 2023; Qadir, 2023), which, in an educational context, can pose significant obstacles to an effective learning. The objective of the experiments performed with the humans was to verify whether a dialogue-based interaction was more effective than a graphical representation, such as the transition table, for VIP to access the FSA. The DS was built using AIML (McTear et al., 2016), and the users were able to interact with the system via a web interface. The web interface used is compliant with the Web Content Accessibility Guidelines (WCAG) 2.1, ensuring full accessibility for VIP participants.[3]

For this experimentation, 32 volunteers were recruited, among which 6 VIP and 26 non-VIP. In the invitation letter the requirements to participate in the experimentation were fluency in English and a proper knowledge of FSA.

We would like to emphasize that involving VIP in experimentations is significantly difficult, especially if it requires them to have specific competencies. Indeed, the studies concerning assistive technologies for VIP have an average number of participants of 22.3 and they often are sighted persons (30% of the studies analyzed in (Brulé et al., 2020)).

Out of the 32 participants, all of them are non-native English speakers; as regards their demographics, 67.7% fall within the age range of 25–34, 6.5% are between 45–64 years, and the remaining 12.9% are evenly distributed between the age groups of 35–44 and 18–24. One participant holds a high school diploma, one participant has a PhD, and the remaining participants hold a Bachelor of Science. All participants have a background in Computer Science. Among them, 83.9% identify as male, and 16.1% as female.[4]

We collected 32 human-machine dialogues, for a total of 706 turns (i.e. 353 user turns and 353 DS turns) consisting on average of 22.06 turns per dialogue ($min = 8$; $max = 56$; $SD = 10.34$). In Table 1, we summarize the main corpus statistics, including both the overall data and the the ones pertaining to the two groups of participants.

	VIP	Non-VIP	All
# Dialogues	6	26	32
# Turns	194	512	706
Turns/Dialogue	32.33	19.69	22.06
Tokens/User's turn	3.87	5.91	5.53

Table 1: Basic corpus statistics. Third and fourth rows report average values.

From an initial observation of the collected dialogues, it can be seen that they are mainly question-answer pairs, probably due to the task setting—i.e. we have users who want to extract information

[2]Release notes available here: https://openai.com/blog/chatgpt.

[3]Accessibility guidelines available here: https://www.w3.org/TR/WCAG21/.

[4]For the sake of inclusivity, we added as answers also *I prefer not to say* and a free form text field.

about the FSA by interacting with the DS. The table also highlights some differences in the interactions of the different subgroups: it can be observed that VIP interactions are on average longer (i.e., they involve more turns), but the respective messages are shorter (in terms of number of tokens[5]). This is also exemplified in Figure 1, that shows the same questions posed by users from the different subgroups.

VIP	Non-VIP
Initial state	Which is the initial state?
q0 is both the initial and final state	
Arc from q0	Is there an arc from q0?
From q0 there is only one transition to q2 marked with 1.	

Figure 1: Two questions exemplifying the differences in question formulation between the two subgroups.

We compared VIP and non-VIP participants. Considering the number of turns per dialogue, $min_{nonVIP} = 8$, $max_{nonVIP} = 34$, $SD_{nonVIP} = 8.27$; $min_{VIP} = 20$, $max_{VIP} = 56$, $SD_{VIP} = 12.86, p = 0.005$. Considering the average number of tokens per user's turn, $min_{nonVIP} = 2.67$, $max_{nonVIP} = 8.67$, $SD_{nonVIP} = 1.46$; $min_{VIP} = 2.64$, $max_{VIP} = 4.82$, $SD_{VIP} = 0.84$, $p = 0.003$. In both cases the difference between VIP and non-VIP interactions appear to be extremely statistically significant using Welch's t-test for unequal sample sizes and unequal variances.

Data availability Due to the nature of the data at hand, obtained through the involvement of human participants, the corpus will be released for the sole purposes of academic research, and upon request. Researchers interested in accessing the data will be required to declare, via a dedicated form, their commitment to use such data only for the permitted purposes. The corpus metadata and the form to request the data are available in the following repository on Zenodo: https://zenodo.org/records/10822733.

