AI

MDPI

Review

A Review of Natural-Language-Instructed Robot Execution Systems

Rui Liu [1,*], Yibei Guo [1], Runxiang Jin [1] and Xiaoli Zhang [2]

[1] Cognitive Robotics and AI Lab (CRAI), College of Aeronautics and Engineering, Kent State University, Kent, OH 44240, USA; yguo27@kent.edu (Y.G.); rjin4@kent.edu (R.J.)
[2] Intelligent Robotics and Systems Lab, Department of Mechanical Engineering, Colorado School of Mines, Golden, CO 80401, USA; xlzhang@mines.edu
* Correspondence: ruiliu.robotics@gmail.com

Abstract: It is natural and efficient to use human natural language (NL) directly to instruct robot task executions without prior user knowledge of instruction patterns. Currently, NL-instructed robot execution (NLexe) is employed in various robotic scenarios, including manufacturing, daily assistance, and health caregiving. It is imperative to summarize the current NLexe systems and discuss future development trends to provide valuable insights for upcoming NLexe research. This review categorizes NLexe systems into four types based on the robot's cognition level during task execution: NL-based execution control systems, NL-based execution training systems, NL-based interactive execution systems, and NL-based social execution systems. For each type of NLexe system, typical application scenarios with advantages, disadvantages, and open problems are introduced. Then, typical implementation methods and future research trends of NLexe systems are discussed to guide the future NLexe research.

Keywords: robot execution systems; robot control; natural language; task execution; social execution

Citation: Liu, R.; Guo, Y.; Jin, R.; Zhang, X. A Review of Natural-Language-Instructed Robot Execution Systems. *AI* **2024**, *5*, 948–989. https://doi.org/10.3390/ai5030048

Academic Editor: Gianni D'Angelo

Received: 9 May 2024
Revised: 6 June 2024
Accepted: 21 June 2024
Published: 26 June 2024

1. Introduction

1.1. Background

Human-robot cooperation facilitated by natural language (NL) has garnered increasing attention in human-involved robotics research. In this process, a human communicates with a robot using either spoken or written instructions for task collaboration [1–4]. The use of natural language enables the integration of human intelligence in high-level task planning with the robot's physical capabilities, such as force, precision, and speed [5], in low-level task executions, resulting in intuitive task performance [6,7].

In a typical NL-instructed robot execution process, a human gives spoken instructions to a robot to modify robot executions for improved performance. The sensors mounted on a robot capture human voice and translate it to written NL using speech recognition techniques. NL understanding is then executed to analyze user intention to perform human-expected tasks. A robot's action is decided by multiple factors including current robot status, dialog instructions, user intention, and robot memory. Using NL generation and speech synthesis, a robot answers a user's question and asks for help if needed. With speech recognition sensors and environmental sensors, human's task understanding will be integrated into robot decision making with improved performance and reliability.

Human-instructed robot executions using tactile indications have many applications, such as identifying human-robot contact states (tactile states between robot and human body) based on contact location [8], measuring hand control forces for robust stable control of a passive system [9], visual indications, such as intention estimation by understanding human attention based on gestures or body pose [10–12] and human behavior understanding based on motion detection for transferring skills to humanoid robots [13–15]. In comparison, robot executions using spoken NL indications have several advantages.

First, NL makes the human-instructed robot executions natural. For the aforementioned traditional methods, it is necessary for the human involved in robot executions to undergo training in order to use specific actions and poses that facilitate comprehension [16–20]. While in NLexe, even non-expert users without prior training can cooperate with a robot by using natural language intuitively [21,22]. Second, NL describes execution requests accurately. Traditional methods that rely on actions and poses offer only limited patterns to approximate execution requests, primarily due to the information loss inherent in the simplification of actions and poses (e.g., the use of markers to simplify actions) [23–25]. While in NLexe, execution requests related to action, speed, tool and location are already defined in NL expressions [6,26,27]. With these expressions, execution requests for various task executions are described accurately. Third, NL transfers execution requests efficiently. The information-transferring method using actions/poses requires the design of various patterns for different execution requests [23–25]. While existing languages, such as English, Chinese and German, already have standard linguistic structures, which contain various expressions to serve as patterns [28,29]. NL-based methods do not need to design specific patterns for various execution requests, making human-instructed robot executions efficient. Lastly, since the instructions are delivered verbally, instead of physically involving a human, human hands are set free to perform more important executions, such as "grasp knife and cut lemon" [30,31]. NLexe has been widely researched in the realm of automation due to its numerous advantages. Its applications span across various areas, including task planning and handling unexpected events for daily assistances [31,32], voice communication and control of knowledge-sharing robots in medical caregiving area [33–36], NL assisted manufacturing automation with failure learning to minimize human workload [6,37], parametrized indoor/outdoor navigation planning with human commands [2,38,39], and human-like interaction for social accompany for children with emotion recognition [40–42]. These scenarios are shown in Figure 1.

