

# Dialogue Games for Benchmarking Language Understanding: Motivation, Taxonomy, Strategy

David Schlangen

Computational Linguistics / Department of Linguistics

University of Potsdam, Germany

david.schlangen@uni-potsdam.de

## Abstract

How does one measure “ability to understand language”? If it is a person’s ability that is being measured, this is a question that almost never poses itself in an unqualified manner: Whatever formal test is applied, it takes place on the background of the person’s language use in daily social practice, and what is measured is a specialised variety of language understanding (e.g., of a second language; or of written, technical language). Computer programs do not have this background. What does that mean for the applicability of formal tests of language understanding? I argue that such tests need to be complemented with tests of language use embedded in a practice, to arrive at a more comprehensive evaluation of “artificial language understanding”. To do such tests systematically, I propose to use “Dialogue Games”—constructed activities that provide a situational embedding for language use. I describe a taxonomy of Dialogue Game types, linked to a model of underlying capabilities that are tested, and thereby giving an argument for the *construct validity* of the test. I close with showing how the internal structure of the taxonomy suggests an ordering from more specialised to more general situational language understanding, which potentially can provide some strategic guidance for development in this field.

## 1 Introduction

*Steff is sitting at a desk, intently focussed on the piece of paper in front of them. The task is to read short paragraphs of text and then to answer questions about them, and how well Steff does at this will determine their “language proficiency score”, and thus contribute to whether they will get admission to the University of their choice or not – for it is a requirement to understand the language that is being tested. Steff gets up and heads toward the door – “the other one” shushes the proctor. Outside, Steff sees their friend, looking at them,*

*and greets them with “next time”; the reply comes immediately: “drinks?”*

The subfield of “Natural Language Understanding” (NLU) within the field of Natural Language Processing (NLP) uses tests of the first kind—written responses to written material—to measure the degree to which a technical artefact can be said to possess the *ability* of understanding natural language. More recently, NLP has expanded towards tackling more situated and less abstracted cases of language use—as in the second part of the story, if not quite as social—, under the headings “language and vision (navigation)” or “embodied AI” (Duan et al., 2022; Gu et al., 2022; Sundar and Heck, 2022),<sup>1</sup> with evaluation practices not yet fully established.

This paper aims to systematise already ongoing efforts in this direction and to support future ones, by first asking how these kinds of language understanding settings—formal, and situated—relate. Coming to the conclusion that Situated Language Understanding (SLU) requires different testing approaches, and that NLU evaluation has proceeded somewhat haphazardly, I will describe the design choices for creating situated language use activities, relating them to a particular, but abstract, model of situated language understanding; thereby addressing for this new field the concern of Schlangen (2021) that progress cannot be measured without clarity about underlying theoretical commitments. More specifically, I want to show a way how Dialogue Games can be integrated into a sound methodology for computational research on meaning, by providing explicit information about relations between research objects (see Figure 1).

It might be useful to mention at the outset what this paper is *not* aiming to do, which is to make rec-

<sup>1</sup>The field has moved *back* to this, one should say, as of course situated language used to be much more in the center, as for example in the very early SHRDLU system (Winograd, 1972).

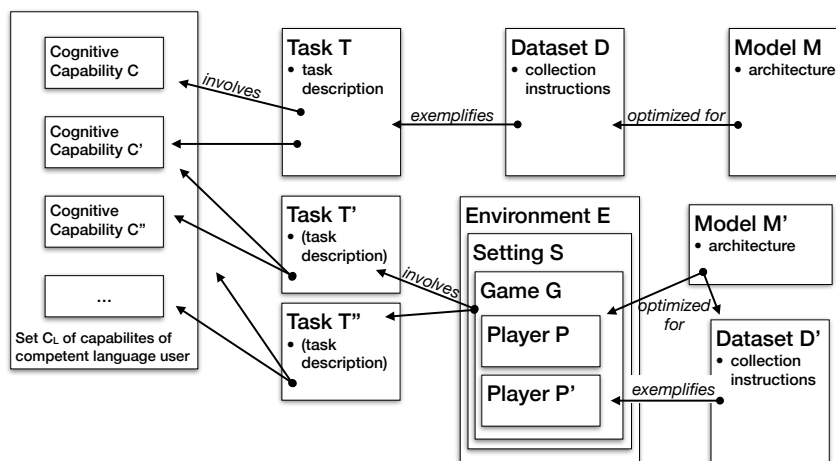


Figure 1: The structure of relations between the research objects *model*, *dataset*, *task*, *environment*, *setting*, *game*, and *cognitive capability*. Adapted from (Schlangen, 2019b).

ommendations for *how* SLU should be modelled, in the technical sense. While I see value in being able to understand the components of a task and how they interact (which suggests modularity in design), nothing precludes attempting benchmarks of the types described here with monolithic models, and even, insofar as the requisite information can be represented in the right way, with “general purpose” models such as Large Language Models.

## 2 Background On Measurement

*Language understanding*, as a psychological process, is observable only in its reflections in behaviour.<sup>2</sup> But not any behaviour counts, and not any behaviour is *measurable*—and measurement is our goal here. Experimental psychology has developed many ways to deal with the problem of measurement of unobservables in a principled manner. A central notion here is that of *validity* of a measurement instrument: Does the instrument indeed measure the unobservable *construct* that it is set up to measure?

This is not the place to give a full introduction into that field,<sup>3</sup> so I will concentrate on those aspects of validity that I see as attainable through the methodology described below. A first claim for validity of an instrument is via an appeal to its *face validity*: That it intuitively appears to capture the construct. Being able to count in a text the occurrences of the letter “o”, for example, would lack such face validity for the construct “language

understanding”, while being able to answer questions about it may be argued to have it. (Although our intuitions leave us quickly here: What if some questions are answered well, but others bizarrely badly? More on this below.) A second element is *ecological validity*, an argument for how closely the measure resembles the use of the construct in the domains in which it ordinarily shows. Measures of situated language understanding (as will be developed here) can arguably make a claim for high ecological validity—this kind of language understanding plays a large role in people’s lives—but more abstract or formalised understanding tasks do occur, in situations as described above. Lastly, quantifiable support comes from *convergent validity*, as different measures that are purportedly addressing the same construct can be expected to correlate, and if they do so, mutually support their validity.

What is important to note is that behind all these aspects of validity there is a argumentative connection to the construct and its structure, lending a kind of network character to the notion: “the measure is valid if there is evidence that it fits into the nomological network – the network of predicted relationships with other constructs and their measures” (Frank et al., 2023). We will see below that this is something that is missing in the evaluation practices in NLU, and it is something that I will try to develop here for SLU. (In Figure 1, this is the box on the left.)