3.2. Annotation Design

We annotated DAs applying the ISO 24617-2 Standard (Bunt et al., 2017, 2020). This scheme includes the annotation of 9 dimensions (e.g. *task*, *dialogue structuring*, *own communication management*) and a number of general-purpose (e.g. *set question* which can be used in *dialogue structuring* or *task* dimension) or dimension-specific (e.g. *self correction* which can be used only in *own* or *partner communication management*) communicative functions.

We used the more user-friendly annotation format called DiAML-TabSW, which is compliant with the ISO Standard but, instead of XML less-readable tags for the annotation, it uses a tabular format in a spreadsheet.[6] The main challenges in applying this ISO Standard to our data are: being created having spoken language in mind, some labels are not perfectly matching (e.g. the communicative function *stalling* in *time management* dimension indicating DAs aimed at gaining time during a conversation); since they have been created for human-human dialogues, recovering techniques or dialogue infelicities typical of human-agent dialogues are not pinpointed.

To tackle the latter issue mentioned above, we further enriched the annotation by applying an annotation layer which highlights conversation *errors*. For this purpose, we relied on the annotation scheme proposed in Sanguinetti et al. (2020), whose guidelines have been made available both in Italian and English.[7] That framework, in fact, was primarily motivated by the need to encode and easily identify interaction patterns in Italian human-DS dialogues in a customer-care setting. Nevertheless, the possibility of accessing labels that in turn were partially borrowed from previous schemes available in the literature (as described in Section 2), as well as the availability of labels that covered both users' and DS errors made us opt for this choice, despite the different domain (educational vs customer care) and language (English vs Italian) involved in our project. Additionally, we included the *Off-topic* label, which was not originally provided in the reference scheme, because it has been necessary to report a few cases in which the user asked completely irrelevant questions, mistakenly using the DS as an open-domain chatbot. Finally, all the turns where no errors have been identified by the annotators were labeled using the tag *None*.

Annotation example. An excerpt of annotated interaction that displays both dimensions is shown in Figure 2. In the reported example there are 8 turns—4 from one of the users and 4 from the DS—and the corresponding DAs and errors annotations. Also, the number of the actual turn within the dialogue has been included, with the prefix T before the turn text. Please note that users' turns are collected after the removal of special characters and whitespaces, punctuation included. This pre-

[5]The tokenization step has been carried out using one of the English models of spaCy (Honnibal et al., 2020) (en_core_web_sm). Moreover, punctuation symbols have been excluded from the token count.

[6]Template available here: https://dialogbank.lsv.uni-saarland.de/?page_id=1233.

[7]Guidelines available here: https://cutt.ly/cdMcnyM.

		DIALOGUE ACTS	ERRORS
T15	How many transitions are there	Ta:setQuestion	None
T16	There are a total of 3 arcs. Try asking me if there is a particular pattern among them.	Ta:answer; Ta:suggest	None
T17	Is there a transition from q0 to q5	Ta:propositionalQuestion	Ignoring question/feedback
T18	From q0 there is only one transition to q2 marked with 1.	Ta:answer	Indirect response
T19	According to your understanding of Automaton 2 what is its optimal spatial representation	Ta:setQuestion	None
T20	I dont' know. Try asking something like, "Describe the automaton."	AutoF:autoNegative; DS:suggest	Topic change
T21	The automaton accepts a language allowing words made of an odd number of 0s and 1s	Ta:checkQuestion	Ignoring question/feedback
T22	The language accepted by this automaton is made of zero or more words formed by a sequence ofa pair of 1s followed by a 0.	Ta:answer	Indirect response

Figure 2: Example of annotated user-DS interaction, with actual turn numbers (prefixed with T) in the dialogue.

processing is implemented to maximize the DS ability to trigger the right pattern and consequent response. This is the reason why punctuation is only present in DS turns. This fragment of interaction first highlights the bipartite question-answer character typical of the dialogues that make up this corpus, and shows a recurrent mode of question formulation that characterizes the most frequent DAs. In the question in $T15$ the user is asking about the elements of a specific set, i.e. that of states; in the question in $T17$ the user wants to know whether the content of a given proposition is true, while in the question in $T21$ the user has an uncertain opinion about the truth of the proposition and asks the DS to confirm it. Please note that we considered it to be a question from the textual and extra-textual context and not based on syntax, compliant to the ISO annotation guidelines. In the DS turns, we find the answers to the user's questions. In addition in $T16$ we annotated also the DS initiative in prompting for a follow-up question about the transitions, which is annotated in *task* dimension. In $T20$ instead, the DS suggestion is classified as *dialogue structuring* as it is not purely intended to advance the task, but to prevent the dialogue from stagnating. Looking at annotated errors, we marked $T17$ and $T21$ as *Ignoring question/feedback* as the user is not accepting the DS suggestions, but they ask a different question than the recommended one. Then, the other errors are annotated on DS turns, signaling *indirect responses* both in $T18$ and $T22$. In fact, in $T18$ the desired answer would be 'No, there is no transition from q0 to q5.' and since we are in an educational context, it could be also preferable to add that q5 is not a state of the automaton. In $T22$, again the direct answer would be 'No' plus the reason provided in the DS answer.