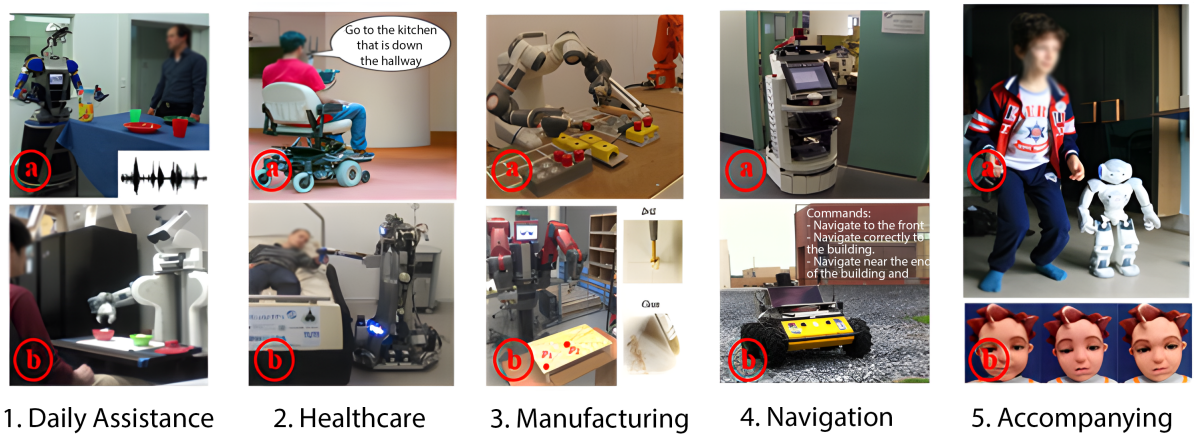


Figure 1. Typical areas utilizing NLexe systems include various applications. NL-based robotic daily assistance is exemplified by studies such as (**1a**) [8] and (**1b**) [1]. NL-based healthcare applications are demonstrated in works like (**2a**) [33] and (**2b**) [34]. NL-based intelligent manufacturing is represented by research such as (**3a**) [37] and (**3b**) [6]. NL-based indoor/outdoor navigation is explored in studies like (**4a**) [38] and (**4b**) [39]. NL-based companion systems are investigated in research such as (**5a**) [43] and (**5b**) [41].

1.2. Two Pushing Forces: Natural Language Processing (NLP) and Robot Executions

From a realization perspective, the recent advancements in NLexe have been facilitated by the progress made in natural language processing (NLP) and robotic execution technologies.

Different from NLP for text semantic analysis, NLP for robot executions needs to consider execution-specific situation conditions, such as human intentions, task progresses, environmental context, and robot working statuses. Understanding human NL in a robot

execution scenario is challenging for the following reasons. (1) Human NL requests are ambiguous that human desired objects, such as 'cup', are unclearly referred to in the expression 'a cup of water', without specific descriptions of object location and physical attributes [44]. (2) Human NL requests are abstract that high-level plans, such as 'prepare coffee, then deliver the coffee to me', are usually instructed without robot execution details, such as 'hand pose, movement speed' [45]. (3) Human NL requests are information-incomplete. Key information, such as 'coffee type, cup placement location', in the above example is missing [46]. (4) Human NL requests are real-world inconsistent that human instructed object 'cup' may be unavailable in a practical situation [47]. To solve these problems, natural language processing (NLP) techniques, which automatically analyze semantic meanings of human NL, such as speech and written text, are adopted. Supported by machine learning techniques in classification (use predefined classes to classify what type an object might be), clustering (identify similarities between objects in the scene) and feature extraction (select variables into features to represent the input and reduce set of features to be processed), NLP has been developed from simply syntax-driven processing, which builds syntax representations of sentence structures, to meaning-driven processing, which builds semantic networks for sentence meanings [48]. Improvements in NLP methods enable robots to understand human NL accurately, further enhancing the naturalness of its executions.

Recent developments in robot execution have evolved significantly over time. Initially, it began with low-cognition-level action research, wherein actions were designed and selected based on human instructions. This phase primarily focused on executing predefined tasks without much autonomy. Subsequently, research progressed to a middle-cognition-level interaction stage, which involved a basic understanding of human motions, activities, tasks, and safety concerns [49]. The most recent advancements have focused on high-cognition-level human-centered engagement research. In this stage, the robot's performance considers human psychological and cognitive states, such as attention, motivation, and emotion, to enhance the effectiveness and naturalness of human-directed robot actions [50]. The increasing emphasis on human factors in robot execution has led to a closer integration between robots and human users, thereby improving the intuitiveness of robot operations.

For a robot to collaborate with human users, techniques like text generation, speech synthesis, and multi-domain sensor fusion were developed to enrich communication. Recent text generation techniques utilize sequence-to-sequence models like variational autoencoders (VAE) [51] and generative adversarial network (GAN) [52] to generate NL sentences which increases Human-robot Interaction (HRI) naturalness and facilitates users' understanding. In NLexe scenarios like human asks robots for information, these methods need to combine with the robots' specific tasks to generate text with pragmatic information from knowledge for human users to understand [53]. To achieve a fluent and intuitive human-robot interaction, robots need speech synthesis techniques that convert the generated text to sound that a human hears easily, especially for people with visual impairment [54]. The current trending method, WaveNet [55], uses a probabilistic and autoregressive deep neural network to generate raw audio waveforms. Moreover, Ref. [56] further reduced the network's complexity to make it possible to synthesize voice in a low-performance device. For a typical NLexe task, a robot accepting NL instructions from its user usually has sensors like a camera and a distance sensor to assist in executing its task, which brings the multi-sensor fusion techniques. By combining the information from multiple sensors and instructions, the robot disposes of the information and decides the best solution in the current moment [57–59].