Summarising this brief review, to avoid “Questionable Measurement Practices”, Flake and Fried (2020) propose a number of questions to which experiment designers must be able to give a good

<sup>2</sup>This is independent of whether you think that it is a process resulting in a specific psychological state, or a behavioural disposition (Ryle, 1949).

<sup>3</sup>See (Frank et al., 2023; Flake and Fried, 2020; Sireci and Sukin, 2013) for some recent overviews.

answer (Flake and Fried, 2020, p. 459):<sup>4</sup>

1. What is your construct?
2. Why and how did you select your measure?
3. What measure did you use to operationalize the construct?
4. How did you quantify your measure?

The questions shall serve as a guide for the discussion below.

### 3 Current Practices in Measuring NLU

The practice of *benchmarking* in NLP / AI is curiously disconnected from that of measurement in experimental psychology, even if it sets itself what looks like rather closely related goals (for example, to provide a “General Language Understanding Evaluation”, as indicated in the name of the GLUE corpus, (Wang et al., 2019b)).

Evaluation in NLU centers on the *language task*, a functional mapping between input and output, where at least one of these involves language.<sup>5</sup> For a given NLU evaluation corpus, this mapping is typically characterised verbally; e.g., “the text labelled ‘answer’ is a correct answer to the question in the text labelled ‘question’, given the context in the text labelled ‘passage’ ”, as this description could go for the example in Figure 2. It is this verbal (or *intensional*) description that enters into an intuitive appeal to face validity—surely, answering questions must require understanding them. However, it can be remarked that the notion of understanding in NLU evaluation typically remains an intuitive one and no further attempt is made at specifying the construct.

In any case, the actual measurement instrument is one step further removed, as the task needs to be operationalised via *instances* collected into a dataset; this then serves as the *extensional* definition of the task. As observed in (Schlangen, 2021), to not lose connection to the validity argument (which goes via the intensional description) requires care in setting up the dataset, which sometimes is missing. (For example if the collected instances do not span the domain in the way claimed by the intensional description.) For specialised ma-

<sup>4</sup>These are the first four of the six questions they give; the latter ones concern pre-registration, the use of which in NLP would be a topic for another paper and is glossed over here.

<sup>5</sup>The discussion in this section follows Schlangen (2021), which however did not yet use the language of measurement from experimental psychology, however; this connection is helpfully made in (Raji et al., 2021).

<p><b>Passage:</b> Barq’s – Barq’s is an American soft drink. Its brand of root beer is notable for having caffeine. Barq’s, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq’s Famous Olde Tyme Root Beer until 2012.</p> <p><b>Question:</b> is barq’s root beer a pepsi product</p> <p><b>Answer:</b> No</p>
--

Figure 2: An Example of a GLUE-type task (from the BoolQ subset, (Clark et al., 2019), as cited in (Wang et al., 2019a))

chine learning models, a further challenge is posed by the fact that they are typically trained on a (set aside) portion of the dataset. Machine learning methods are very good at identifying predictors that optimise performance, regardless of whether these predictors are related to the construct that is to be measured (Lapuschkin et al., 2019). (Note again that for humans, tests of language understanding happen on the tacit and unquestioned background of already existing general language competence, acquired through material distinct from the testing material.)

With respect to these concerns, practices in NLU evaluation have not much improved. With existing tests saturating when probing newer models, the response has become to go bigger, and efforts such as BigBench (Srivastava et al., 2022) and HELM (Liang et al., 2022) invested in bringing in very many different evaluation sets. While this may be seen as potentially improving convergent validity (if a model achieves high performance on so many tests, it must be doing *something* right), there is still little concern about what exactly the underlying construct is. This we can then take with us to the next section: NLU evaluation centers on language tasks, and relies on the face validity of the task, without making much effort to connect to any further specified construct.

It is also worth noting the criticism of this approach put forward by Raji et al. (2021), which is that the aim of measuring understanding performance in the abstract (without tasks that have extrinsic value beyond their role in the test) through datasets is misguided, conflating as it does a language ability with a recall test on the necessarily open-ended world knowledge that enters into many of these tasks. We will see below how the methodology developed here can answer this reservation, through controlling the world knowledge required to perform. First, we shall look more closely at

how NLU and SLU differ.

#### 4 SLU is Different From NLU

To give us more examples of situated language use, here is another short story:

*You are assembling flat-packed furniture, with the help of your friendly household robot. You send the robot to “fetch the box cutter from the drawer in the other room.”<sup>(1)</sup> “Which one, it’s not in the one with the tools”<sup>(2)</sup>, you hear it shout from the other room. Later, the both of you look over the instructions – why are the pictograms always so obscure? – and discuss how to proceed. Having reached step 24, you look at a screw and wonder whether it is of type 35784, of which there were supposed to be 12 in package A, but the robot just says, “no, the other one”<sup>(3)</sup>. “Alright, so can you pass me the torx?”<sup>(4)</sup>, you say. “Sure, here you go. That’s a torx then?”<sup>(5)</sup>*

This—obviously constructed, but nevertheless hopefully coherent—story showcases several features of situated language use unlikely to be found in monological text corpora: Example (1) contains referring expressions that express an *exophoric reference*—a reference to singular objects outside of the discourse itself, but to its immediate situational context—, and it realises a *request* speech act, for which one sign of understanding is compliance through (non-verbal) action. (2) realises a *clarification request*, which is another way understanding can be signalled, albeit a partial understanding only. This hints at the processual nature of understanding in interaction, different from the single-shot framing as in the example in Figure 2. (3) highlights how the syntax of situated language can be different from the edited written language found in NLU corpora; it is a “*non-sentential*” or “*fragmental*” utterance of a kind which is frequent in dialogue and not at all syntactically or semantically malformed (Schlangen and Lascarides, 2002; Fernández and Ginzburg, 2002). It also shows that the acts that are to be understood need not be linguistic ones; here, the fragment itself is a reaction to a presumed mental state. (4) again is a request for action, this time in the guise of a question as an *indirect speech act*. (5) finally shows that an outcome of understanding in situated interaction can be that understanding

itself can be *adapted*—here, we would expect an agent that has real understanding to be able to later use the term that at that point was new to them.