3.3. Inter-Annotator Agreement

Two annotators (among the authors of this paper) were involved in the annotation of both DAs and errors. A training was provided on how to annotate them. They were also provided with annotation guidelines and related papers. The two annotators worked independently and asynchronously on the whole corpus. At the end of the annotation 12 labels were annotated, among all the possible combinations foreseen in the ISO.[8] Both the annotators used maximum three DA labels per turn.

As reported above, the DA annotation was carried out in two phases. Phase 1 was carried out on the data collected in Spring 2023, Phase 2 on the data collected in August-September 2023. The disagreement captured in both phases does not distinguish apparent disagreement due to annotator's errors from real disagreement. We measured disagreement using Cohen's kappa (κ) (Cohen, 1960). The obtained κ in the first phase is 0.74, which corresponds to substantial agreement, using Landis and Koch (1977) terms. This score is obtained considering the whole DA annotation (i.e. dimension and communicative function). We also computed κ separately on dimension and communicative function annotation. From this analysis it results that there was more agreement on communicative function selection ($\kappa = 0.91$), than dimension ($\kappa = 0.56$).

In Phase 2, κ is 0.96, which corresponds to almost perfect agreement. The reason behind this higher score obtained in Phase 2 data is twofold: first, it is motivated by the fact that, before the annotation of these new data, Phase-1 disagreement was already resolved, thus disagreement com-

[8]The annotated labels are: *AutoF:autoNegative*, *DS:opening*, *DS:suggest*, *OCM:selfCorrection*, *SOM:initGreeting*, *Ta:answer*, *Ta:checkQuestion*, *Ta:propositionalQuestion*, *Ta:request*, *Ta:setQuestion*, *Ta:suggest*, *TUM:turnAccept*.

Dialogue Acts	
Users	
Ta:setQuestion	51.54
Ta:request	27.45
Ta:propositionalQuestion	16.25
Ta:checkQuestion	1.96
OCM:selfCorrection	1.40
DS:opening	0.56
TuM:turnAccept	0.56
SOM:initGreeting	0.28
DS	
Ta:answer	48.14
AutoF:autoNegative	24.79
DS:suggest	24.79
Ta:suggest	2.27

Table 2: Distribution (in %) of single DAs in the users' and DS turns.

Errors	
Users	
Repetition	42.75
Ignoring question/feedback	28.24
Grammatical error	13.74
Non-understandable	4.58
Off-topic	3,82
Lack of information	3.05
Non-cooperativity	2.29
Ill-formed	1.53
DS	
Topic change	62.30
Straight wrong response	14.66
Indirect response	10.47
Excess of information	7.33
Lack of information	4.71
Ignoring question/feedback	0.52

Table 3: Distribution (in %) of single errors in the users' and DS turns. The percentage of error labels is computed over the total number of errors.

mented and annotation doubts clarified; second, Phase-2 data (6 dialogues) are less than Phase-1 data (26 dialogues). Also in this case we computed κ separately on dimension ($\kappa = 0.9593$) and communicative function ($\kappa = 0.9610$). The obtained scores confirm that dimension is the label in which there was more disagreement, although in this case it was a minor difference. The κ score obtained on the whole corpus (Phase 1 + Phase 2 data) is 0.76.

Differently from the 2-step DA annotation, errors were annotated in a unique phase. The agreement achieved is $\kappa = 0.56$, which is moderate agreement. The total number of labels annotated is 18 (considering also *None* label).

3.4. Annotation Results

This subsection discuss the annotation results focusing first on the differences between users and DS. And then focuses on users, highlighting differences between VIP and non-VIP users.