1.3. Systematic Overview of NLexe Research

Advancements of NLP, such as modeling semantic meanings for complex NL expressions [60,61], interpreting implicit NL expressions by extracting the logic like "I need food" from commands like "I am hungry" [62,63], and enriching abstract NL expressions by

associating them with extra commonsense knowledge [64,65], support an accurate task understanding in NLexe. Advancements of robot execution capabilities, such as replicating or representing the action of human [50,66], human intention inference and reaction based on reinforcement learning [67–69], and the ability to understand human's implicit intention with the environment and external knowledge [70,71], support intuitive task execution in NLexe. A Common architecture for an NLexe system includes three key parts: instruction understanding, decision-making, and knowledge-world mapping. In instruction understanding, the user's verbal instructions were acquired by sensors and processed by speech recognition and language understanding systems to perform a comprehensive semantic analysis. In the decision-making phase, various algorithms such as deep neural networks were used to collect and construct the task-related knowledge using information from the robot knowledge base or NL knowledge and supported robot decision-making in various manners. Through the process of knowledge-world mapping, information patterns were accurately contextualized within real-world scenarios, allowing for the resolution of incomplete knowledge gaps. This facilitated the successful execution of the NLexe process. With supporting techniques from both NLP and robot execution, NLexe has been developed from a low-cognition-level symbol matching control, such as using "yes" or "no", to control robotic arms, to a high-cognition-level task understanding, such as identifying a plan from the description "go straight and turn left at the second cross". Combining the advancements of both NLP and robot executions improves the effectiveness of both communication and execution in NLexe.

As a result of NLexe research, a substantial number of NLexe projects were launched, including the following. (1) "collaborative research: jointly learning language and affordances" from Cornell University, which interprets objects by natural-language-described affordances, such as 'water: drinkable; cup: containable', to help robot with its manipulation [72]; (2) "robots that learn to communicate with human through natural dialog" from University of Texas at Austin, which enables robots to directly learn task executions from NL instructions with a user-friendly manner [73]; (3) "collaborative research: modeling and verification of language-based interaction" from MIT, which uses NL to interpret human interactions and understand human perspective in physical world for finally integrating humans and robots [74]; (4) "language grounding in robotics" in University of Washington, which maps theoretical knowledge, such as object-related descriptions, to practical sensor values, such as sensor-captured object attributes [75]; (5) "semantic systems" from Lund University, which uses NL to describe industrial assembly processes [76]. NLexe research is frequently disseminated through prestigious international journals, including International Journal of Robotics Research (IJRR), IEEE Transactions on Robotics (IEEE TRO), The Journal of Artificial Intelligence (AIJ) and IEEE Transactions on Human-machine Systems (IEEE THMS), and international conferences, like International Conference on Robotics and Automation (ICRA), International Conference on Intelligent Robots and Systems (IROS) and AAAI Conference on Artificial Intelligence (AAAI). By using the following keywords, NLP, human, robot, execution, speech, dialog, and natural language, human-robot interaction, HRI, social robotics, about 4410 papers were retrieved from Google Scholar. Then with a focus on NL-instructed robot executions, about 1390 papers were finally kept for plotting the publication trend, shown in Figure 2. The steadily increasing publication numbers demonstrate the increasing significance of NLexe research.

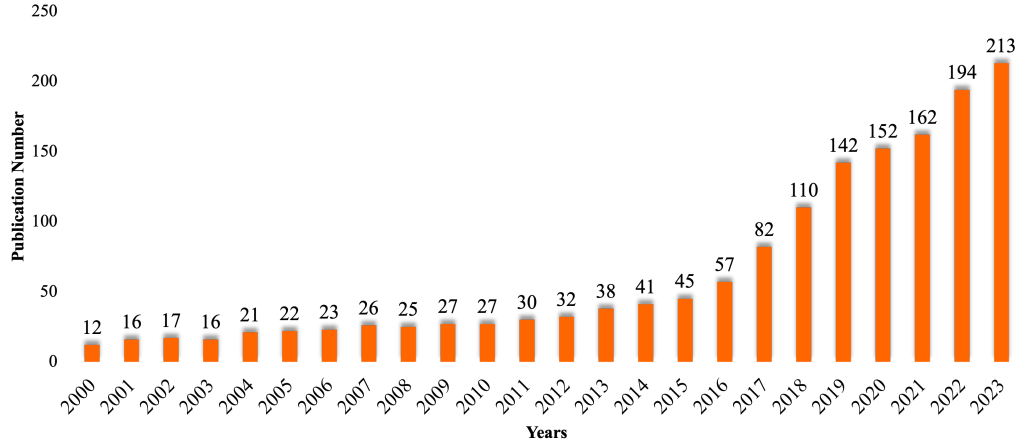


Figure 2. The annual amount of NLexe-related publications since the year 2000. In the past 23 years, the number of NLexe publications has steadily increased and reached a historical high.

2. Scope of This Review

Compared with the existing review papers about human-instructed robot executions using communication means of gesture and pose [77], action and motion [78], and tactile [79], a comprehensive review of NLexe, which employs natural language for command delivery and execution planning, is currently lacking in the literature. Considering the significant potential and growing interest in NLexe, it is imperative to provide a summary of the state-of-the-art systems in this domain. Such a review would highlight the current research advancements and guide future research directions in NLexe. NLexe is a process of interactive information sharing and execution between humans and robots [7]. In this paper, the focus is on robotic execution systems with natural language facilitation. This review systematically presents the implementations of NLexe, encompassing topics such as motivations, typical systems, realization methods, and emerging trends, as illustrated in Figure 3.