What this short example has shown, when contrasted with the example in Figure 2, is that SLU differs both on the side of the “input” (the act that is to be understood) as well as on the side of the “output”, where the action space is much larger—in fact, infinite (but compositional). That is, SLU poses language tasks that do not occur in the text corpora used in NLU research. Even more importantly, the individual acts of understanding (from one turn to the next) are embedded in the general goal-directed structure of the interaction as a whole; something that cannot be captured in the i.i.d. (independent and identically distributed) nature of a static dataset. This argues for finding a measurement instrument that provides not only richer context information in a static way (as could be recorded in a richer dataset), but also an active embedding of language use in varying, goal-directed interactions.

#### 5 SLU Requires Different Benchmarking Methods: Dialogue Games

The scenario described above makes for a good use case—having such a robot would be useful!—but a bad measurement instrument. One reason for that is that it simply is far out of reach of current technology (not just in the language abilities, but also in the physical abilities that it suggests). Any attempts at approximating such abilities with current technologies would require making design choices that are more driven by the specific goal rather than by testing language abilities. This also suggests a second problem, which is that this scenario does not isolate the language abilities well enough to serve as a good test. We hence need more controlled situations in which the situated language use can be modelled, while preserving the goal-orientation exhibited by this scenario, as the structure it provides is, as argued above, a crucial element that is not captured by dataset-based methods.<sup>6</sup>

This discussion motivates the use of what I will call *Dialogue Games* as benchmarking instrument,

<sup>6</sup>This can be seen as a restriction compared to the general phenomenon of Situated Language Use: Not every language use situation must necessarily be understood as goal-directed. However, if interaction episodes generally are seen as having a beginning and an ending (Clark, 1996) and the notion of activity types (Levinson, 1979) is accepted, a broad goal of getting from beginning to ending can be assumed to be active in general.



where:<sup>7,8</sup>

A *Dialogue Game* is a constructed activity with a clear beginning and end, in which *players* attempt to reach a pre-determined *goal state* primarily by means of producing and understanding linguistic material.

It is the goal-orientation and the constructed nature of the activity, as we will see, that makes it possible to target particular aspects of Situated Language Understanding, without conflating understanding with recall performance on general world knowledge.

Before moving on to how such games can be constructed in such a way that they make clear connections to (assumptions about) underlying capabilities, and how they can be organised into a strategic plan for making progress, we need to register a cautionary note from the (long) history of this type of approach. In 1972, Minsky and Papert introduced the notion of “micro-world”, as a way to explore “intelligence” problems in context: “we see solving a problem often as getting to know one’s way around a ‘micro-world’ in which the problem exists” (Minsky and Papert, 1972). The most famous of these micro-worlds is the “blocks-world” of the system SHRDLU (Winograd, 1972)—which would count as a Dialogue Game according to the definition above. SHRDLU seemed to demonstrate what I have called here Situated Language Understanding quite well, but criticism of the approach soon arose, of which the following quote is representative (see also Dreyfus (1981); Marr (1982)):

*SHRDLU performs so glibly only because his domain has been stripped of anything that could ever require genuine wit or understanding. [...] Neglecting the tangled intricacies of everyday life while pursuing a theory of common sense is not like ignoring friction while pursuing the laws of motion; it’s like throwing the baby out with the bathwater. A round frictionless wheel is a good approximation of a real wheel because the deviations are comparatively small and theoretically localized; the blocks-world “approximates” a playroom more as a paper plane approximates a duck. (Haugeland, 1985, p. 190)*

<sup>7</sup>Named of course with a nod to Wittgenstein’s *language games*: “I shall call the whole, consisting of language and the activities into which it is woven, a ‘language game’” (Wittgenstein, 1984 [1953], §7); with another inspiration coming from Levinson’s “activity types” (Levinson, 1979).

<sup>8</sup>Note that this definition is general enough to cover “book a train ticket” or even “interactively instruct agent to summarise a text” under the name “game” as well.

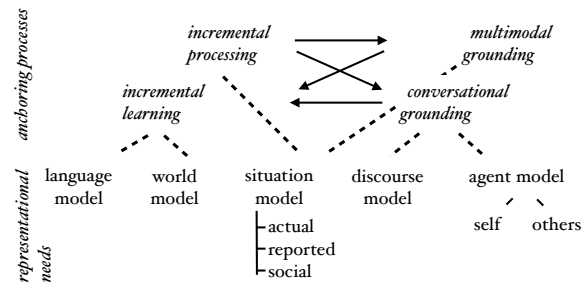


Figure 3: Representational Domains (bottom) and Anchoring Processes (top) Structuring the Situated Agent

This gives us a warning to bring with us to the further discussion, which is to take care that any abstractions made in simulation shall not abstract away the real challenges. We will come to a delineation of the design space in which we can search for Games that meet this challenge in a moment, but first I will sketch a model of the capabilities underlying SLU, to which we can then connect the Game taxonomy.

## 6 The Construct: A Model of Capabilities Involved in SLU

The methodology described above (illustrated in Figure 1) rests on the benchmarking instruments being explicitly grounded in (assumed) capabilities that are being tested. Elsewhere, I have developed a model that distinguishes between various kinds of capabilities involved in SLU (Schlangen, 2023); this will serve us here as the “nomological network of relationships between constructs” from Section 2 above.

This model, illustrated in Figure 3, assumes that the agent represents what I call “knowledge domains”, and maintains “anchoring processes” that operate on them. The knowledge domains are as follows: the *language model* (here meant to collate only linguistic knowledge about the form/meaning mapping; updated rarely), the *world model* (concepts, concept hierarchies, script knowledge, etc.; also updated rarely), the *situation model* (details of the current conversational situation and/or the reported situation; updated continuously), the *discourse model* (what has been said so far, and how it relates; discourse referents; updated continuously), and finally the *agent model* (of the beliefs, desires, intentions of agents, and recursively what it represents of the participating agents; also updated continuously).

