PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Review



Cite this article: Wahlster W. 2023 Understanding computational dialogue understanding. *Phil. Trans. R. Soc. A* **381**: 20220049. https://doi.org/10.1098/rsta.2022.0049

Received: 6 February 2023 Accepted: 20 March 2023

One contribution of 10 to a discussion meeting issue 'Cognitive artificial intelligence'.

Subject Areas:

artificial intelligence human-computer, interaction

Keywords:

dialogue systems, human-like understanding, multi-modality, chatbots, large language models, anticipation feedback

Author for correspondence:

Wolfgang Wahlster e-mail: wolfgang.wahlster@dfki.de

Understanding computational dialogue understanding

Wolfgang Wahlster

German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91c, D-10559 Berlin, Germany

(D) WW, 0000-0002-8785-5567

In this paper, we first explain why human-like dialogue understanding is so difficult for artificial intelligence (AI). We discuss various methods for testing the understanding capabilities of dialogue systems. Our review of the development of dialogue systems over five decades focuses on the transition from closed-domain to open-domain systems and their extension to multi-modal, multi-party and multilingual dialogues. From being somewhat of a niche topic of AI research for the first 40 years, it has made newspaper headlines in recent years and is now being discussed by political leaders at events such as the World Economic Forum in Davos. We ask whether large language models are superparrots or a milestone towards human-like dialogue understanding and how they relate to what we know about language processing in the human brain. Using ChatGPT as an example, we present some limitations of this approach to dialogue systems. Finally, we present some lessons learned from our 40 years of research in this field about system architecture principles: symmetric multi-modality, no presentation without representation and anticipation feedback loops. We conclude with a discussion of grand challenges such as satisfying conversational maxims and the European Language Equality Act through massive digital multi-linguality-perhaps enabled by interactive machine learning with human trainers.

This article is part of a discussion meeting issue 'Cognitive artificial intelligence'.

1. Introduction

Social interaction based on natural language dialogue is one of the defining characteristics of our species. Dialogues are written or spoken conversational exchanges between two or more agents consisting of a coherent sequence of dialogue acts. The computer simulation of human dialogue behaviour is one of the most ambitious scientific goals of this millennium. It requires multi-disciplinary collaboration of experts in artificial intelligence (AI), computational linguistics (CL) and cognitive science.

(a) Human-like dialogue understanding is Al-complete

During the past 50 years, our understanding and simulations of the complex processes underlying human dialogue behaviour have become more and more precise, so that today the ancient dream of mankind, to be able to converse with things, is coming into reach.

However, human dialogue understanding is not an isolated capacity, but is embedded into many other aspects of cognition. This means that many subfields of AI beyond natural language understanding are relevant for achieving human-like dialogue behaviour: vision, reasoning, planning, learning, plan recognition, user modelling, knowledge representation and robotics [1].

Therefore, we claim that dialogue understanding is AI-complete. Although a formal and broadly accepted definition of AI-completeness does not yet exist, we strongly believe that an AI system cannot achieve human-like dialogue behaviour without having many other cognitive competences that are the basis for human intelligence. Our 40 years of experience in the realization of dialogue systems has shown that, compared to human-like behaviour, our systems were often limited by the lack of certain cognitive skills outside the narrower field of language processing, such as sensing, reasoning, planning and acting.

This suggests a multi-disciplinary approach to cognitive AI for dialogue systems, which uses theoretical claims and experimental insights from various neighbouring disciplines as architectural constraints for computational models. These models generate new predictions that can be tested by experiments. The resulting data can be used for new predictions and new insights. During the past five decades of research about dialogue systems, results from the following fields have had a significant impact on the AI-driven design of dialogue systems: discourse research, interaction research, social psychology, emotion research, psycholinguistics, neurolinguistics, neuropsychology and general cognitive science.

(b) Why is human-like dialogue understanding so difficult for artificial intelligence?

In the computer modelling of dialogue comprehension, the step-by-step reduction of uncertainties in the interpretation of utterances represents the greatest challenge (figure 1). This is already evident in the mapping of sound waves to words, since there is no one-to-one function for this. For speaker-independent continuous speech recognition, the naïve approach of simply creating a database of all spectra for all words of a language and then performing pattern matching with the spectra of all words of the same language is not possible (see chapter 16 of [2]). One and the same word, even if spoken by the same person, can be pronounced differently, in different pitches and at different speeds, depending, for example, on the emotional state of the speaker and the situation. In addition, there are large differences in pronunciation from person to person, depending on gender, age and the mother tongue of the interlocutor. This high phonetic variability of spoken language utterances must therefore be taken into account in statistical and neural models. A complicating factor is that many people speak a dialect, resulting in a multitude of further pronunciation variants that lead to interpretation uncertainties.

However, the greatest uncertainty for dialogue understanding systems arises from the many types of ambiguity in natural language. For example, without further context, a statement such as 'We were on the phone with friends in Japan' remains ambiguous, whether it means that we telephoned in Japan with some friends or whether we called friends in Japan from another country. Such ambiguity in the binding of a prepositional phrase like 'in Japan' is a typical form of syntactic ambiguity that can only be resolved by consulting other sources of knowledge in dialogue understanding. Particularly difficult and usually impossible without additional knowledge of the world is the resolution of semantic ambiguities in the area of natural language quantifiers and their scope, as is evident in the utterance 'All students own a computer', where it



Figure 1. Uncertainty reduction in dialogue: the cylinders symbolize models as data sources that are used on the three processing levels. The triangle in the middle symbolizes the funnel created by the incremental reduction of interpretation hypotheses outlined on the right.

is unclear whether a single computer belongs to all students or whether each student has his or her own computer.

Since alternative interpretations accumulate through ambiguities across all levels of processing, there is the danger of a combinatorial explosion of the hypotheses for the intended meaning. This must be contained at an early stage by filtering out non-sensical combinations in the dialogue context during the understanding process. If one adds the uncertainties outlined above during speech recognition, the result is a flood of hypotheses that ultimately must be reduced to a single representation of the intended meaning.

Other important sources of uncertainty are the semantic and pragmatic vagueness of many utterances and the incompleteness of knowledge sources like the user models and the domain model.

In more complex dialogue situations, it is quite common that a final decision about the intended meaning of an utterance can only be made after a further sequence of subsequent utterances has been understood, so that a kind of wait-and-see understanding strategy has to be implemented by the dialogue system. This requires the possibility of storing underspecified meaning representations in the discourse memory in order to allow inference and response generation without having achieved full understanding, e.g. in the case of a cataphor as an expression that co-refers with a later, more specific, expression in the discourse.

User: Any Volkswagen Beetle would do for me.

System: If you are more specific about the production date, I will present it to you. User: The 1980 Beetle.