Users and DS. Statistics about the gold standard annotated corpus are reported in Tables 2 and 3, for DAs and errors, respectively. Both tables group DAs occurring in users' or in DS turns. Looking at users' DAs (Table 2), it is possible to confirm the conversation dynamics mentioned in the previous subsections. In fact, the three most occurring DAs are all pertaining to the *task* dimension and elicit information (in the syntactic structure of *question* or *request*). In addition, it is possible to notice that, being a text-based dialogue, *self corrections* are limited. Moreover, our participants, knowing that they are interacting with a DS for FSA exploration, did not feel the necessity to strictly follow *social obligation* rules. This information is then complemented by the DAs annotated in DS turns, which are basically answers to users' questions (including non-answers indicated by *auto feedback: auto negative*, which are followed by *suggestions* by design to help the dialogue moving on).

As for the results of the error annotation, we observed that 42.78% of the overall number of turns (of both parties) includes at least one error. The distribution of error labels, which is further calculated over the total number of errors for each party (DS and users), is shown in Table 3. Such distribution highlights infelicities in users' turns and a strategy adopted to cope with DS non-answers (*repetition*). In fact, we used the error label *repetition* to highlight in the dialogues not only pure repetitions, but also reformulations aimed at being understood by the DS and obtaining an answer. In the DS part, apart from being evident the design strategy to cope with unmatched users' utterances (*topic change*), thanks to error annotation it is possible to highlight infelicitous pattern-matching (*straight wrong response*) and other minor issues.

To assess the strength of association between the two annotation dimensions (i.e. DAs and errors), we calculated Cramer's V (Cohen, 1988), which revealed a moderate association with a value of 0.44.

VIP and non-VIP: DAs. Non-VIP account for 72.52% of total DAs (26 users); VIP users for 27.48% of annotated DAs (6 users). There is more variety of annotated DAs in non-VIP than VIP. In fact, 5 DAs are annotated only in non-VIP dialogues (i.e. *DS:opening*; *DS:opening* and *SOM:initGreeting* in the same turn; *Ta:checkQuestion*; *Ta:request* and *Ta:setQuestion* in the same turn; *TuM:turnAccept*), the other 4 DAs occur in both subgroups with different percentages (i.e. *OCM:selfCorrection* 60% non-VIP, 40% VIP; *Ta:propositionalQuestion* 94.83% non-VIP, 5.17% VIP; *Ta:request* 48.52% non-VIP, 51.58% VIP; *Ta:setQuestion* 76.37% non-VIP, 23.63% VIP). It is noteworthy how, even though VIP are in smaller numbers, *OCM:self-correction* and *Ta:request* annotations are fairly evenly distributed between the two subgroups, with *Ta:request* having a higher percentage among VIP compared to non-VIP. This reflects the trend we have already reported in Section 3.1, where VIP, as they are using fewer tokens, tend to formulate more requests than their counterparts.

VIP and non-VIP: Errors. Non-VIP account for 66.96% of total errors (26 users); VIP for 33.04% of annotated errors (6 users). 69.92% of non-VIP turns do not contain errors compared to 60.82% of VIP. The most annotated errors in the non-VIP turns are *repetition* (29.87%), *ignoring question/feedback* (22.08%), *grammatical error* (18.18%), and *ignoring question/feedback; repetition* (12.99%). In contrast, in the *VIP* turns, *repetition* accounts for 52.63% of the errors, while *lack of information*, *grammatical error*, and *ignoring question/feedback* each account for 10.53%. As far as *lack of information* error is concerned, it is worth noticing that it is usually associated to DS answers, all occurring in non-VIP dialogues. However, we also marked it in 4 VIP turns as they were formed of a single token not providing enough information (e.g. *Describe*, which could be referred to the FSA, its spatial representation, the states, the alphabet, etc.).

4. Validating Users' DA Annotation with DIET

In this section we introduce some preliminary experiments in classifying the DAs applied in our corpus, in order to further validate the annotated data, and to create a baseline for a DA classification task. With a view to the further development of the DS, we addressed the task of recognizing users' dialogue acts structuring this process as an intent detection task.