As anchoring processes (which “bind the agent to the here, now, and us”), there is *incremental pro-*

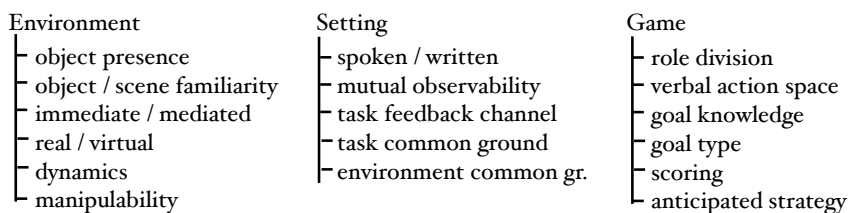


Figure 4: The main components of the proposed taxonomy

cessing (updating situation and agent model, based on minimal units of observation), *conversational grounding* (the process of negotiating shared understanding, for example through asking for clarification, if necessary), *incremental learning* (which ranges from the establishment of spontaneous local conventions, e.g. on how to refer to objects, established during the conversational grounding, to learning facts from observation and from testimony, and, crucially, from discussion and disagreement), and *multimodal grounding* (resolving references to objects in the shared surrounding, as well as deriving meaning from non-verbal actions such as gestures).<sup>9</sup>

What this gives us is a finer-grained picture of the construct: *understanding* means applying, building up and maintaining these representations, via these processes.<sup>10</sup> Not all acts of understanding rely on all aspects equally, and this makes it possible to develop a strategy for working towards modelling the overall capability, as we will see. With this in hand, we can now come to the design aspects of Dialogue Games, and how they may put certain of these capabilities (knowledge domains and anchoring processes) more or less into focus. You can read the following section also as advice on best practices in experiment design, drawn from extensive experience in setting up what is here systematised as Dialogue Game (Fernández et al., 2006; Kennington et al., 2013; Zariëß et al., 2016; Ilinykh et al., 2019; Attari et al., 2019; Schlangen, 2019a).

## 7 A Taxonomy of Game Types

The most salient aspect of a Dialogue Game might be the task that it poses to the players; that is, the goal state and how to get there. For what could

<sup>9</sup>For a more detailed description and a justification of this way of analysing the SLU agent, including references to prior work making use of related concepts, see the original paper (Schlangen, 2023).

<sup>10</sup>Let me stress again that I am not making any claims about whether such representations should be built into models of situated language understanders or not; the (falsifiable) claim is just that something like them will be found in such models.

be called the “furniture assembly game” from Section 4, this would be *have the furniture fully assembled* (goal state) and *provide required assistance* (game “play”); for a more realistic Dialogue Game (of the type “reference game”, see below), that could be *bring cards into same order* (goal state) and *ask each other which cards there are, and jointly decide on order* (game play). But hidden behind descriptions like these there is a large number of additional design decisions that need to be made before the game can be played. These decisions have many degrees of freedom, but all come with subtle influences on the shape of the interaction, and on the phenomena that one can expect to see in protocols of the game play.

I distinguish here between three major areas in which decisions must be made when setting up a concrete Dialogue Game: Environment, Setting, Game Proper; with many sub-aspects within. A comprehensive overview is given in Appendix A. The discussion here presents these design features and directly links them to the model of the underlying construct(s) from the previous section.

### 7.1 Environment

This section groups together all design decisions that influence **what the relevant entities and actions in the game are, and how they are presented to the players.**

A high-level decision here is whether the game requires talking about **objects that are currently present** (in some form), or not. (An example of a task that is not about currently present objects would be booking a train ticket, which does require talking about entities such as train stations, without them needing to be present. Such a conversation is still situated in the sense that the interlocutors share time, but it is at the boundaries of what I consider here.) This decision influences how the *situation model* is constructed (e.g., from visual evidence or not) and how the *world model* is challenged (because for non-present objects, agreement on referents must come from prior common ground). A different design dimension concerns

**prior knowledge of these objects**, whether they are (expected to be) **familiar** to the players or not. Whether something is familiar or not depends on the *world model*, and whether it is assumed to be *mutually* familiar on the *agent model*; successfully referring to unfamiliar objects means more effort in *conversational grounding*.

Another decision is whether **access** of the players to the objects is **immediate or mediated**, and if mediated, if the **objects are real or computer simulated**. (So, a video call would be mediated but real; operating with representations on a computer screen would be mediated and virtual.) Virtual environments make further abstractions possible, for example by discretising changes (the world “jumps” from one state to the next), or reducing the action space (what can be done to and with objects). The difference here is less one in what capabilities are challenged than in the control that is given over the situation; for example, a **static** environment will force fewer updates to the *situation model* as one with **discrete** updates, which in turn may require slower changes than one that is fully **dynamic**.

## 7.2 Setting

This dimension collects decisions about **how the players can interact with each other**—which of course determines to a large extent what kind of data can be expected.

A first high-level decision here is whether the verbal interaction is done via **speech**, or through **typed messages**. Written language, even in the dynamic form that it can take in chat interactions, is a restricted channel compared to spoken language (where prosody and other para-linguistic information provides a channel for *multimodal grounding*); the interaction also slows down and is, at least in typical setups, more discretised (as messages need to be sent before they are seen; this influences the degree to which *incremental processing* is challenged). Finally, turn-taking, which is an essential process in the organisation of free interaction (de Ruyter et al., 2006), works differently in chat communication than in spoken interaction. On the other hand, practical advantages in choosing typed messages are also clear, in that written language is typically easier to store, and (for artificial agents) to process, and to generate.

Another set of decisions concern **what the players see of each other**, and **what they see of their actions** in the environment. *Multimodal ground-*

*ing* of the signal in actions of the interlocutor is an important aspect of meaning making (Holler and Levinson, 2019); disabling it through hiding the interlocutor forces more meaning into the verbal channel (which can be desired, but reduces ecological validity). Similarly, other aspects of interaction management get harder when there is no visual contact between interlocutors (Brennan, 2000), but at same time become more visible in the linguistic material.

I have also grouped under this heading questions of how **common ground** between the participants (other than what they can see of each other) can form. If the players knowingly play repeated rounds of the game (from some initial state to a respective goal state), they can build up personal common ground (Clark, 1996), a form of *incremental learning* influencing their *agent model*.<sup>11</sup> When players (knowingly) **share the same environment** (be that a simulated and mediated one or a real one; where even looking at the same image would count as sharing the environment), there is an automatic assumption that large parts of the respective *situation models* are shared (and represented as thus in the *agent model*); if this is not the case, or not knowingly so, linguistic labor must be performed to reach such common ground (if the tasks requires it).

## 7.3 Game

The decisions grouped here concern the game in the narrow sense: **how initial state and goal state are defined**, but also **what the players know about this**, and **how the games defines roles and suggests strategies**. In the terminology of Suits (1978), a game must subordinate under a *preludatory goal*, which can be stated independently of the game (e.g., in football (soccer), “make the ball be in the opponent team’s goal”); it is further defined by *constitutive rules*, which make reaching that goal more difficult than necessary (e.g., by disallowing to just grab the ball and carry it to the goal); it must also trigger in the players a *lusory attitude*, which is the acceptance of the complications posed by the constitutive rules. Doing the latter successfully can increase the quality of the collected data, as players with higher engagement can be expected to show a wider range of behaviours (von Ahn and Dabbish, 2004). Inspiration can be taken here from the lit-

<sup>11</sup>These days in Artificial Intelligence more typically called “theory of mind”, see e.g. (Bara et al., 2021).

erature on the design of games (e.g., (Adams and Dormans, 2012)); ultimately, however, the purpose of a Dialogue Game in the sense developed here is to provide data and a testing environment in a principled way, and not primarily enjoyment.