(c) Testing dialogue understanding through system reactions

AI systems for dialogue understanding extract the intended meaning of digital input data in a given context, so that they can produce an output that is judged to be as an appropriate reaction by the dialogue partner.



Figure 2. Testing successful understanding: the ellipses on the left symbolize the different types of input data that an AI system is supposed to understand. The ellipses on the right list the main tasks that can be used to validate correct understanding.

One of the simplest heuristics for testing the understanding of an utterance, which should be understood as a request to an agent to perform an action, is to observe whether the appropriate action is performed or not.

There are many other tasks for testing the successful understanding of the AI system that rely on a human evaluator to judge the adequacy and validity of the systems output for a given input (figure 2). If human evaluators judge that the communicative reaction of the dialogue system to an input utterance is human-like in the given discourse context, then in a behaviouristic view based on the observable behaviour of the AI system, it is postulated that the dialogue system has 'understood' the input. In a similar way, teachers test a student's 'understanding' of a text based on the adequacy of the student's summary, paraphrase or translation of the input text.

Testing understanding by asking questions about the digital input was one of the first tasks used as evidence for meaning comprehension already in the early days of NL interfaces to databases (such as SIR [3]) and is still used for testing current systems like ChatGPT [4]. Today, captioning images and videos with automatically generated natural language descriptions is another popular understanding test.

The structure of the rest of this article is as follows. The next section analyses the development of dialogue systems over the last five decades, using German-language systems as examples. Section 3 discusses the role of large neural language models (LLMs) for the next generation of dialogue systems. Finally, §4 presents two fundamental insights for the system architectures of multi-modal dialogue systems and the most important research trends and challenges for the next generation of dialogue systems.

There are many feature-based classifications of dialogue systems, such as task-oriented, goaloriented, plan-based, mixed-initiative, multi-modal, multi-party or multi-lingual, which are often combined and overlap to varying degrees for a given implementation [2]). In §§2 and 3, we discuss examples of dialogue systems, that show these characteristics in a wide range of combinations up to the new generation of chatbots.

2. Five decades of natural language dialogue systems

Work on natural language dialogue systems has grown enormously and rapidly. Although there were always some contributions on dialogue systems in the most important conferences on AI and CL during the first 40 years, the topic has never been as central to the public discussion on AI applications as it is today. From being a niche topic in the public press, chatbots have made it into

1976	19	86 19	96 20	006 2	2016 202	26
	closed-	perceptually grounded dialogue systems	speech-to-speech dialogue translation	open-domain dialogue systems	massively multimodal dialogue	
	domain dialogue systems	multimodal dialogue systems	conversational characters	empathic virtual agents	hybrid team interaction	
		task-oriented dialogue systems	embodied dialogue systems	multiparty dialogue systems	chatbots based on large language models	
	HAM-RPM, HAM-ANS	VITRA, XTRA	VERBMOBIL, SMARTKOM, MSA	SMARTWEB, THESEUS, VIRTUAL HUMAN	MADMACS, HYSOCIATEA, OPENGPT-X	
	selected natural language dialogue systems for German					

Figure 3. The development of dialogue understanding systems in Germany from 1976 to 2026.

the best daily newspapers and are now even discussed at the World Economic Forum in Davos. Since a complete review of the development of dialogue systems (but see [2,5,6]) and the insights gained, in terms of both the breadth of architectures and the depth of methods, is impossible within the page limits of this paper we will base the discussion on our own research experience in this area of AI.

In the following, we illustrate the development of natural language dialogue systems using examples drawn mainly from the author's own project work as an AI researcher and/or a principal investigator on many dialogue systems (figure 3) over the last 48 years.

(a) Early closed-domain dialogue systems

Already 51 years ago, Terry Winograd for his PhD thesis developed the SHRDLU system at MIT [7]. It could answer simple questions, execute commands and accept information in an interactive dialogue with a human user. The dialogues were semantically grounded in a microworld domain: the famous BLOCKS world, which was also used in the early days of AI for research in vision and robotics.

SHRDLU dealt in an integrated, but very simplified way with important aspects of language: syntax, semantics and inference. The system included a parser, a recognition grammar of English, programs for semantic analysis, and a very simple general reasoning and planning system.

It could remember and explain its plans and actions as well as carrying them out with a simulated robot gripper in a CAD model of the BLOCKS world (figure 4). The user could follow the actions of SHRDLU by direct observation in an animated graphical representation of the domain of discourse. This allowed the user to verify SHRDLU's understanding of his commands and the ground truth behind the answers given by SHRDLU.

SHRDLU was a typical closed-domain dialogue system, which was based totally on symbolic methods, with complete knowledge about the state of the microworld as its domain of discourse and no machine learning components. It could ask clarification questions to the user, answer follow-up and Why-questions, resolve anaphoric references (e.g. the pronoun resolution for 'it' in figure 4) and point out incorrect presuppositions. As a demo the system was very impressive but was not robust, scalable or general enough for realistic applications.

The next generation of closed-domain dialogue systems (e.g. our HAM-RPM system [8]) achieved a certain level of domain independence by a clear separation of specific knowledge about a particular domain of discourse and the general knowledge and algorithms for its discourse processing components (HAM-RPM was tested in two completely different domains). In contrast with SHRDLU, systems like HAM-RPM and TEAM [9] were designed for easy transportability, so that they can be easily retargeted to other domains.

6



Figure 4. SHRDLU dialogue about the BLOCKS microworld. (Online version in colour.)

But a general-purpose natural language dialogue system should be adaptable to applications that differ not only with respect to the domain of discourse, but also to dialogue type, user type and intended system behaviour. In [10], we call such systems, which are transportable and adaptable to diverse conversational settings, 'transmutable systems'. A first attempt to build a transmutable system was our design of the experimental dialogue system HAM-ANS [11] whose dialogue behaviour can be switched from a cooperative mode (e.g. the system answers questions about a traffic scene) to a persuasive mode (e.g. the system tries to persuade the user to book a room in a particular hotel).

Until the late 1980s, natural language dialogue systems were mainly based on typed input and output. They were based on purely symbolic AI methods. These early systems used no machine learning and no subsymbolic neural methods. Only since the first spoken language dialogue systems evolved were statistical and neural methods used, mainly for speech analysis which was viewed as a pattern recognition problem.

(b) The emergence of multi-modal and multi-lingual dialogue systems

Since the early 1990s, the research on dialogue understanding shifted from unimodal systems, which only allow text or speech for input and output, to multi-modal systems, which allow the freedom to combine all senses for interaction. Today, multi-modal dialogue systems combine speech with gestures, eye- and head-gaze, facial expressions and many other aspects of body language [12]. In many applications of dialogue systems like driver or worker assistance systems, it turned out that multi-modal discourse in comparison to unimodal discourse is faster and shorter, more robust, less ambiguous, more expressive and more coherent [13].