To do so, we ran the Natural Language Understanding component available in RASA (Bocklisch et al., 2017), a well-known framework for the development of DSs. Specifically, we used the Dual Intent and Entity Transformer (DIET) classifier (Bunk et al., 2020), a multi-task model for intent classification and entity extraction, which is the default classifier provided with the platform. Due to the considerable skewness in the distribution of the various DAs in this corpus,[9] an effective classification was only possible on a smaller portion of the data, namely those annotated for the three most frequent users' DAs: *Ta:request*, *Ta:setQuestion* and *Ta:propositionalQuestion*. This selection was made because the remaining DAs lacked sufficient data instances for training purposes, as per our analysis using the DIET tool. The classifier was tested on the intent detection task only (thus excluding entity recognition) and using slightly different pipelines: one with a bag-of-word representation of character n-grams (1 to 4), already set in the default configuration, and one using BERT pre-trained embeddings.[10] In both configurations the classifier was trained for 100 epochs, a 5-fold cross-validation was performed, and it was repeated for 3 runs; the results were finally averaged over all runs.[11] Table 4 shows the results obtained, in terms of F-score, on all three individual classes. Finally, the macro-F1 score was calculated to provide an indicator of the overall performance of the models.

	BoW	BERTemb
Ta:setQuestion	0.9852	0.9963
Ta:request	0.9645	0.9655
Ta:propositionalQuestion	0.9293	0.9583
macro-F1	0.9597	0.9734

Table 4: Classification results of the three most frequent DAs in the corpus with the DIET classifier and using two different vector representations.

As the table shows, the results obtained in the experiment are very promising using both configurations, despite the rather small number of training instances; they also align overall with the agreement results obtained by the annotators on this dimension. We attribute such high results mainly to the linguistic nature of these data: even though the subgroup analysis described in Section 3.1 revealed significant differences between VIP and non-VIP, the questions posed by the users are more generally characterized by a low diversity, both from a syntactic and a lexical point of view. Concerning the latter in particular, the type/token ratio was calculated on the questions labeled with these three DAs, namely to assess such degree of diversity.[12] The results show a value equal to 0.15 for *Ta:setQuestion* (which is also the class that obtained the highest classification results), 0.18 for *Ta:propositionalQuestion* and 0.23 for *Ta:request*. This low linguistic variability can be partly explained

[9]Please note that it is a common aspect in DA-annotated corpora (Lin et al., 2023).

[10]All the the layers of BERT model were initialized from the following model checkpoint: `https://huggingface.co/rasa/LaBSE`.

[11]The RASA configuration files have been made available in the dataset repository.

[12]The index has values ranging from 0 to 1. The closer the value is to 0, the lesser the vocabulary variation.

by the restricted domain (the one relating to FSA), which requires the use of a pre-defined specialized language.

5. Annotating DAs with LLMs

We explore in this section the LLM performance in annotating the DAs in a set of dialogue turns. To do this, we chose three different models—ChatGPT, LLama 2 (Touvron et al., 2023), and Tk-Instruct-11B (Wang et al., 2022)—all instruction-tuned LLMs, which improve their performance by relying on reinforcement learning from human feedback (Ouyang et al., 2022).

We conducted the experiments with ChatGPT 3.5 through the web interface,[13] and with Llama 2 and Tk-Instruct-11B by downloading the pre-trained models and prompting them. All prompts were executed using the default settings of each model.

Prompt Engineering. When constructing a prompt for instruction-tuned models, Fu et al. (2022) suggest that the most effective prompts are those focused on reasoning, where the prompt explains the reasoning necessary to perform the task. This prompt style performed better than other 5 prompt styles in intent-slot and goal tracking annotations (Addlesee et al., 2023). A similar prompt type, called multi-step tree-like annotation prompt (including yes/no questions) seemed to be the best approach when compared to direct annotation (let the model decide a label among the set of labels) and step-by-step scheme (developing on intermediate labels) (Ostyakova et al., 2023). It is worth noticing that the logic behind reasoning and multi-step tree-like annotation prompts is similar. In fact, both are based on the explanation of the logical reasoning necessary to attain the answer. The only difference is that in multi-step tree-like annotation prompts, this is done by posing propositional questions. For this reason, we used propositional questions to list reasoning steps trying to combine the reasoning prompt and the multi-step tree-like annotation prompt.