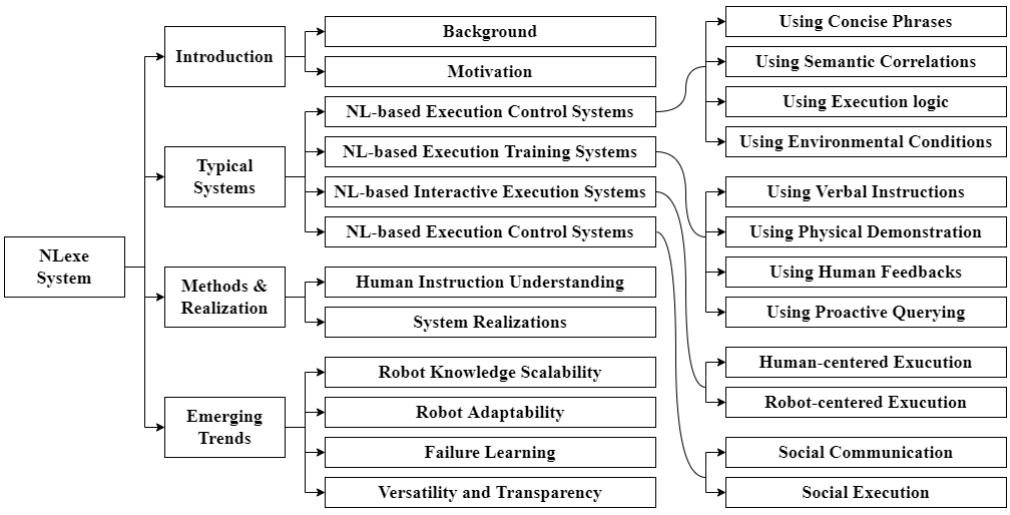


Figure 3. Organization of this review paper. This review systematically introduces NLexe implementations, encompassing topics such as motivations, typical systems, realization methods, and emerging trends. The NLexe systems are comprehensively categorized into four groups: NL-based execution control systems, NL-based execution training systems, NL-based interactive execution systems, and NL-based social execution systems. Within each category, typical application scenarios, knowledge manners, knowledge formats, as well as their respective advantages and disadvantages, are summarized.

NL is used as a communication means to realize interactive information sharing, which further facilitates interactive task executions between robots and humans. NL instructions during task executions could deliver information, including human intentions, task execution methods, environmental conditions, and social norms which are embedded in human oral speech. By extracting different contents from human NL instructions in different manners, different NLexe systems are designed for supporting human-guided robot task executions in various scenarios. To perform a simple task in the lab environment, a robot only needs a low-level cognition to execute the control commands given by its human users. In this situation, robots do not involve decision-making because it is unnecessary, which represents an NL-based execution control system. As the environment and user become different, robots need to understand the different situations and human preferences, so they can make use of a middle-level cognition to learn new execution knowledge to adapt its logic, which represents an NL-based training system. Because the execution process is not always the same in real-life production scenarios, human users may actively take charge of the execution procedure or change their minds to execute other tasks. Also, robots may produce errors and need help from human users, so the ability to describe the current situation is required. Thus, robots need a higher cognition level, so human users communicate with robots updating plans or progress, which represents an NL-based interactive execution system. When robots are used by ordinary people in daily life, communicating via NL in the execution process becomes essential. The robots thereby need to understand commonsense as humans since ordinary people do not have expertise in robotics and may get confused when the abnormal behavior of a robot occurs. Robots then need a remarkably high cognition level to understand what an ambiguous user instruction implies and figure out the optimal plan to execute or change its behavior pattern, which represents an NL-based social execution system.

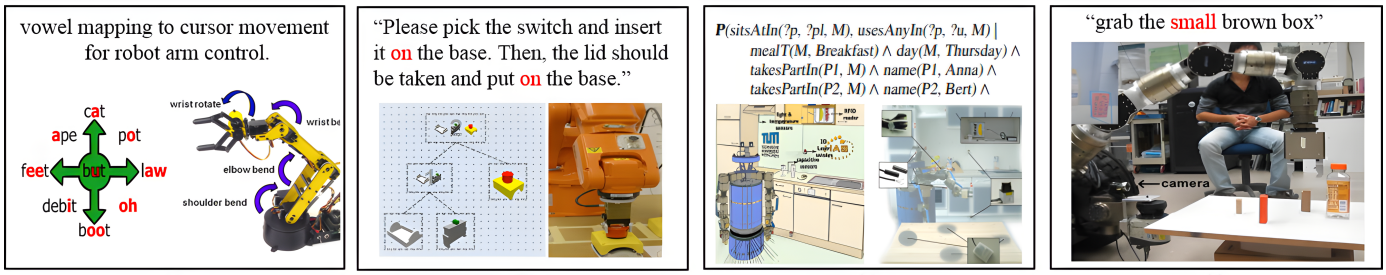
Based on the level of robotic cognition during execution, NLexe-based robotic systems can be classified into four primary categories. (1) NL-based execution control systems. In these systems, NL is used to convey human commands to robots for remote control. The robots receive only NL-format control symbols, such as "start", "stop", and "speed up", without comprehensive instruction understanding. Consequently, robots operate with a low cognition level, lacking execution planning. The human-robot role relation is "leader-follower", where humans assume all cognitive burdens in task planning, while robots undertake all physical burdens during task execution. (2) NL-based robot execution training systems. In these systems, NL transfers human execution experiences to robots for knowledge accumulation, a process termed robot training. Robots must consider environmental conditions and human preferences to adjust execution plans, including action and position adjustments, thus exhibiting a middle cognition level. The human-robot role relation remains "leader-follower", with humans handling the major cognitive burdens, such as defining sub-goals and execution procedures, while robots follow human instructions and assume minor cognitive burdens. (3) NL-based interactive execution systems. In these systems, NL is used to instruct and correct robot executions in real-world situations. Robots operate with a high cognition level, as they can infer human intentions and assess task execution progress. The human-robot role relation is "cooperator-cooperator", with both robots and humans sharing major cognitive burdens. (4) NL-based social execution systems. In these systems, NL facilitates verbal interaction between humans and robots to achieve social purposes, such as storytelling. Robots require the highest cognition level to understand social norms and task execution methods. The human-robot role relation is "cooperator-cooperator". Table 1 allows clear comparisons between four types of NLexe systems' cognition levels and execution manners, and shows human and robot involvements for the four types, respectively.