Two pioneering European projects for multi-modal dialogue systems were ALFRESCO and XTRA. ALFRESCO [14] focused on the interactive exploration of multi-media information about Italian frescoes. It promoted multi-modality by combining language, gestures and hypermedia, coordinated through a coherent management of all types of communication acts. Our XTRA system combined natural language and pointing in an intelligent user interface to an expert system for tax advice. The relevant tax form is displayed on the screen so that the user can refer to regions of the form by various tactile gestures, including different granularities, inexact pointing and pars-pro-toto deixis [15,16]. After more than 30 years of successful research on multi-modal dialogue systems, it is surprising to see that some of the latest systems such as ChatGPT, which are most present in the public press, are still based on a very restricted communication channel, based on a teletype paradigm similar to SMS interaction.

When seeing a series of TV pictures showing a part of a freeway where several hundred vehicles are lined up one behind the other, each one moving forward only at a snail's pace, we

can sum up the scene with the expression 'traffic-jam'. This is a typical example of a large class of situations that can be described with the statement 'one word says more than a thousand pictures'—a reversal of the classical saying [17]. VITRA (Visual Translator) is one of the first systems, which is perceptually grounded since all discourse objects are linked to visual objects and events recognized by the system's vision system. VITRA translates the output of an image sequence analysis system of football scenes into natural language descriptions and can serve as a football radio reporter. Such a live captioning of videos is becoming increasingly popular for live streaming sports [18]. The user can also engage in a multi-modal dialogue with VITRA about the ongoing football game.

SMARTKOM is one of the first multi-modal dialogue systems that was designed for mobile devices and supports users' on-the-go lifestyles [19]. SMARTKOM supports face-to-face dialogic interaction between two agents that share a common visual environment: the human user and Smartakus, an autonomous embodied conversational agent. Smartakus is a self-animated interface agent with a large repertoire of gestures, postures and facial expressions.

SMARTKOM covers the full spectrum of multi-modal dialogue phenomena like

- mutual disambiguation of modalities
- multi-modal deixis resolution and generation
- cross-modal reference resolution and generation
- multi-modal anaphora resolution and generation
- multi-modal ellipsis resolution and generation
- multi-modal turn-taking and backchannelling

Smartakus uses body language to notify users that it is waiting for their input, that it is listening to them, that is has problems in understanding their input, or that it is trying hard to find an answer to their questions.

The mobile shop assistant (MSA) is an early example of an embodied dialogue understanding system, in which humans can manipulate, and converse with physical objects in their surrounding environment via coordinated speech, handwriting and gesture. In this dialogue scenario, shopping items have human-like characteristics and such anthropomorphized objects include embodied dialogue systems [20]. MSA provides a tangible user interface mediating spatially manipulable physical objects with digital twins of these objects. For example, grasping a particular shopping item from the shelf may start a self-introduction of the anthropomorphized object (figure 5) as a first turn in a mixed-initiative dialogue with a customer.

VERBMOBIL [21,22] was one of the first speaker-adaptive and bidirectional speech-to-speech translation systems for multi-lingual cell phone dialogues between four languages (English, German, Japanese and Chinese). It has been trained on a variety of business-oriented dialogues, including appointment scheduling and travel planning. The system integrates five processing threads, from shallow processing based on end-to-end machine learning using transcribed natural dialogues to deep semantic transfer based on an interlingua. Transfer rules are applied to underspecified semantic representations for a precise translation, which is less robust than the statistical and neural translation methods, which VERBMOBIL runs in parallel.

In VERBMOBIL's multi-blackboard and multi-engine architecture, the results of concurrent processing threads are combined incrementally. All results from concurrent processing modules come with a confidence value so that statistically trained selection modules can select the most promising result. Uncertainties are captured at each processing stage, so that they can be reduced by linguistic, discourse and domain constraints as soon as they become applicable.

In the evaluation metrics for the dialogue translation system, the task-completion rate plays a major role, so that the achievement of the main dialogue goal of the two human participants, e.g. to finding an acceptable time slot for their planned meeting, is the most important criterion. After training VERBMOBIL with 181.6 h of transcribed human dialogue, the system achieved a task-completion rate of 90% in end-to-end evaluations with real users.



Figure 5. Example of an embodied dialogue system. (Online version in colour.)

At the end of their conversation, the participants can request a written summary of the main dialogue results in their own language. The summaries are based on the semantic representation of all dialogue turns stored in VERBMOBIL's dialogue memory as a by-product of the deep processing thread. The most specific and mutually accepted decisions from the discussion are selected to generate the multi-lingual summaries.

One of the main lessons learned from the VERBMOBIL project is that the problem of speechto-speech translation of spontaneous dialogue language can be tackled with the combination of deep and shallow processing approaches.

(c) Towards open-domain and multi-party dialogue systems

When the first successful information extraction systems emerged, a first generation of opendomain dialogue systems was implemented that were capable of answering user queries based on information retrieved from the open web. SMARTWEB [23] was one of these systems which combined domain knowledge represented in knowledge graphs with the results of a series of web search commands and statistical methods for information extraction from the retrieved multimodal documents. If the answer to a user query cannot be found or inferred from the system's domain knowledge, then open-domain information extraction methods based on deep machine learning are used. Of course, answers based on knowledge graphs come with a higher confidence and trust level than answers based on the open web. Distinguishing features of SMARTWEB are its introspective capabilities used for predicting how long an answer will take and which confidence level can be expected based on experience with similar questions and the open sources of information used. SMARTWEB [23] provides not only an open-domain question-answering machine but a multi-modal web service interface for coherent dialogue, where questions and commands are interpreted according to the context of the previous conversation. For example, if the driver of our Mercedes-Benz test car asks SMARTWEB 'Where is the closest Italian restaurant?' it will access a web service to find an appropriate restaurant and show its location on a digital map presented on the large dashboard display. The user may continue his dialogue with a command like 'Please guide me there with a refuelling stop at the lowest price gas station'. In this case, SMARTWEB combines a navigation service with a special web service that finds low gas prices. SMARTWEB includes plan-based composition methods for semantic web services, so that complex tasks can be carried out for the mobile user.

8



Figure 6. Multi-party dialogues with virtual humans. (Online version in colour.)

For many years, multi-modal dialogue systems were primarily used for information-seeking tasks or dialogic access to cyber-physical systems in the Internet of Things. However, when the Internet of Services enabled the automatic composition of services in so-called mashups a new task for multi-modal dialogue systems emerged. In THESEUS, the first dialogue systems for composing new services in a multi-modal and cross-device style were implemented in a wide range of applications [24] such as medical diagnosis based on radiological data or car insurance services including damage appraisal.