Our prompts are all made of the task definition (i.e. annotating DAs in turns), the context (i.e. the dialogue takes part between a student and a DS programmed to answer about a specific FSA), the annotation constraints (e.g. one or more DAs can be assigned to each turn, additional co-text[14]), the presentation of the label set,[15] and their decomposition into dimensions and communicative functions, and the propositional questions to guide the selection of the most appropriate dimension and communicative function (e.g. 'Does it concern the underlying task? If yes then it's Task dimension, otherwise go ahead', 'Is the turn aiming at knowing if the proposition expressed in it is true? Then, it is propositionalQuestion. If the sender seems to know that the addressee knows the answer, then it is checkQuestion'). Finally, the last part of the prompt presents the turn to annotate—which is the only part differing between the prompts—in which we ask to substitute the [MASK] token with the chosen annotation.

[13] https://openai.com/chatgpt - ChatGPT 3.5 September 25 Version.

[14] Co-text is the linguistic, verbal context in which something occurs.

[15] We provided only the 12 labels actually annotated in the corpus.

Prompt	ChatGPT	Llama 2	Tk
1	0.33/1	0.00/1	0.00/1
2	0.33/1	0.00/1	0.50/1
3	1.00/1	0.00/1	0.50/1
4	0.00/1	0.17/1	0.00/1
5	0.50/1	0.00/1	0.00/1
6	0.17/1	0.00/1	0.50/1
7	0.83/2	0.00/2	0.50/2
8	1.00/1	0.00/1	0.50/1
9	1.00/2	0.00/2	0.50/2
10	1.17/3	1.00/3	0.50/3
average	0.63/1.40	0.11/1.40	0.35/1.40
SD	0.41/0.70	0.31/0.70	0.24/0.70
success	45.24%	8.33%	25.00%

Table 5: Results of ChatGPT 3.5, Llama 2 and Tk-Instruct-11B on each Prompt. n/m represents a score of n out of m, where m is the maximum attainable score for the prompt.

Experimental Protocol. For this experiment, we selected 10 turns to annotate. The DA annotation task was divided into two main steps: first, identifying the dimension, second, assigning the communicative function. We ran the same prompt three times and independently on each turn. The prompts do not contain examples, so we tested the performance of off-the-shelf LLMs in a zero-shot setting.

For each turn the maximum attainable score is computed, considering 1 point per each gold DA (0.5 points per subtask: dimension and communicative function). To evaluate the predictions we assigned 1 point per complete, correct DA, 0.5 if only one subtask is correct, and 0 if the annotation is incorrect or if the model failed to interpret the prompt correctly.

LLM Results. Table 5 displays the averaged results on the 3 runs per prompt in n/m form, where n is the averaged attained score and m the maximum achievable score per prompt. $Success$ reports the percentage of success rate (correct annotated DA) per LLM. As it can be observed, the results for each model are notably low. However, ChatGPT achieves a moderate success rate, which would speed up the annotation procedure if used as draft for human correction. Llama 2 provided annotations only in 2 out of 10 turns. The remaining times, it generated fictitious dialogues or hallucinated information using the turn to be annotated. Finally, Tk-Instruct-11B provided annotation labels, but never complete ones, it mainly provided communicative functions (e.g. *initGreeting*)—always one per turn—and only in one occasion it provided only the dimension (i.e. *task*).

Although these results are far from being comparable to human performance, they are indeed a good starting point for a semiautomatic annotation of the DAs —especially as far as ChatGPT is concerned—as they are elicited in a zero-shot setting.

6. Conclusion

This paper reported the study, the design and the realization of a corpus of human-machine dialogues aimed at verbally exploring FSA. This is a first necessary step for the development of a complete DS for allowing VIP to learn from graphical structures. We described the development of the corpus and its evaluation. We reported and discussed the main features regarding DA annotations, further enriched with conversation error annotation, as well as a subgroup analysis on DAs and errors comparing users and DS turns, and VIP and non-VIP users. Moreover, we performed a technical validation of DA annotations which uses the built-in classifier available in RASA, in two configurations, trained on the annotated corpus. Finally we reported the results of an exploratory experiment that evaluate the annotation capabilities of three LLMs for a future extension of the corpus.

In the future, we want to fine-tune some LLMs to develop an automatic annotation tool for accelerating the annotation process and expand the size of the corpus. Additionally, we want to leverage the DA annotation to enhance the proactive capabilities of the DS.