To classify games, one high-level aspect concerns the **goal type**, where we can distinguish **reference games** (a time honoured instrument in Psycholinguistics, going back at least to (Krauss and Weinheimer, 1964); see (Ji et al., 2022) for a recent overview), which focus on reference and hence *multimodal grounding* and alignment of *situation models*; **information games**, which center on the requesting and giving of information, which depending on the domain can lead to demands on the *world model* and/or the *situation model*; **construction games**, which go beyond reference and information in that they require the execution of actions, and hence require coordination of the anchoring processes to a higher degree; **navigation games**, which center spatial language and spatially complex *situation models*; **negotiation games**, which focus on explicit coordination of *agent models*; and finally **teaching games**, which make explicit the *incremental learning* and how it updates the *world model*. This is not a complete categorisation, and each concrete game will contain elements of more than one of these types; but this does represent good coverage of types of games typically used. (We are currently preparing a comprehensive survey of the field, which will provide a plethora of references for representatives of all types.)

Some more final subdimensions. In the design of the game, the player can be assigned **distinct roles** with different responsibilities, such as for example an assigned *questioner* paired with an assigned *answerer*, or an *instruction giver* with an *instruction follower*. The stricter these roles are, the lower the coordination effort required, deemphasising functions such as *conversational grounding* and the keeping of detailed *agent models*. Goal-directed games naturally come with a notion of **success**, but beyond that, **scoring** functions can be introduced (for example, “faster is better”; or a reconstruction loss for construction-instruction tasks). Making the score known to players introduces incentives that can change the dynamics of the interaction (e.g., prioritizing speed over accuracy, or vice versa).

Finally, the design of the game can also make a desired **strategic behaviour** more salient. A **cooperative** player would be one who does their

best in understanding intents behind requests (e.g., through replying correctly to indirect speech acts, or to providing partial information when a question cannot be answered fully), whereas a **collaborative** player is one who takes their goal to be shared with the other player, and who hence has an interest in being proactive as well—likely to be more challenging to the anchoring processes and to the alignment of the *agent models*.<sup>12</sup>

## 8 Dialogue Games as Evaluation Instrument

Let us assume that we now have designed a Dialogue Game, starting from ideas about which aspects of the construct we particularly want to challenge and making careful decisions on all the design features mentioned above. Of the questions listed above in Section 2, we have an answer to numbers 1 to 3 (Q1: What is the construct?—A: The model in Section 6); Q2,3: Why and how did we select measure? What measure to operationalise?—A: By making a decision to focus on some aspects, and selecting according to Section 7). This leaves one crucial element, Q4: Deciding on how to quantify the measure. This for us translates into how to use the dialogue game to quantify the abilities of an artificial model, which is what this section will look into.

Dialogue games can be used as a means for data generation, simply by collecting game play from people playing the game. Using a dialogue game promises to offer some control over the data that is to be expected, insofar as the connection between properties of the game to underlying capabilities (as discussed in the previous section) is also reflected in properties of the language use. For example, a reference game will make the use of referring expressions likely; a game where mutual understanding is particularly forwarded will make linguistic devices for conversational grounding prominent (Schlangen, 2019a). This can be interesting for the study of these linguistic phenomena (see, e.g.,

<sup>12</sup>We note here that a demand for cooperation can lead to increased coordination effort, which can result in the players negotiating in the game to follow a merely *cooperative* strategy (with one instruction giver and one instruction follower), if this seems more efficient to them. This is something that we have experienced with the MeetUp game (Ilinykh et al., 2019), where two players moving in separate copies of the same virtual environment must manage to meet up in the same room (without seeing each other), and which we had designed to trigger collaborative interactions; it however turned out to be a frequent strategy for one player to just stop moving and only answer questions by the other.



(Fernández et al., 2006; Schlangen and Fernández, 2007)). Moreover, the control over environment and setting makes it possible to record rich contextual information alongside with the language use (Kousidis et al., 2012).

This does not mean, however, that the use ends with the recording of richer corpora, to then be used in the same way as the NLU corpora mentioned above. In fact, as we have touched on above, using static corpora for research on SLU is problematic, as here the recorded actions can count even less as reference than they do in other generative tasks, such as machine translation.<sup>13</sup>

In this section, I will highlight some uses that go beyond the train/val/test dataset paradigm typical of NLP, all revolving around the ability of Dialogue Games—at least those with simulated environments—to serve as *execution environments*.

**Rollout** One such use, which in a way stands between the use of static corpora and the evaluation of agents in interactive game play, is nicely exemplified by the benchmarks built on the TEACH dataset (Padmakumar et al., 2022). The game, according to our taxonomy, uses a simulated, dynamic environment with familiar objects (household objects), is a Navigation Game with elements of an Information Game (as a Commander, for whom the environment is fully observable, instructs via a written channel a Follower, for whom the environment is only partially observable, to perform a task in the environment, with the Follower getting an opportunity to ask for clarification). Two tasks are defined that one may call *rollout* tasks (our terminology), in the sense that they require the prediction of actions, based on partial or complete history. Crucially, since the environment simulator is provided, the predicted actions can be executed, and the evaluation target is whether the required state changes have been affected. This abstracts away at least to a certain extent from what is recorded in the corpus, as only the state changes (and not all actions) serve as reference. This type of evaluation is only possible in a Dialogue Game setup and not with a static corpus alone.

**Agent/Agent Play** If artificial agents for all roles in the game are provided, another mode of evaluation comes available, that of fully simulated play.

<sup>13</sup>Where metrics that compare model predictions against a reference have long been criticised, see *inter alia* (Turian et al., 2003); see (Liu et al., 2016) for an early extension of this critique to the evaluation of “dialogue systems”.

Games as defined above will come with a some sort of score that measures success in reaching the pre-defined goal state, and this can then serve as the evaluation target. If a human/human reference corpus is available, the produced language itself can then furthermore be evaluated along formal parameters (e.g., average turn length and distribution, vocabulary size, etc.).