One of the most complex spoken dialogue understanding tasks is multi-party dialogues between more than two virtual agents and a group of humans. Our VIRTUAL HUMAN system [25] is based on a complex agent-based dialogue management platform, which allows multimodal turn-taking as well as emotional involvement in controversial and argumentative group discussions. One of the demonstration scenarios for multi-party dialogue understanding of the VIRTUAL HUMAN system allows two human candidates to take part in a football quiz in a virtual sports studio (figure 6).

The three virtual characters take on two different roles: that of a moderator and that of two football experts who help the human candidates by giving hints. Just as in a TV quiz, the real and the virtual participants can discuss among themselves, ask for help and answer questions. In the studio, the characters behave like actors following independently the stage directions that guide the narration. For each character, dynamic dialogue goals are set up that fit that character's role. While pursuing these goals, the characters plan their multi-modal behaviour autonomously. Depending on the course of the interaction, the moods and emotions of the characters change. Just as with humans, the affective state of the characters is reflected in their body language including gestures, facial expressions and colouring, which is rendered in real time for each animated virtual character.

Today, cyber-physical environments are omnipresent, for instance as smart homes, cars, shopping environments, business facilities, Industry 4.0 factories and smart cities. Characterized by a large number of individual systems and devices with their sensors and actuators, the interaction paradigm from the user's perspective is shifting towards system-environment interaction. Following this principle in MADMACS [12], a single user or user groups can freely choose an interaction modality to address the environment, which responds in a multi-adaptive manner. This new interaction paradigm requires massively multi-modal dialogue platforms such as SiAM-dp [26], which offers a free choice of up to nine different modalities in any combination.



Figure 7. Softbot and cobot dialogues for worker assistance. (Online version in colour.)

In the new generation of smart factories, dialogue understanding plays a major role in the communication between human workers and robots or softbots. Industry 4.0 [27], the fourth industrial revolution, requires hybrid teams of workers and collaborative robots with different skill sets enabling mass customization of products in highly adaptable production environments. A new form of team robotics is needed that focuses on human–machine interaction led by skilled human workers. To solve complex manufacturing tasks, they work hand in hand with robots as a team, that plans and acts together [28]. In our HYSOCIATEA system [29], softbots (i.e. software robots that act as intelligent communicative agents) give cognitive assistance for new assembly tasks in addition to physical worker assistance provided by cobots. As shown in figure 7, these worker assistants are often visualized as animated characters that serve as multimodal dialogue partners. Since hybrid teams must agree on a particular task allocation to the robots, softbots and humans, they use spoken multi-modal dialogue to agree and commit to certain tasks. In HYSOCIATEA, a blackboard is used in a multi-agent architecture for dialogue management.

3. Large neural language models: super-parrots or human-like understanding?

Recently, LLMs have become important tools for realizing scalable AI systems for natural language understanding in open-domains. LLM-powered language generators are now able to produce snippets of text that are often difficult to distinguish from human-written text. This is undoubtedly a significant advance in natural language processing and generation.

Dialogue Systems such as LaMDA (short for 'Language Model for Dialogue Applications', Google [30]) or ChatGPT (GPT short for 'Generative Pre-trained Transformer', OpenAI [4]) attempt to answer the user's questions about everything under the sun, write essays on any topic, tell stories, explain jokes, solve math word problems and perform other increasingly varied and complex tasks in conversation with a human user.

The dialogue format makes it possible for systems like ChatGPT to answer follow-up questions, admit their mistakes, challenge incorrect premises and reject inappropriate requests.

Unlike most previous natural language dialogue systems, discussed above, they respond very quickly. This enables a fluent and coherent dialogue with the human user.

However, LLM-powered dialogue systems such as LaMDA, PaLM (Google [31]) or ChatGPT have not yet achieved human-like dialogue understanding and behave more like 'super-parrots' because they do not have an explicit representation of the communicative intent of their utterances [32,33], have only a very limited grounding of meaning in the physical world, and are unable to deal with complex meta-dialogues due to their very limited introspective and metacognitive capabilities. Talking birds such as parrots are animals that can imitate human speech, some up to 2000 words. But LLMs such as LaMDA are trained with 1.56 trillion words so that they cover the entire vocabulary of native speakers of a language.

(a) ChatGPT's output is not easily distinguishable from dialogue contributions of humans

ChatGPT (Generative Pre-trained Transformer) is built on top of the GPT-3 family of large language models. It is fine-tuned with both supervised learning based on annotated dialogues between human AI trainers and reinforcement learning techniques based on user feedback. The basic training was performed on 570 gigabytes of data obtained from books, web texts, Wikipedia articles and other documents on the Internet. In sum, 300 billion words were fed into the system. The GPT-3 LLM of OpenAI has 175 billion parameters, derived from 45 terabytes of training data parameters with a memory size of the model exceeding 350 gigabytes, and close to \$12 million in compute costs. An efficient initial training of LLMs is only possible on massively parallel supercomputers based on Graphical Processing Units (GPUs) provided by NVIDIA or Tensor Processing Units (TPUs) provided by Google. In machine learning terminology, a parameter is a value that can be adjusted in a neural model to optimize its performance. In the case of ChatGPT, the model's 175 billion parameters were learned during the training process. These parameters include the weights of the neural network's connections, which determine how much influence each input has on the output. The parameters also include the biases of the neurons in the multi-layered neural network in an advanced transformer architecture.

Although the use of LLMs for dialogue systems is at a very early stage so that there is not enough empirical evidence, it seems that, based on current technology and electricity prices, each chat costs 'single-digit cents'. It is therefore reasonable to speculate that serving just 100 million chats per month could cost millions of dollars. However, with dedicated hardware accelerators beyond GPUs and TPUs, as well as more efficient algorithms for the deployment phase that puts an LLM into production after the training phase, the cost can certainly be reduced.

LLM-driven dialogue systems often repeat back segments of the texts they have been trained on, so that they are often criticized as 'super-parrots'. But because they probabilistically select relevant words or text snippets from the language model, they don't engage in simple plagiarism.

Google's recent PaLM system, with 540 billion model parameters, was built by fine-tuning Transformer-based neural language models trained on 6144 TPU chips as ASICs for accelerating highly parallel machine learning tasks. However, unlike GPT-3, PaLM is not an open-access system. The question remains whether the number of model parameters is a good metric for the quality of LLMs or whether the training computational budget or the average perplexity are better proxies.

However, it is interesting to note that when trying to solve math word problems such as 'Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4pm. He then travels back at a rate of 6 mph. How long does it take him to get back?' with a 62 billion parameter model PaLM generates the wrong answer '18h', whereas the 540 billion parameter model generates the correct answer '5h' [31].

Convincing evidence that the output produced by systems such as ChatGPT is formally of high quality and not always easily distinguishable from texts written by human authors is, ironically, the ban on their use for submissions to a leading machine learning conference. The Program Chairs of the International Conference on Machine Learning, ICML 2023, have banned papers that include text generated by LLMs in the Call for Papers (see https://icml.cc/Conferences/2023/CallForPapers): 'Papers that include text generated from a large-scale language model (LLM) such as ChatGPT are prohibited unless this produced text is presented as a part of the paper's experimental analysis.'