Table 1. Comparison of typical NLexe systems.

NLexe Systems	Application Scenarios	Robot Cognition Level	Human-Robot Role Relations	Human Involvement	Robot Involvements
Execution Control	action selection, manipulation pose adjustment, navigation planning	minimal	leader-follower	cognitive burden	physical burden
Execution Training	assembly process learning, object identification, instruction disambiguation, speech-motion mapping	moderate	leader-follower	heavy cognitive burden	heavy physical burden
Interactive Executions	assembly, navigation in unstructured environments	elevated	cooperator-cooperator	cognitive and physical burden	cognitive and physical burden
NL-based social execution systems	restaurant reception, interpersonal spacing measures, kinesic cues learning in conjunction with oral communication, human-mimetic strategies for object interaction	maximum	cooperator-cooperator	partial cognitive/physical burden	partial cognitive/physical burden

3. NL-Based Execution Control Systems

To alleviate the physical demands on humans and allow their hands to be free for other tasks, NL was initially employed to replace physical control mechanisms such as joysticks and remote controllers, which traditionally necessitated manual operation [31]. NL-based execution control is a single-direction NLexe in which a human mainly gives NL control commands on task executions, and a robot follows the control commands to execute the plan. In NLexe control systems, only speech recognition and direct language pattern mapping are needed without high-level semantic understanding. A robot receives simple user instructions and directly map into the actions for task execution. These simple instructions include the following. (1) Simplest instructions, with which a user gives are symbolic words identified by a robot without semantic understanding using predefined instruction-action mapping. (2) Phrases (word groups) or short sentences are complicated as the user gives a semantic correlation to describe a task. (3) Logic structure mapping, with which a user indicates an object with details like "color or shape of that object" and execution steps of a task to guide robot executions. During the controlling process, NL functioned as an information-delivering medium, conveying the robot's actions as desired by the human operator, as illustrated in Figure 4. The human user primarily made decisions during this process, while the robot was directed by recognizing and executing commands expressed through NL.



(a) (b) (c) (d)

Figure 4. Typical NL-based execution control systems. (**a**) Control using concise phrases: The motion directions of a robot arm's joints were controlled by mapping symbolic vowels from human spoken instructions [80]. (**b**) Control using semantic correlations. A robot was verbally directed by a human to perform assembly tasks by mapping semantic correlations of the components from spoken instructions [81]. (**c**) Control using execution logic. The execution of cooking tasks was logically defined. By mapping the logical structure of human spoken instructions, a robot was able to perform kitchen tasks [82]. (**d**) Control using environmental conditions. During task execution, a robot took into account practical environmental conditions, such as "object availability" and "objects" relative sizes", to perform tasks like "grabbing the small brown box" [31].

3.1. Typical NLexe Systems

According to NL command formats, systems for execution control based on NL can be primarily classified into four distinct categories.

3.1.1. NL-Based Execution Control Systems Using Concise Phrases

To realize simple NLexe, NL-based execution control systems based on short commands or keywords, uttered by users to adjust robot executions, were designed. In this system, speech recognition and word-to-action mapping are required, which brings the difficulties of accurate recognition of words and predefinition of commands. These systems are designed for concise execution with simple and straightforward words or phrases instruction. Typical tasks include asking the robot to accelerate by using the word "faster" and confirming your command by saying "yes". The NL-based execution control for planning execution procedures involved a word-to-action mapping process. This process was discrete, with the associations between words and actions predefined in the robot's database. In this framework, human operators were limited to providing the robot with symbolic and simple NL commands. The robot was required to accurately recognize the symbolic words in human speech and subsequently associate these words with the predefined actions or sequences of actions.

Typical symbol words involved in NLexe systems include action types, such as "bring, come, go, send, take" [83], motion types, such as "rotate, slower, accelerate, initialize" [84,85], spatial directions, such as "left, right, forward, backward" [86,87], confirmation and denial commands, such as "yes, no" [88], and objects, such as "cup, mug" [89].

Typical NLexe systems using symbolic word control include the following. (1) Robot motion behavior control by verbally describing motor skills, such as "find object, track object, give object, move, set velocity", for daily object delivery [90]. (2) Robotic gripper pose control by verbally describing action types, such as "open or close the grip", for teleoperation in manufacturing [91,92]. (3) Mobile robot navigation behavior control by using location destination, such as "kitchen, there", to plan robot routine [88,93]. NL-based execution control was employed to specify detailed action-related parameters, including action direction, movement amplitude, motion speed, force intensity, and hand pose status. This NL-based execution control for action specification involved a word-to-value mapping process, which was continuous, with value ranges predefined in the robot database. During the development of NL-based execution control, two mainstream mapping rules were established: fuzzy mapping, use terms, such as "move slowly, move far" [94], and strict mapping, such as "rotate 180 degrees, go to the laser printer" [95].

3.1.2. NL-Based Execution Control Systems Using Semantic Correlations

To perform relatively complex NLexe, NL-based execution control systems using semantic correlations between verbal instructions defining a series of robot actions, were designed. In such systems, speech recognition and linguistic structure analysis are required, which brings the difficulties of correctly recognizing the relationship between words. This system is designed for execution with an understanding of action-property-object-location relations. Typical tasks include giving a simple description of an object like "white cube", where robots distinguish the white cube from all cubes. To enhance the robustness of NL-based execution control in human-guided task execution, semantic correlations among control symbols were investigated through an analysis of the linguistic structures of NL commands. The control symbols encompass various types of actions, hand poses, and objects.