7. Limitations

We believe that this corpus can represent a useful resource in supporting the development of assistive technologies for VIP; however, we are also aware of the limitations of the study. One in particular concerns the unbalanced distribution of the annotated labels, especially the ones concerning errors, that consist in the higher frequency of few labels on the one hand, and the greater sparsity of the remaining ones on the other. This does not allow at the current state of the resource the proper training of models for the recognition of all possible utterances (and not just a portion of them) that a user may produce in such a context, let alone for early identification of possible errors that would allow effective recovery strategies to be implemented.

We also add another important limitation related to the choice of using ChatGPT among the models for the experiments in Section 5. As well known, ChatGPT is a closed model, and this can pose challenges to the reproducibility of the experiments we conducted. Nonetheless, especially given the exploratory and comparative nature of the experiment in question, we believe that including this model as well could provide a further contribution to the wider debate about its actual capabilities with respect to specific tasks.

8. Ethical Considerations

The availability of data produced by real users, while being a crucial factor in the creation of language resources and their exploitation for NLP applications, requires special attention as to how they are collected, used, and possibly disseminated. As regards their collection, we have worked to ensure that each participant in the experiments was properly informed about the scope of the research and their rights and requirements as participants. Together with the invitation letter, a consent form was provided in which the user declared, among other things: *i*) to be aware of the objectives of this research; *ii*) to participate on a voluntary basis; *iii*) to be of legal age; *iv*) to be aware that the study was in line with current data processing and protection regulations, on both national and EU level; *v*) to be aware of the possibility of withdrawing from the study at any time, without explanation, without any penalty and obtaining the non-use of the data.
Concerning the use of the data, the personal information provided by each participant was only used for statistical purposes and in aggregated forms, while for the annotation task the sole content of the interactions with the DS was taken into account.
Finally, as also mentioned in Section 3.1, the release of our corpus admits use for scientific purposes only and following mechanisms aimed at enforcing the accountability of anyone accessing the data.

9. Acknowledgments

We thank the anonymous reviewers for their constructive comments. This research has been developed for the project NoVAGraphS (Non-Visual Access to Graphical Structure), and is partially supported by the grant Progetto CRT 2021.1930 from Fondazione CRT (Cassa di Risparmio di Torino).

10. Bibliographical References

John Aberdeen and Lisa Ferro. 2003. Dialogue patterns and misunderstandings. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with llms: Comparing pre-training, fine-tuning, and prompt engineering.

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. Dialogue acts in verbmobil-2 second edition. *VM-Report 226*.

Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2020. Conversational error analysis in human-agent interaction. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Christoph Benzmüller, Armin Fiedler, Malte Gabsdil, Helmut Horacek, Ivana Kruijff-Korbayová, Manfred Pinkal, Jörg Siekmann, Dimitra Tsovaltzi, Bao Quoc Vo, and Magdalena Wolska. 2003. Tutorial dialogs on mathematical proofs. In *Proceedings of IJCAI-03 Workshop on Knowledge Representation and Automated Reasoning for E-Learning Systems*, pages 12–22.

Niels Ole Bernsen, Laila Dybkjær, and Hans Dybkjær. 1996. User errors in spoken human-machine dialogue. In *Proceedings of the ECAI '96 Workshop on Dialogue Processing in Spoken Language Systems*, ECAI '96, Berlin, Heidelberg. Springer-Verlag.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *CoRR*, abs/1712.05181.

Emeline Brulé, Brianna J Tomlinson, Oussama Metatla, Christophe Jouffrais, and Marcos Serrano. 2020. Review of quantitative empirical evaluations of technology for people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Mark Buckley and Magdalena Wolska. 2008. A classification of dialogue actions in tutorial dialogue. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 73–80.

Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. DIET: lightweight language understanding for dialogue systems. *CoRR*, abs/2004.09936.

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

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The iso standard for dialogue act annotation. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*, pages 549–558. European Language Resources Association (ELRA).

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the iso 24617-2 standard. In Deborah A. Dahl, editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*, pages 109–135. Springer.

Jon Z Cai, Brendan King, Margaret Perkoff, Shiran Dudy, Jie Cao, Marie Grace, Natalia Wojarnik, Ananya Ganesh, James H Martin, Martha Palmer, et al. 2023. Dependency dialogue acts–annotation scheme and case study. *arXiv preprint arXiv:2302.12944*.