(b) Large language models with more computational units than the human brain?

Neurons are the computational units of the brain. An adult male human brain has an average of 86 billion neurons. But it would be wrong to conclude that then PaLM, with its 540 billion model parameters, has more computational power than the human brain, because each neuron has, on average, about 7000 synaptic connections with other neurons and a total of about 100 trillion synapses. Synaptic connections are an essential part of the human brain's information-processing engine. LLMs with 100 trillion synapses are still a long way off. Of course, only parts of the brain are involved in understanding and producing speech or language. But with today's brute-force approach to deep learning on massively parallel machines, training a model with 100 trillion parameters would be extremely resource-intensive in terms of hardware and power.

By contrast, the human brain is powered only by 20 Watts and energy consumption is only 6 kcal/day per billion neurons. It consumes 20% of the total body energy budget [34].

Language processing in the brain is organized in several cascaded modules, some of which are likely to work in parallel [35]. Acoustic-phonological processes take place during the first 100 ms after acoustic stimulation. Semantic and syntactic relations in a sentence are processed between 300 and 500 ms after the stimulus onset, possibly in parallel systems. Syntactic and semantic integration processes take place approximately 600 ms after the stimulus input and beyond. Thus, the simple end-to-end transformer architectures of LLM-based dialogue understanding are not in line with the physiological and cognitive findings in neuroscience.

(c) Some limitations of ChatGPT

The developers of ChatGPT have placed a great emphasis on educating users about the limitations and potential dangers of gullibly accepting texts produced by the system. They point out in their user guidelines that it is important to keep in mind that the output may sometimes be inaccurate, untrue, and otherwise misleading at times as this is a direct result of the system's design (i.e. maximizing the similarity between the output and the dataset on which the models were trained). ChatGPT will occasionally make up facts or 'hallucinate' outputs. ChatGPT does not have access to the Internet, search engines, databases or any other sources of information outside of its own model. So it cannot verify facts or provide references to relevant new sources of information. If you ask for today's train schedule, it will reply 'I don't have the up-to-date information about the schedule of trains, as my knowledge cut-off date is 2021'.

While ChatGPT has a very simple discourse memory, there is a limit to how much information it can retain (approximately 3000 words from the current conversation). Ideally, the ChatGPT would ask clarifying questions when the user entered an ambiguous query. In our tests with PP-attachment ambiguities, ChatGPT (January 2023 demo release) did not ask clarification questions, when multiple readings of a user query are possible. In the following examples of question-answer turns, ChatGPT seems to work with the two readings caused by the ambiguous attachment of 'with a gun', whether the man killed by the police was armed with a gun (see 2) or the police officers killed the man with a gun (see 4). However, the user is not asked a clarification question.

- (1) User: Short response please. Did Cops kill a man with a gun?
- (2) ChatGPT: Yes, it is possible that police officers have killed someone who was armed with a gun in certain circumstances. The justification of the use of force would depend on the specific circumstances of the case, and it is subject to investigations and legal process.
- (3) User: Can Cops kill a man with a gun? Short response please.
- (4) ChatGPT: Yes, under certain circumstances, police officers may use deadly force, including shooting someone with a gun, if they perceive an imminent threat of death or serious bodily injury to themselves or others.

12

- (5) User: Why have you picked the meaning that the police officers used a gun in my last query? Short answer please.
- (6) I interpreted your last query to be asking about the possibility of police officers using a gun to kill someone, as that is what you have been asking in previous questions.

It's also interesting to note that ChatGPT does not use the colloquial term 'cops' in its answer, but the politically correct term 'police officer'.

Metacommunicative questions about the system's interpretation of user input (see 5) are answered, but they do not provide a convincing explanation of why a particular reading of the attached prepositional phrase was picked (6). However, the phrase 'in my last query' in (5) is interpreted correctly, based on ChatGPT's simple discourse memory.

4. Concluding remarks: architectural insights and grand challenges for future research

The research discussed in the previous sections on a wide variety of dialogue system prototypes has provided some fundamental insights into adequate system architectures which, after a long test-of-time, are still valid today for the design of future advanced multi-modal dialogue systems. Two of these principles will be discussed in the first two paragraphs (a) and (b) of this section, before we give an outlook on grand challenges for future research in dialogue systems.

(a) The principles of symmetric multi-modality and no presentation without representation for dialogue understanding

A central notion for advanced dialogue systems is symmetric multi-modality, where all input modes (e.g. speech, gestures and facial expressions) are also available for output, and vice versa. A dialogue system with symmetric multi-modality must not only understand and represent the user's multi-modal input, but also its own multi-modal output [36]. Only mixed-initiative dialogue systems with symmetric multi-modality create a natural experience for the user in the form of everyday human-to-human communication, by allowing both the user and the system to combine the same range of modalities.

When the generation component generates some graphics or displays images and videos on the screen as part of its response to a user input, the dialogue system needs to have a symbolic representation of the current screen content and the intended dialogue act, to be able to refer to it in the next dialogue turns. Without such a representation of the generated multi-modal presentation, the user's anaphoric, cross-modal and gestural references could not be resolved.

Furthermore, multi-modal turn-taking, back-channelling and metacommunicative interaction as important components of human-like dialogue behaviour can only be realized if the principle of representing the communicative intent of all utterances of the dialogue partners in the discourse memory is followed.

Multi-modal fusion maps the combined output of the mode-specific analysis components onto an internal representation of the intended dialogue act [37]. The modality fission component provides the inverse functionality of the fusion component (figure 8). It is responsible for the situation-aware allocation of mode-specific generators to segments of the intended communicative act and for the synchronization of their multi-modal output. Multi-modal fusion enables a reduction of uncertainties and a mutual disambiguation of the components of a multi-modal utterance in dialogue understanding by excluding non-sensical combinations of possible interpretations. Symmetric multi-modality is a fundamental design goal for human-like dialogue behaviour, which is difficult to achieve today, because existing mode-specific generators, e.g. for facial expressions, do not yet reach human levels.



Figure 8. Multi-modal fusion and fission. The flow of information from the user's multi-modal input to an internal symbolic representation and back to a multi-modal output as the system's communicative reaction is central to symmetric multi-modality based on semantic fusion and fission components for a wide range of modalities.



Figure 9. Anticipation feedback loops. The graphics shows the dialogue partner model as a prerequisite for tailoring a planned utterance to an individual user. The loop is repeated until the anticipated communicative intention has reached a maximum degree of similarity with the intention of the system.