Typical semantic correlations used for robot execution control include action-object relations, such as "put-straw", object-location relations, such as "straw-cup" [89], object-property relations, such as "apple-red, apple-round" [96], action-action relations, such as "open–move" [97], and spatial-spatial relations, such as "hole constraint-shaft constraint" [81].

Typical NLexe systems using semantic correlation control include the following. (1) flexible object grasping and delivery by using action-object-location relations, such as "put book on desk" [98]. By introducing semantic correlations into object-related executions, human daily commonsense about object usage was initially introduced into NL-supported systems. (2) Accurate object recognition by using correlations, such as "move the black block" [96]. Semantic correlations reveal existing associations between object properties. (3) General task execution procedures, such as "open gripper—move to gripper", were defined to plan assembly procedures for industrial robotic systems [81]. (4) Flexible indoor navigation by using general motion planning patterns, such as "go-to location" [2], improving robots' adaptability towards instruction varieties.

3.1.3. NL-Based Execution Control Systems Using Execution Logic

To flexibly adjust execution plans, NL-based execution control systems using execution logic were designed, in which hierarchical logic is verbally specified to define execution preconditions and procedures. In this system, speech recognition and logic level NL understanding are required, which brings the difficulty of accurately understanding the coordination of consecutive instructions. This system is designed for tasks with an ordered execution or a conditional execution. Typical tasks include instructing with logic words like "turn left, then go forward", where robots figure the order and turn left to go straight. Despite the improvements in flexibility brought about by the integration of semantic correlation in NL execution mapping, control performance in dynamic situations remains limited due to the neglect of control logic within these semantic correlations. This oversight renders robots unable to adapt to environmental changes and hampers their ability to intuitively reason about execution plans. For instance, the NL instruction "fill the cup with water, deliver it to the user" inherently includes the logical sequence: "search for the cup, then use the water pot to fill the cup, and finally deliver the cup". Ignoring this logic during the control process can lead to incorrect executions, such as "use the water pot to fill the cup, then search for the cup", or can restrict proper executions, such as "deliver the cup, then use the water pot to fill the cup", in dynamic environments. This results in the incorrect removal or addition of execution steps. To address this issue, the study explores logic correlations—encompassing temporal logic, spatial logic, and ontological logic—among controlling symbols to enhance the adaptability of NL-based execution control methods.

Typical NLexe systems that employ NL-described logic relations include the following: (1) Using NL commands, such as "if you enter a room, ask the light to turn on and then release the gripper", to modify the robot trajectory in different situations with self-reflection on task executability [99–101]. (2) Designing robot manipulation posts based on fuzzy action type and speed requirements, exemplified by commands like "move up, then move

down" [88,94]. (3) Serving meals by considering object function cooperation logic, such as "foodType—vesselShape" relation [82,102]. (4) Assembling industrial parts by considering assembly logic, such as "first pick part, then place part" [81,103].

3.1.4. NL-Based Execution Control Systems Using Environmental Conditions

To realize human-robot executions with consideration of the surrounding environment, NL-based execution control systems using environmental conditions were designed, allowing users to give verbal notifications to the robot, describing the environmental conditions to consider in a situation-aware control context. In such systems, speech recognition, logic level NL understanding, and knowledge-world mapping are required, which brings the difficulty of correctly mapping descriptions in instruction to environmental conditions. This system is designed for tasks that real-world conditions are included in NL instructions. Typical tasks include asking a robot to identify environmental information like "box under the table", where robots locate the table to understand which box the user is referring to. In addition to the logical relationships among control commands, it is imperative to consider environmental conditions for the practical execution of NL-based execution control in real-world scenarios. Achieving an intuitive NLexe requires a balance between human commands, robot knowledge, and environmental conditions.

Typical environment conditions verbally described for control include the following. (1) human preferences on object placement location, obstacle avoiding distances, and furniture set up manners [84,104]; (2) human safety-related context, such as arms' reachable space and joint limit constraints [97]; (3) environmental conditions, such as 'door is closed, don't go to the lounge' [19]; (4) object properties, such as "mug's usage is similar to cup, mug's weight and mug's cylinder shape" [89]; (5) robot working statuses, such as "during movement (ongoing status), movement is reached (converged status)" [105].

Typical NLexe systems using practical environmental conditions include the following. (1) By using NL-described safety factors, such as "open hand before reaching, avoid collision", a robotic arm was controlled for grasping with considerations of both robot capabilities, such as "open griper, move gripper", and runtime disturbances, such as "encountering a handle, distance with an obstacle is reduced" [84,97]. (2) Outdoor navigation robot planned path by considering the location and building matchings, such as "navigate to the building behind the people" [39]. (3) Food serving robots served customers with consideration of user locations [106], and path conditions [89]. In these systems, real-world conditions were embedded in NL commands to improve robot control accuracy.

3.2. Open Problems

In the context of NL-based robot control, human-robot interaction was facilitated through verbal communication. The control commands in natural language were issued either individually or in a hybrid manner, wherein the NL commands were integrated with visual and haptic cues. A human was the only information source, guiding the whole control process. A robot was designed to simply map the human NL commands to the knowledge structure in robot databases, or to the real-world conditions perceived by the robot's sensors. With physical and mental work assignments for robots and humans, current efforts in NL-based execution control focus on improving control accuracy, decreasing human users' cognition burdens, and increasing robots' cognition burdens. However, some open problems are still existing. (1) The cognition burdens of humans in NL-based execution control were at a high level and the robot cognitions were at a low level. A human user was required to lead the execution and a robot was required to follow human instructions without understanding the task executions. The big cognition-level difference between a human and a robot restrained the intuitiveness and naturalness of NLexe with NL-based execution control systems. (2) Low robot cognition level endows robots with limited reasoning capability, restraining NLexe systems' autonomous level. (3) The robot knowledge scale was small, limiting a robot's capabilities in dealing with unfamiliar or dynamic situations where user varieties, task complexities, and real-world uncertainties

were involved in disturbing the performances of NL-based execution control systems. Detailed comparisons among NL-based execution control systems are shown in Table 2 to present four types of systems and their corresponding instruction manner, instruction format, application, pros, and cons.