Note however that the agent/agent mode leads to a even further deviation from the “test is like training” approach described above for NLU, as, unlike in non-communicative game like arcade-type games—one of the early successes of new-generation reinforcement learning, e.g. (Mnih et al., 2013)—this setup can here not be used for learning the agents: In the language case, a competent player needs to already exist from which competent language use can be learned.<sup>14</sup>

We note that Dialogue Games (in our sense, as combination of environment, setting, and game) provide an interesting perspective for the evaluation of (supposedly) general-purpose “foundation models” (Bommasani et al., 2021), which have been claimed to be able to function as general simulators of agents (Andreas, 2022).

**Human/Agent Play** The most informative evaluation mode of interactive system remains evaluation in actual online interaction with human players.<sup>15</sup> Beyond the measures defined by the game (measuring task parameters) and other such measures, which in the dialogue systems community usually are called *objective measures* such as dialogue length, etc. (see (Walker et al., 1998) for a seminal reference), this mode also makes it possible to evaluate the interactive *experience*, through *phenomenological* or *subjective measures* elicited in questionnaires (Kocaballi et al., 2019).<sup>16</sup>

One possible objection against this evaluation mode is quickly dismissed: While there may be

<sup>14</sup>A problem that has led to the research area of “language emergence” in what could be called Dialogue Games between deep reinforcement agents, where however the agents are allowed to coordinate on their own language system, folding the language evolution and language learning problem into one. (See e.g., (Lazaridou et al., 2017).) We concentrate here on the setting of evaluating agents that have acquired (to the extent needed for the Game) an existing natural language.

<sup>15</sup>Compare to what is called “human evaluation” in many fields of NLP (see (Howcroft et al., 2020) for recent critical overview of human evaluation practices in Natural Language Generation).

<sup>16</sup>Although not many fully validated scales exist; a popular one is Goodspeed questionnaire from the neighbouring field of Human/Robot interaction (Bartneck et al., 2009).

Environment	Setting	Game
<ul style="list-style-type: none"> <li>• present y - n</li> <li>• familiar y - n</li> <li>• real &gt; simulated</li> <li>• high fidelity - low</li> <li>• dynamic &gt; static</li> </ul>	<ul style="list-style-type: none"> <li>• spoken &gt; typed</li> <li>• embodiment y &gt; n</li> <li>• repeated y &gt; n</li> <li>• view shared - part - diff</li> </ul>	<ul style="list-style-type: none"> <li>• role equality &gt; div.</li> <li>• action space unrestr. &gt; restr.</li> <li>• symmetry &gt; asymmetry</li> <li>• negot. - instr. foll. &gt; inf. &gt; ref.</li> <li>• collab. &gt; coop. &gt; control</li> </ul>

Figure 5: A partial order on the space of Dialogue Games.  $\sim$  denotes “similar complexity”,  $>$  denotes “leading to higher complexity”.

a superficial similarity to Turing’s imitation game (Turing, 1950)—what is now known as the “Turing Test”—in that quality of a conversation is an evaluation criterion, it is important to note that *deception* about whether a player is human or machine need not be, and in most cases is not, part of the evaluation; and this is what is commonly criticised (see e.g., (Levesque, 2014)). To keep clear of the deception incentive, however, it is important that such subjective measures do not on their own become optimization targets and are always combined with objective, task-oriented measures (Edlund et al., 2008).

**Representation Probing** Another way to gain information about a game-playing model is via *representation probing*. To cite just one example, Madureira and Schlangen (2022) use existing dialogue models as what could be called “overhearers” of dialogues from a corpus, and probed whether they, at a given state of the overheard dialogue, represented information about the assumed common ground status of propositions. Techniques such as these can help further validate claims about the link between capabilities and games (if the assumption is that the game challenges a certain capability, we would expect to find information that it is based on to be represented during the model processing).

Let us take stock. Our long journey has taken us from an argument that Situated Language Understanding needs to be evaluated in the context of, well, situated interactions, through the claim that the underlying construct is multi-faceted, to a recipe for constructing measurement instruments—Dialogue Games. For a given Dialogue Game, the recipe provides at least the beginnings of a validity argument, through the links between taxonomy and construct. We also now know that there is a variety of ways of making use of the instrument and deriving from it a quantified measure, which

in turn facilitates (with the usual caveats) a comparison between models. (“Bigger is better.”) The introduction promised more, however, namely the derivation from this of a general strategy that might get us, eventually, from simple games to scenarios like the robo-helper described above. This is what the next section will discuss.

## 9 Strategy: From Simple to Complex Dialogue Games

In this paper, I have tried to use insights from assessment in experimental psychology to suggest improvements in practices in assessment in Artificial Intelligence. There is a point, however, at which the similarities end. AI, as a constructive discipline, aims to *build* the artefacts it studies; experimental psychology aims to understand existing (biological) systems. I specifically made the point in the introduction that situated interactions of the kinds discussed here come easy to most people; as assessments of understanding abilities for people, Dialogue Games would not have much purchase.<sup>17</sup> On the other hand, if we look at the current state of the field, it is clear that we are still at the relatively low end of game complexity. To pick two examples, “visual dialogue” (Das et al., 2017) represents perhaps one of the simplest game types: player A sees an image and a caption, player B only the caption; player B asks 10 questions about the image—without any further goal—which player A must answer). The TEACH benchmark that was mentioned above (Padmakumar et al., 2022) still relies on a relatively simple game (the tasks are things like “fetch a potato” or “put all plates on the table”), and still, the performance of even the best models is quite modest. The historical development that has led to these games is easy to reconstruct and

<sup>17</sup>Perhaps more so when players are restricted to what for them is not their first language; but we leave this unexplored here.

leads to them from non-interactive tasks (image captioning, natural language navigation), extended to the “adjacent possible”.

The taxonomy described here offers a view towards what is not only adjacent and possible, but also “uphill”. To make a given game more complex, a simple parameter is to increase the variability of the entities that it involves. Then, restrictions can be removed successively, e.g. by going from a static to a dynamic environment, from typed to spoken interaction, and so on. (Figure 5 suggests a partial complexity order by ranking possible feature values.) Following the discussion above, many of these changes will also shift or extend the way the game challenges understanding, and a model capable of this change thus shows itself to reach a higher level. In this way, the taxonomy also suggests a way to construct a meta-benchmark on which (hypothetical) general models of SLU can be measured. (Among two models that perform comparable on one game, that one will rank higher that performs better on a more complex task.)