(b) Anticipation feedback loops for user-adaptive behaviour of dialogue systems

A declarative representation of the communicative intention of a planned utterance is also an important prerequisite for the simulation of user-adaptive utterance generation based on anticipation feedback loops as control structures in human-like dialogue systems. The basic idea behind anticipation feedback is to use the system's analysis components during the system's utterance generation process (figure 9). The aim is to avoid misunderstandings and communication failures when the intended dialogue act is not successful in a given discourse situation. Before the system realizes a planned utterance, it checks whether the anticipated interpretation s2 of the dialogue partner approximately corresponds to its communicative intention s1 (see 5 in figure 9). If not, the generation component is called again to replan the utterance.

To simulate the dialogue partner's interpretation of a planned utterance the system uses its user model, which it has built up during the current conversation or previous interactions with this particular dialogue partner. This means that anticipation feedback loops are based on the implicit assumption that the system's language analysis and interpretation procedures (but 14

not necessarily the content of the sources of knowledge used by these procedures) are similar to those of the user. Since most people cannot be assumed to have an accurate model of the comprehension capacity of their dialogue partners, it is plausible that they use the similarity assumption as a default rule in everyday conversation [10,16]. Although the current state of the art in dialogue systems by no means provides sufficient evidence for assuming a far-reaching similarity, anticipation feedback loops can considerably improve the quality of present dialogue systems.

During the last five decades such anticipation feedback loops were used to cope with a variety of phenomena in many dialogue systems, such as the generation of ellipsis and anaphora in HAM-ANS [38], the generation of deictic expressions combined with pointing gestures in XTRA [15], the generation of spatial descriptions in VITRA [17] and the generation of paraphrases [39]. Currently, similar ideas are used in systems that apply Deep Reinforcement Learning for paraphrase generation [40].

(c) Extensions to more human-like dialogue situations

Although enormous progress has been made in the last five decades in the breadth of coverage of natural dialogue phenomena, there is no system today that can handle all complex dialogue situations in a human-like manner.

We currently see seven major research trends for the next generation of computational dialogue systems, all of which realize a shift from a simpler task to a more complex conversational setting:

- from closed-domain to open-domain dialogue systems
- from single-initiative to mixed-initiative dialogue systems
- from unimodal to multi-modal dialogue systems
- from single-task to multi-task dialogue systems
- from monolingual to multi-lingual dialogue systems
- from dyadic dialogues to multi-party conversations
- from emotionless dialogues to emotionally charged conversations

Two other trends of a more methodological nature have been increasingly adopted in dialogue research over the last five years:

- The move from black-box dialogue processing to transparent dialogue systems, that can
 explain their own processing architecture, knowledge sources and their limitations in
 order to become a trustworthy conversational assistant
- Moving from purely symbolic or neural methods to hybrid neuro-symbolic methods that combine the best of these two approaches for advanced dialogue systems

The scalability of neural language models based on deep learning is a clear advantage for opendomain applications [6]. They clearly outperform previous approaches on most international benchmarks. However, there are not enough publicly available large annotated corpora for building data-driven multi-modal dialogue systems [41]. The high computing and memory requirements and the associated huge energy consumption for the training phase of deep learning make the current approaches implausible as models of human-like dialogue understanding, at least from the perspective of cognitive science and brain science.

The ultimate AI system that can serve as a ubiquitous conversational partner must be able to handle all these different extensions.

(d) Chatbots trying to satisfy Grice's conversational maxims

Paul Grice in his essay 'Logic and Conversation' introduced his well-known Cooperative Principle for dialogue behaviour, according to which a speaker should make his contribution such

as is required by the accepted purpose or direction of the dialogue in which he is engaged, at the stage at which it occurs [42]. He introduced four conversational maxims, which will generally produce results in accordance with the Cooperative Principle:

- Maxim of quantity: be informative
- Maxim of quality: be truthful
- Maxim of relation: be relevant
- Maxim of manner: be clear

Grice's maxims are much too vague to be implemented directly in a dialogue system, but there are early implementations in some of the systems discussed in §2, which are based on algorithms inspired by some of these maxims for specific generation tasks. For example, the answering of why-questions in the early HAM-RPM dialogue system [43] or the generation of referring expressions in the IDAS system [44] adhered already to the Gricean maxims.

Today, even tutorials for the professional design of commercial chatbots recommend taking Grice's Conversational Maxims into account as useful conversation design tips to create more natural conversation flows and improve the customer experience [45].

When chatbot dialogues are generated based on machine learning of large language models, as in Google's LaMDA system [30], metrics are used to guide the training phase. We found that LaMDA's quality metrics are very similar in their spirit to the Gricean maxims. For example, Sensibleness, Specificity and Interestingness (SSI), which are scored by human raters, are dimensions used for the quality metric. Sensibleness refers to whether the model produces responses that make sense in the dialogue context (Maxim: Be relevant). Specificity is measured by judging whether the system's response is specific to the preceding dialogue context, rather than a generic response that could apply to most contexts (Maxim: Be clear). Interestingness measures whether the model produces responses that are also insightful, unexpected or witty (Maxim: Be informative).

The LaMDA classifiers are trained to predict the SSI ratings for the response candidates proposed by the LaMDA generator. The LaMDA generator is trained to predict the next token on a dialogue dataset. For a given multi-turn dialogue context, the LaMDA generator first generates several candidate responses and then the LaMDA classifiers predict the SSI scores for each candidate response. The candidate utterances are ranked according to their SSI scores, and the top result is selected as the response. This means that the prioritized output utterance at least approximately satisfies the Gricean maxims.

(e) Supporting ubiquitous multi-linguality for dialogue understanding systems

For many of the languages spoken in Europe (24 official and 10 co-official EU languages), the digital tools and services needed for natural language dialogue systems (e.g. software for speech and text processing, translation and natural language generation) are not supported or only to a very limited extent (e.g. for Irish or Maltese as official EU languages and Basque or Catalan as regional co-official EU languages). A major obstacle to the use of large language models in dialogue understanding systems is the lack of open language resources for machine learning, fine-tuning and evaluation for many European languages. For example, open multi-modal corpora for languages such as Slovenian, Greek and Hungarian do not yet exist in a sufficient quantity and size (see [46] p. 164]).

The main digital language technologies and corpora are only available for English at a satisfactory level. For German, French and Spanish, there is a moderate support, which compared to English is not yet sufficient. For the rest of the official or co-official languages, the situation is unsatisfactory and not in line with the European Language Equality Act [46]. Fortunately, in the EU, the Parliament supports multi-lingualism as a positive asset for the continent and supports the ambitious long-term goal of establishing digital language equality in Europe by 2030.

In many countries outside of Europe, the situation is often worse, as there is only marginal support for digital language technologies and many of these small language communities in Africa, South America or Asia are at risk of digital language extinction. One of the grand challenges is to develop multi-lingual digital language tools and dialogue technologies so that in the medium term multi-modal dialogue understanding systems become available for all of the more than 7000 different natural languages in the world.