Table 2. Summary of NL-based execution control systems.

	Concise Phrases	**Semantic Correlations**	**Execution Logic**	**Environmental Conditions**
Instruction Manner	predefined	predefined	predefined	sensing
Instruction Format	symbolic words	linguistic structure	control logic formulas	real-world context
Applications	object grasping, trajectory planning, navigation	object grasping, navigation	sequential task planning, hand-pose selection,assembly	safe grasping, daily assistance, precise navigation
Advantages	concise, accurate	flexible	flexible	task adaptive
Disadvantages	limited adaptability	limited adaptability	ignore real conditions	lack commonsense
References	[83,85,86,90,92]	[81,89,96–98]	[99–103]	[39,97,104–106]

4. NL-Based Execution Training Systems

To scale up robot knowledge for executing multiple tasks under various working environments, NL was used to train a robot. During the training, knowledge about task execution methods was transferred from a robot or a human expert to targeted robots. An NL-based execution training system uses human knowledge in the form of NL to train an inexperienced robot for task understanding and planning. In NLexe training systems, with the support of speech recognition technologies, robots have the capability of initial language understanding based on interpreting multiple sentences. User instructions include simple words like the name and property of an object and the action of robot motion and grasping. Human users give verbal descriptions, and robots extract information and task steps as well as their meaning according to the way instructions are interpreted. A robot also proactively asks its human users questions for object or action disambiguation, bringing the instruction complexity to a higher level since the robots need both language understanding and generation during training. Typical NL-based execution training systems are shown in Figure 5. Human commonsense knowledge was organized into executable task plans for robots with consideration of the robot's physical capabilities, such as force, strength, physical structure, and speed, human preferences, such as motion and emotion, and real-world conditions, such as object availability, object distributions, and object locations. With the executable knowledge, a robot's capabilities in task understanding, environment interpretation, and human-request reasoning were improved. Different from NL-based execution control, where robots were not involved in advanced reasoning, in NL-based execution training, robots were required to reason about human requirements during human-guided executions.

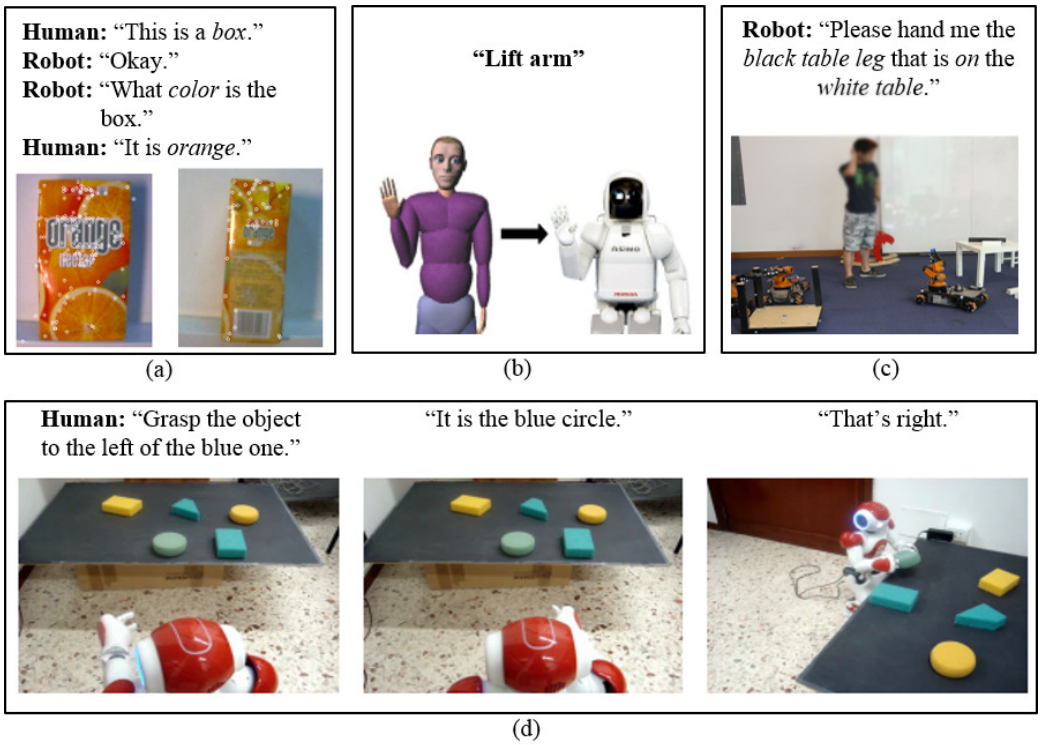


Figure 5. Typical NL-based execution training systems. (**a**) Training using human instructions. By describing the physical properties of the objects, the robot-related knowledge was transferred from a human to a robot, enabling the robot to recognize objects in the future [107]. (**b**) Training using human demonstrations. With a human's physical demonstrations of actions, a robot learns to perform an action by imitating the motion patterns, such as trajectory, action sequence, and motion speed [108]. (**c**) Training using proactive robot querying. A robot proactively detected its missing knowledge and proactively asked human users for knowledge support [109]. (**d**) Training using human feedback. During the execution process, a human proactively interfered with the execution and gave timely feedback to improve a robot's performance [110].