## 10 Related Work

What I call Dialogue Games here has been used for a long time as instrument in driving forward research on situated language modelling. This is not the contribution of this paper—there is a rich, and ever more strongly growing, literature making use of such games (see (Duan et al., 2022; Gu et al., 2022; Sundar and Heck, 2022); and our forthcoming general survey). What there is less work on is on this instrument itself. Bisk et al. (2020) make general points about types of information the availability of which during learning might improve the “understanding” of AI models. Fried et al. (2022) cover similar ground to this paper, but more generally focus on the kinds of contexts needed for certain pragmatic phenomena.

There are now several simulation environments that support setting up Dialogue Games in simulated, continuous environments with high fidelity (Gu et al., 2022). In my research group, we have focussed on the development of a flexible environment that makes the implementation of Dialogue Games and the collection of game play for example through crowd sourcing easier (Götze et al., 2022; Schlangen et al., 2018).

For a related argument for how SLU differs from a monological perspective, taking a wider Cognitive Science perspective, see (Dingemane et al.,

2023).

## 11 Conclusions

After briefly reviewing how experimental psychology thinks about the validity of assessments, I reviewed practices of evaluation in NLU, in the light of these considerations of validity. This served as the foil on which to develop a guide for evaluating what I call *Situated Language Understanding* (SLU). I argued that SLU is different from NLU (Section 4) and hence requires different evaluation instruments: *Dialogue Games* (Section 5). As an element in the argument for the validity of this instrument, I reviewed a model of the capabilities involved in SLU, that is, of the internal structure of the construct that is to be measured (Section 6). This was then followed by the description of a detailed taxonomy of Dialogue Game *types* (Section 7), which can be read as a guide for constructing a particular game in such a way that it can serve as an assessment instrument focussing on particular aspects of the construct. Different ways of using a given game for evaluation were then discussed in Section 8, before Section 9 brought together these insights into a discussion of how this suggests a ordering of assessments from simpler to more complex, thereby suggesting a possible development strategy for models in this space.

## Ethics Statement

Let us address the ethical elephant in the room. Should we even attempt to build systems that can do this kind of situated language understanding? Should research be conducted on increasing the complexity of the tasks in which they can be used, as indicated in Section 9? It should be clear that there are potential enormously beneficial use cases, for example where such systems are used to restore the physical reach of humans that have lost abilities (or never had them). But understanding of the kind discussed here shows in action, making systematic failures or biases of such systems potentially more directly harmful than for language-only systems: What holds for language models holds even more for models with arms. The hope is that the methodology described here of starting with simpler settings *and thoroughly evaluating performance on them* before moving on might provide one way in which this research can be made safer—although of course further work is needed on working out whether this argument holds water.

## References

- Ernest Adams and Joris Dormans. 2012. *Game Mechanics: Advanced Game Design*. New Riders Games.
- Jacob Andreas. 2022. [Language models as agent models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nazia Attari, Martin Heckmann, and David Schlangen. 2019. From explainability to explanation: Using a dialogue setting to elicit annotations with justifications. In *Proceedings of SIGdial 2019, Short Papers*, Stockholm, Sweden.
- Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. [MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. [Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots](#). *International Journal of Social Robotics*, 1(1):71–81.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. [Experience grounds language](#). *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 8718–8735.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avatika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R’e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. [On the opportunities and risks of foundation models](#). *ArXiv*.
- Susan E. Brennan. 2000. Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, China.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.
- Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. [Learning cooperative visual dialog agents with deep reinforcement learning](#). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979.
- J.P. de Ruyter, H. Mitterer, and N.J. Enfield. 2006. Projecting the end of a speaker’s turn: a cognitive cornerstone of conversation. *Language*, 82(3):504–524.
- Mark Dingemans, Andreas Liesenfeld, Marlou Rasenberg, Saul Albert, Felix K. Ameka, Abeba Birhane, Dimitris Bolis, Justine Cassell, Rebecca Clift, Elena Cuffari, Hanne De Jaegher, Catarina Dutilh Novaes, N. J. Enfield, Riccardo Fusaroli, Eleni Gregoromichelaki, Edwin Hutchins, Ivana Konvalinka, Damian Milton, Joanna Rączaszek-Leonardi, Vasudevi Reddy, Federico Rossano, David Schlangen, Johanna Seibt, Elizabeth Stokoe, Lucy Suchman, Cordula Vesper, Thalia Wheatley, and Martina Wiltschko. 2023. [Beyond single-mindedness: A figure-ground reversal for the cognitive sciences](#). *Cognitive Science*, 47(1):e13230.
- Hubert L. Dreyfus. 1981. From micro-worlds to knowledge: AI at an impasse. In John Haugeland, editor, *Mind Design*. MIT Press.



- Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. [A Survey of Embodied AI: From Simulators to Research Tasks](#). *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.
- Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50:630–645.
- Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, USA. ACL Special Interest Group on Dialog.
- Raquel Fernández, Tatjana Lucht, Kepa Rodríguez, and David Schlangen. 2006. [Interaction in task-oriented human–human dialogue: The effects of different turn-taking policies](#). In *Proceedings of the First International IEEE/ACL Workshop on Spoken Language Technology*, Palm Beach, Aruba.
- Jessica Kay Flake and Eiko I. Fried. 2020. [Measurement Schmeasurement : Questionable Measurement Practices and How to Avoid Them](#). *Advances in Methods and Practices in Psychological Science*, 3(4):456–465.
- Michael C. Frank, Mika Braginsky, Julie Cachia, Nicholas Coles, Tom Hardwicke, Robert Hawkins, Maya Mathur, and Rondeline Williams. 2023. [Experimentology: An open science approach to experimental psychology methods](#). Website.
- Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. [Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches](#). *CoRR*, abs/2211.08371.
- Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen. 2022. [The slurk interaction server framework: Better data for better dialog models](#). In *Proceedings of the Language Resources and Evaluation Conference*, pages 4069–4078, Marseille, France. European Language Resources Association.
- Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. 2022. [Vision-and-language navigation: A survey of tasks, methods, and future directions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, Dublin, Ireland. Association for Computational Linguistics.
- John Haugeland. 1985. *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge, Mass.
- Judith Holler and Stephen C. Levinson. 2019. [Multi-modal Language Processing in Human Communication](#). *Trends in Cognitive Sciences*, pages 1–14.
- David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. [Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Nikolai Ilinykh, Sina Zarriß, and David Schlangen. 2019. [Meetup! a corpus of joint activity dialogues in a visual environment](#). In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2019 / LondonLogue)*, London, UK.
- Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. 2022. [Abstract visual reasoning with tangram shapes](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 582–601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. [Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information](#). In *Proceedings of the SIGDIAL 2013 Conference*, pages 173–182, Metz, France. Association for Computational Linguistics.
- Ahmet Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. 2019. [Understanding and Measuring User Experience in Conversational Interfaces](#). *Interacting with Computers*, 31(2):192–207.
- Spyros Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. 2012. [Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data](#). In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog at Interspeech 2012*, pages 39–42, Stevenson, WA, USA.
- Robert M. Krauss and Sidney Weinheimer. 1964. [Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study](#). *Psychonomic Science*, 1:266–278.
- Sebastian Lopuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. 2019. [Unmasking Clever Hans predictors and assessing what machines really learn](#). *Nature Communications*, 10(1):1–8.
- Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. [Multi-agent cooperation and the emergence of \(natural\) language](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. OpenReview.net.
- Hector J. Levesque. 2014. [On our best behaviour](#). *Artificial Intelligence*, 212(1):27–35.