Promising research directions to address this problem are new resource-efficient training methods, in particular transfer learning approaches. A new cross-lingual and progressive learning (CLP-Transfer) approach, which uses a large and pre-trained model in a source language (e.g. English), helps to generate a model of the same size in a target language (e.g. German), which is not trained from scratch. Instead, the overlapping vocabulary of the source and target languages is used to initialize the transfer learning process for the large target model [47]. Only a small pre-trained model of the target language is used, with much fewer parameters than the large model of the source language. Since the vocabulary overlap between German and English is significant, the CLP-Transfer achieves an 80% reduction in training effort compared to training from scratch for the same model size. Using the publicly available BLOOM LLM for English, the openGPT-X project derived a 6.4 billion parameter model for German using the CLP-Transfer method, achieving training efficiency by exploiting publicly available LLMs for other languages.

Although transfer learning of a new language is a promising approach, human children cannot acquire a language simply by listening to it. According to [33] the process of acquiring a linguistic system, like human communication in general, relies on joint attention and intersubjectivity: the ability to be aware of what other humans are attending to and guess what they are intending to communicate. Human language learning is thus an interactive process, and conversational machine learning is a promising research trend for dealing with languages that lack digital resources. It is important to understand the limitations of current large language models, put their success in dialogue systems in context and reduce the hype [32].

Two of the important aspects of dialogue understanding (e.g. for textual entailment) on which LLMs perform poorly today are logical reasoning and its explanation for deductive and inductive inference tasks. However, recent research [48] has shown that using a neuro-symbolic approach, it is not impossible to create multi-step reasoning and to generate inference traces in natural language that are easy for humans to interpret. But there are many open questions, such as when to stop the reasoning process in the LLM, how to reduce fine-tuning, and how to avoid false inferences.

While SHRDLU was able to explain its closed-domain responses to the user input, most of today's open-domain systems are unable to do so. The limited trustworthiness of these systems due to the lack of an explanation component is therefore a fundamental challenge for applications that have a critical societal impact. While neural systems can be easily integrated with machine learning and provide scalable and robust dialogue behaviour, symbolic systems can be easily integrated with logical reasoning and provide explainable and trustworthy dialogue behaviour.

This is the way forward: to achieve more human-like behaviour from dialogue systems, we can use interactive machine learning with human trainers to gradually expand the communicative capabilities of these systems and implement them as hybrid neuro-symbolic systems for advanced dialogue understanding.

Data accessibility. This article has no additional data.

Author contribution. W.W.: conceptualization, funding acquisition, methodology, software, supervision, validation, visualization, writing—original draft and writing—review and editing.

Conflict of interest declaration. I declare I have no competing interests.

Funding. I received no funding for this study.

References

1. Oviatt S, Schuller B, Cohen PR, Sonntag D, Potamianos G, Krüger A (eds). 2019 *The handbook of multimodal-multisensor interfaces*, vol. 1–3. San Rafael, CA: ACM and Morgan & Claypool.

- 2. Jurafsky D, Martin JH. 2023 Chatbots and dialogue systems. In *Chapter 15 and 16 in speech and language processing*. Draft. Stanford, CA: Stanford University.
- 3. Raphael B. 1964 SIR: A computer program for semantic information retrieval. Ph.D. Thesis, TR 220, MIT, Cambridge, MA, USA.
- OpenAI. 2022 ChatGPT: optimizing language models for dialogue. See https://openai.com/ blog/chatgpt/.
- 5. McTear MF. 2004 Spoken dialogue technology. London, UK: Springer.
- 6. Ni J, Young T, Pandelea V, Xue F, Cambria E 2023 Recent advances in deep learning based dialogue systems: a systematic survey. *AI Rev* 56, 3055–3155.
- 7. Winograd T. 1972 Understanding natural language. *Cognit. Psychol.* **3**, 1–191. (doi:10.1016/0010-0285(72)90002-3)
- 8. Hahn W, Hoeppner W, Jameson A, Wahlster W. 1978 HAM-RPM: Natural dialogues with an artificial partner. In *Proc. ECAI-78, Hamburg, Germany July 1978, pp. 122–131.* Bonn/Edinburgh: GI, AISB.
- 9. Grosz BJ, Appelt DE, Martin PA, Pereira FC. 1987 TEAM: an experiment in the design of transportable natural-language interfaces. *Artif. Intell.* **32**, 173–243. (doi:10.1016/0004-3702(87)90011-7)
- Wahlster W, Kobsa A. 1989 User models in dialog systems. In User models in dialog systems (eds A Kobsa, W Wahlster), pp. 4–34. Berlin, Germany: Springer.
- Hoeppner W, Christaller T, Marburger H, Morik K, Nebel B, O'Leary M, Wahlster W. 1983 Beyond domain-independence: experience with the development of a German language access system to highly diverse background systems. In *Proc. of the IJCAI-83, Karlsruhe, Germany, 1 August 1987*, pp. 588–594. San Francisco, CA: Morgan Kaufmann.
- Körber Y, Hahn V, Moniri MM, Schwartz T, Feld M. 2017 MADMACS multiadaptive dialogue management in cyber-physical environments. In *Proc. IEEE Conf on Intelligent Environments*, pp. 184–187. Los Alamos, NM: IEEE.
- Oviatt S. 1999 Ten myths of multimodal man-machine interfaces. *Communications of the ACM* 42, 74–81. (doi:10.1145/319382.319398)
- 14. Stock O. 1991 Natural language and exploration of an information space: the alfresco interactive system. *Proc. IJCAI* **1991**, 972–978.
- Kobsa A, Allgayer J, Reddig C, Reithinger N, Harbusch K, Schmauks D, Wahlster W. 1986 Combining deictic gestures and natural language for referent identification. In *Proc. of the* 11th Intern. Conf. on Computational Linguistics, Bonn, Germany, August 1986, pp. 356–361. Bonn, Germany: IKS.
- 16. Wahlster W. 1991 User and discourse models for multimodal communication. In *Intelligent user interfaces* (eds JW Sullivan, SW Tyler), pp. 45–67. New York, NY: ACM Press.
- Wahlster W. 1989 One word says more than a thousand pictures. On the automatic verbalization of the results of image sequence analysis systems. *Comput. Artif. Intell.* 8, 479–492.
- Herzog G, Sung CK, André E, Enkelmann W, Nagel HH, Rist T, Wahlster W, Zimmermann G. 1989 Incremental natural language description of dynamic imagery. In *Wissensbasierte Systeme* (eds C Freksa, W Brauer), pp. 153–161. Berlin, Germany: Springer.
- Wahlster W. 2006 Dialogue systems go multimodal: The SmartKom Experience. In Smartkom: foundations of multimodal dialogue systems. Cogn, tech. Series (ed. W Wahlster), pp. 3–27. Berlin, Germany: Springer.
- 20. Wasinger R, Wahlster W. 2006 The anthropomorphized product shelf: symmetric multimodal interaction with instrumented environments. In *True visions: the emergence of ambient intelligence* (eds E Aarts, J Encarnação), pp. 291–306. Heidelberg, Germany: Springer.
- 21. Wahlster W. 2000 (ed.). Verbmobil: foundations of speech-to-speech translation. Berlin, Germany: Springer.
- 22. Wahlster W. 2001 Robust translation of spontaneous speech: a multi-engine approach. In *Proc. IJCAI, Seattle, WA, 4–10 August 2001,* pp. 1484–1493. San Francisco, CA: Morgan Kaufmann.
- Sonntag D, Engel R, Herzog G, Pfalzgraf A, Pfleger N, Romanelli M, Reithinger N. 2007 Smartweb handheld — multimodal interaction with ontological knowledge bases and semantic Web services, LNAI 4451, pp. 272–295. Berlin, Germany: Springer.
- 24. Wahlster W, Grallert HJ, Wess S, Friedrich H, Widenka T. (eds) 2014 Towards the Internet of Services: The Theseus Research Program. In *Cognitive technologies*. Berlin, Germany: Springer.