Seehait Chockthanyawat, Ekapol Chuangsuwanich, Atiwong Suchato, and Proadpran Punyabukkana. 2017. Towards automatic diagram description for the blind. In *i-CREATe. The International Convention on Rehabilitation Engineering and Assistive Technology*, pages 1–4.

Alexander Clark and Andrei Popescu-Belis. 2004. Multi-level dialogue act tags. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 163–170.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences. Second edition*. Lawrence Erlbaum Associates.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Suzanne Eggins and Diana Slade. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.

Aysu Ezen-Can and Kristy Elizabeth Boyer. 2015. Understanding student language: An unsupervised dialogue act classification approach. *Journal of Educational Data Mining*, 7(1):51–78.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10749–10757.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online. Association for Computational Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015a. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 87–95, Prague, Czech Republic. Association for Computational Linguistics.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015b. Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Edward Ivanovic. 2005. Dialogue act tagging for instant messaging chat sessions. In *Proceedings of the ACL Student Research Workshop*, pages 79–84.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meter, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1998. Switchboard discourse language modeling project final report. In *Johns Hopkins LVCSR Workshop, 1997*. Citeseer.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Jionghao Lin, Wei Tan, Ngoc Dang Nguyen, David Lang, Lan Du, Wray Buntine, Richard Beare, Guanliang Chen, and Dragan Gašević. 2023. Robust Educational Dialogue Act Classifiers with Low-Resource and Imbalanced Datasets. In *International Conference on Artificial Intelligence in Education*, pages 114–125. Springer.

Bilyana Martinovsky and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the ISCA Workshop on Error Handling in Dialogue Systems*.

Sergio Mascetti, Andrea Gerino, Cristian Bernareggi, and Lorenzo Picinali. 2017. On the evaluation of novel sonification techniques for non-visual shape exploration. *ACM Transactions on Accessible Computing*, 9:1–28.

Alessandro Mazzei, Michele Monticone, Cristian Bernareggi, et al. 2019. Using NLG for speech synthesis of mathematical sentences. In *INLG 2019-12th International Conference on Natural Language Generation, Proceedings of the Conference*, pages 463–472. Association for Computational Linguistics (ACL).

Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices*, 1st edition. Springer Publishing Company, Incorporated.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. Chatgpt vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Volha Petukhova and Harry Bunt. 2007. A multidimensional approach to multimodal dialogue act annotation. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS)*, pages 142–153.

Junaid Qadir. 2023. Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9. IEEE.

Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalerandi, Dario Mana, and Rossana Simeoni. 2020. Annotating errors and emotions in human-chatbot interactions in Italian. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 148–159, Barcelona, Spain. Association for Computational Linguistics.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Abigail See and Christopher Manning. 2021. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alexandria Katarina Vail and Kristy Elizabeth Boyer. 2014. Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes. In *International conference on intelligent tutoring systems*, pages 199–209. Springer.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks.

Qingyang Wu, James Gung, Raphael Shu, and Yi Zhang. 2023. Diacttod: Learning generalizable latent dialogue acts for controllable task-oriented dialogue systems. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 255–267.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14230–14238.

11. Language Resource References

Anne H. Anderson and Miles Bader and Ellen Gurman Bard and Elizabeth Boyle and Gwyneth Doherty and Simon Garrod and Stephen Isard and Jacqueline Kowtko and Jan McAllister and Jim Miller and Catherine Sotillo and Henry S. Thompson and Regina Weinert. 1991. *The HCRC map task corpus*. Sage. PID https://doi.org/10.1177/002383099103400404.

Benzmüller, Christoph and Horacek, Helmut and Lesourd, Henri and Kruijff-Korbayova, Ivana and Schiller, Marvin and Wolska, Magdalena. 2006. *A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material*. European Language Resources Association (ELRA). PID http://www.lrec-conf.org/proceedings/lrec2006/pdf/778_pdf.pdf.

Bunt, Harry and Petukhova, Volha and Malchanau, Andrei and Fang, Alex and Wijnhoven, Kars. 2019. *The DialogBank: dialogues with interoperable annotations*. Springer. PID https://doi.org/10.1007/s10579-018-9436-9.