- Stephen C. Levinson. 1979. Activity types and language. *Linguistics*, 17:365–399.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yükekşönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khatib, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. [Holistic evaluation of language models](#). *CoRR*, abs/2211.09110.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Brielen Madureira and David Schlangen. 2022. [Can visual dialogue models do scorekeeping? exploring how dialogue representations incrementally encode shared knowledge](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 651–664, Dublin, Ireland. Association for Computational Linguistics.
- David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco, USA.
- Marvin Minsky and Seymour Papert. 1972. Progress Report on Artificial intelligence. Technical report, MIT Artificial Intelligence Laboratory, Cambridge, Mass., USA.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. [Playing atari with deep reinforcement learning](#). *CoRR*, abs/1312.5602.
- Aishwarya Padmakumar, Jesse Thomason, Ayush Srivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. TEACH: Task-Driven Embodied Agents That Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2017–2025.
- Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.
- Gilbert Ryle. 1949. *The Concept of Mind*. Hutchinson & Co.
- David Schlangen. 2019a. [Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings](#). *CoRR*, abs/1908.11279.
- David Schlangen. 2019b. [Language tasks and language games: On methodology in current natural language processing research](#). *CoRR*, abs/1908.10747.
- David Schlangen. 2021. [Targeting the benchmark: On methodology in current natural language processing research](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.
- David Schlangen. 2023. [What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis](#). *CoRR*, abs/2302.08590.
- David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarriß. 2018. slurk – A Lightweight Interaction Server For Dialogue Experiments and Data Collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / semdial 2018)*.
- David Schlangen and Raquel Fernández. 2007. Speaking through a noisy channel - experiments on inducing clarification behaviour in human-human dialogue. In *Proceedings of Interspeech 2007*, Antwerp, Belgium.
- David Schlangen and Alex Lascarides. 2002. [Resolving fragments using discourse information](#). In *Proceedings of the 6th International Workshop on Formal Semantics and Pragmatics of Dialogue (EDIALOG 2002)*, pages 161–168, Edinburgh.
- Stephen G. Sireci and Tia Sukin. 2013. Test Validity. In K. F. Geisinger, editor, *APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 4. The American Psychological Association.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazary, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askill, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas

Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Ghohlamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *CoRR*, abs/2206.04615.

Bernard Suits. 1978. *The Grasshopper: Games, Life, and Utopia*. The University of Toronto Press, Toronto, Canada.

Anirudh Sundar and Larry Heck. 2022. [Multimodal conversational AI: A survey of datasets and approaches](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. [Evaluation of machine translation and its evaluation](#). In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59:433–460.

Luis von Ahn and Laura Dabbish. 2004. [Labeling images with a computer game](#). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA. ACM.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, 12(3).

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. [SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems](#). In *NeurIPS*, July, pages 1–30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. [GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding](#). In *ICLR 2019*, pages 1–20.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Ludwig Wittgenstein. 1984 [1953]. *Tractatus Logicus Philosophicus und Philosophische Untersuchungen*, volume 1 of *Werkausgabe*. Suhrkamp, Frankfurt am Main.

Sina Zarrieß, Julian Hough, Casey Kennington, Rames Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. Pentoref: A corpus of spoken references in task-oriented dialogues. In *Proceedings of LREC 2016*, Portoroz, Slovenia.

## A Example Appendix

Table 1 shows the full taxonomy, with explanations of all attributes and values.

Category / Feature	Possible Values	Description
<b>Environment</b>		Characterises the relevant objects and configurations, and how they are presented to players
presence	ready-to-hand / absent	are players talking about objects that are immediately perceivable to them or not?
object familiarity	instance familiar / type familiar / unfamiliar	prior to game, do they know object instance (Barack Obama), type (fridge), or likely not at all (pento pieces)
scene familiarity	instance familiar / type familiar / unfamiliar	same for constellations of objects
access	immediate / mediated	are objects physically present or via interface?
reality	real / virtual	are objects real or computer represented?
fidelity	high / low	realism of representation
dynamics	continuous / discrete / static	how changes of the environment proceed
env action space	object manipulation / viewpoint man. / none	what can be done to environment
<b>Setting</b>		Characterises how the players can interact with each other
verbal action channel	spoken / written	w/ variations on turn taking, e.g. "free turn taking", "push to talk", "turn-based", "round-robin"
mutual observability	real / avatar / none	whether other player is visible / embodied
task action channel	in-environment / symbolic feedback / none	how actions of other player are perceived (symbolic feedback would be e.g. just info whether they picked the right object, w/o player seeing the picking action)
task common ground	single game / repeated games	whether players (knowingly) play repeated rounds
env. common ground	full / partial / none	whether players are in same environment or not
<b>Game</b>		Characterises the goal of the interaction and the constraints on how to reach it
role equality	equal / specialised / sequentially-equal	do players have the same action space or not (e.g., instruction giver / follower); "sequentially-equal" meaning they swap roles
verbal action space	unrestricted / restricted	whether they can talk freely or are limited to range of utterances (e.g., just "yes" or "no"); by player
goal information	symmetric / asym. / complementary / none	whether one player has solution, or both, or both have (different) parts; none means neither player knows more than general goal spec (e.g., like in chess)
goal type	(games can contain several goal types)	
	reference	identify object(s)
	information	request / provide information
	construction	configure objects
	navigation	go somewhere / direct somewhere
	negotiation	agree on something
	teaching	teach / learn something
scoring	binary / graded / none	measure of immediate task success (Interaction as a whole might additionally be evaluated otherwise as well)
score impact	yes / no	whether players are motivated to achieve good score
anticipated strategy	cooperative / collaborative / anti-collaborative	by player; whether player is expected to facilitate other player's goals, or has own goals which coincide (or not) w/ other player and for which other player is needed

Table 1: The proposed fine-grained classification scheme