- Reithinger N, Gebhard P, Löckelt M, Ndiaye N, Pfleger N, Klesen M. 2006 Virtual human: dialogic and affective interaction with virtual characters. In *Proc. 8th Intern. Conf. on Multimodal Interfaces, Alberta, Canada, 2–4 November 2006*, pp. 51–58. New York, NY: ACM.
- 26. Neßelrath R. 2015 SiAM-dp: an open development platform for massively multimodal dialogue systems in cyber-physical environments. Ph.D. thesis, Computer Science Department, Saarland Informatics Campus, University of Saarbrücken, Germany.
- 27. Kagermann H, Wahlster W. 2022 Ten years of Industrie 4.0. *Sci* **4**, 26–36. (doi:10.3390/sci4030026)
- 28. Grosz BJ, Hunsberger L, Kraus S. 1999 Planning and acting together. AI Magazine 20, 23–34.
- 29. Schwartz T *et al.* 2016 Hybrid teams: flexible collaboration between humans, robots and virtual agents. In *Proc. of the 14th German Conf. on Multiagent System Technologies, LNAI, 9872, Klagenfurt, Austria, 27–30 September 2016,* pp. 131–146. Berlin, Germany: Springer.
- Cheng HT, Thoppilan R. 2022 LaMDA: Towards safe, grounded, and high-quality dialog models for everything, Google Research Blog. See https://ai.googleblog.com/2022/01/ lamda-towards-safe-grounded-and-high.html, posted on January 21, 2022.
- Narang S, Chowdhery A. 2022 Pathways Language Model (PaLM): scaling to 540 billion parameters for breakthrough performance. See https://ai.googleblog.com/2022/04/ pathways-language-model-palm-scaling-to.html.
- 32. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021 On the dangers of stochastic parrots: can language models be too big? In *Proc. of Conf. on Fairness, Accountability, and Transparency (FAccT* '21), Online, 3–10 March 2021, pp. 610–623. New York, NY: ACM.
- Bender EM, Koller A. 2020 Climbing towards NLU: on meaning, form, and understanding in the age of data. In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020, pp. 5185–5198. Stroudsburg, PA: ACL.
- Herculano-Houzel S. 2012 The remarkable, yet not extraordinary, human brain as a scaledup primate brain and its associated cost. *Proc. Natl Acad. Sci. USA* 109, 10661–10668. (doi:10.1073/pnas.1201895109)
- 35. Friederici AD. 2011 The brain basis of language processing: from structure to function. *Physiol. Rev.* **91**, 1357–1392. (doi:10.1152/physrev.00006.2011)
- Wahlster W. 2003 Towards symmetric multimodality: fusion and fission of speech, gesture, and facial expression. In *Advances in artificial intelligence, LNAI 2821* (eds A Günter, R Kruse, B Neumann), pp. 1–18. Berlin, Germany: Springer.
- Bunt H, Kipp M, Maybury M, Wahlster W. 2005 Fusion and coordination for multimodal interactive information presentation. In *Multimodal intelligent information presentation. Series text, speech and language technology* (eds O Stock, M Zancanaro), vol. 27, pp. 325–340. Dordrecht: Springer.
- Jameson A, Wahlster W. 1982 User modelling in anaphora generation: ellipsis and definite description. In 1982 European Conf. on Artificial Intelligence, Orsay, France, July 1982, pp. 133– 138. Kaiserslautern, Germany: Univ. of Kaiserslautern.
- 39. Lenke N. 1994 Anticipating the reader's problems and the automatic generation of paraphrases. In *Proc. COLING 1994, Kyoto, Japan, 5–9, August 1994,* pp. 319–323. Stroudsburg, PA: ACL.
- Siddique AB, Oymak S, Hristidis V. 2020 Unsupervised paraphrasing via deep reinforcement learning. In Proc. of the 26th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, Online, 6–10 July 2020, pp. 10–20. New York, NY: ACM.
- 41. Serban IV, Lowe R, Hendersonn P, Charlin L, Pineau J. 2018 A survey of available corpora for building data-driven dialogue systems. *Dialogue Discourse* **9**, 1–49. (doi:10.5087/dad.2018.101)
- 42. Grice P. 1975 Logic and conversation. In *Syntax and semantics 3: speech acts* (eds P Cole, JJ Morgan), pp. 41–58. New York, NY: Academic Press.
- 43. Wahlster W, Jameson J, Hoeppner W. 1978 Glancing, referring and explaining in the dialogue system HAM-RPM. *Am. J. Comput. Ling.* **77**, 53–67.
- 44. Dale R, Reiter E. 1995 Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cogn. Sci.* **19**, 233–263. (doi:10.1207/s15516709cog1902_3)
- Martino C. 2019 Grice's Conversational Maxims applied to Conversation Design. See https:// medium.com/swlh/grices-conversational-maxims-applied-to-chatbot-conversational-uxdesign-e8c4ba670c41.

- 46. Rehm G, Way A. 2023 European language equality. In *the Cognitive Technologies series*. Cham, Switzerland: Springer.
- 47. Ostendorff M, Rehm G. 2023 Efficient Language Model Training Through Cross-Lingual and Progressive Transfer Learning. 11th Internat. Conf. on Learning Representations, Kigali, Rwanda, 1–5 May 2023. New York, NY: IEEE. (doi:10.48550/arXiv.2301.09626)
- 48. Creswell A, Shanahan M, Higgings I. 2023 Selection-inference: exploiting large language models for interpretable logical reasoning. 11th Intern. Conf. on Learning Representations, Kigali, Rwanda, 1–5 May 2023. Appleton, WI: ICLR.