4.1. Typical NLexe Systems

According to the knowledge learning manners, NL-based execution training systems are mainly categorized into four types.

4.1.1. NL-Based Execution Training Systems Using Verbal Instructions

To teach a robot with high-level execution plans, execution training systems that only use verbal instructions were designed. In this way, robots learn from these instructions without any physical participation or help from their users. In this type of system, speech recognition, NL understanding, learning, and reasoning mechanisms are required, which brings the difficulty of transforming high-level human instructions to real-world implementation. They are designed for training robots on how to execute certain tasks using command sequences. Typical tasks include training robots with action combinations like "lift, then move to location", where robots learn this plan and execute it in the future. During the high-level knowledge grounding, human cognition processes on task planning and performance were modeled. The advantage of the training using human NL instructions is that a robot's reasoning mechanisms during NLexe are initially developed; the disadvantage is that the execution methods directly learned from the human NL instructions are still abstract in that the sensor-value-level specifications for the NL commands are lacking, limiting the knowledge implementations in real-world situations.

Given that human verbal instructions can deliver basic information about task execution methods, NL-based execution training initially started with defining simple commonsense by using human NL instructions, enabling a robot with initial reasoning capability

during execution [111]. Typical NLexe execution training systems using human NL instructions include (1) daily assistive robots with NL-trained object identities, such as "cup, mug, apple, laptop" [102,112]; (2) robots with NL-trained object physical attributes, such as "color, shape, weight" [107,113,114]; (3) robotic grippers with NL-trained actions, such as "grasp, move, lift" [115,116].

Instead of modeling correlations among task execution procedures, the knowledge involved in the low-level reasoning was merely piecemeal with only separate knowledge entities, such as "cup, cup color, grasp action". Piecemeal knowledge enables robots with a shallow understanding of motivations and logic in task executions. As information and automation techniques improved, the low-level reasoning method was then evolved into a high-level reasoning method, in which complex NL expressions were grounded into hierarchical knowledge structures for motivation and logical understanding. With hierarchical knowledge, NLexe systems were enabled to learn complex task executions. Typical NLexe systems include the following. (1) Industrial robot grippers with NL-trained object grasping methods, such as "raise, then move-closer" [117]. (2) Executing tasks in unfamiliar situations with NL-trained spatial and temporal correlations, such as "landmark—trajectory—object, computer—on—table, mouse—leftOf—computer" [118–121]. (3) Daily assistive robots with NL-trained object delivery methods, such as "moved forward → moved right → moved forward" [122,123].

4.1.2. NL-Based Execution Training Systems Using Physical Demonstration

To teach a robot with environment-specific execution details during NLexe, execution training systems using physical demonstration were designed to align NL instructions with real-world conditions, in which robots passively learn from their human users while human users need to participate physically. In this system, speech recognition, NL understanding, and sensor value association are required, which brings the difficulty of interpreting sensor data with abstract knowledge from a human. This system is designed for learning task execution methods and extending existing knowledge in real-world scenarios. Typical tasks include training robots with both human action and language instruction like pointing at an object in the room and telling the robot "that is a book", so robots learn the relation between the object and "book". With training in a physical demonstration manner, theoretical knowledge, such as "actions, action sequences, and object weight, object shape, and object color", was associated with sensor values. This theory-practice association enabled a straightforward, sensor-data-based interpretation of the abstract task-related knowledge, improving robot execution capability by practically implementing the learned knowledge.

A general demonstration process was that a human physically performs a task and meanwhile verbally explains the execution intuitions for a robot. The robot was expected to associate the NL-extracted knowledge with sensor data to specify the task executions. Human demonstration enables a robot with a practical understanding of real-world task executions. Compared with robot training using instructions, robot training using demonstrations specified the abstract theoretical knowledge with the real-world conditions, making the learned knowledge executable [124]. However, a robot's reasoning capability was not largely improved since demonstration-based training was actually a sensor-data-level imitation of human behaviors. And it ignored the "unobservable human behaviors", such as a human's subjective interpretation of real-world conditions, a human's philosophy in execution, and a human cognitive process in decision-making.

Typical NL-based execution training systems using human demonstration include the following. (1) Learning object-manipulation methods by associating human NL expressions with sensor data, such as "touching force values, object color, object shape, object size, and visual trajectories" [117,125,126]. (2) Learning human-like gestures by associating NL speech, such as "bathroom is there", with real-time human kinematic model and speeds [108,127]. (3) Learning object functional usages, such as "cup-like objects", by simultaneously considering human voice behaviors, such as "take this", motion behaviors, such as "coordinates of robot arms", and environmental conditions, such as "human

locations" [122]. (4) Learning abstract interpretations of environmental conditions by combining high-level human NL explanations, such as "the kitchen is down the hall", with the corresponding sensor data patterns, such as "robot speed, robot direction, robot location" [20,127]. (5) Adapting new situations by replacing NL-instructed knowledge, such as "vacuum cleaner", with real-world-available knowledge, such as "rag" [6,33,128]. (6) Robot arms learned new actions by interpreting NL instructions, such as "move the green box on the left to the top of the green block on the right", by using visual perceptions which are perceived by cameras [129,130].

4.1.3. NL-Based Execution Training Systems Using Human Feedbacks

To teach a robot to consider human preferences during NLexe, execution training systems using human feedback were designed, in which robots proactively learn from a human according to knowledge needs in specific task execution scenarios and a human decides what knowledge to learn. In this system, speech recognition, NL understanding, and real-time behavior correction are required, which brings the difficulty of proactively understanding human feedback and combining them with the current execution situation. This system is designed to change or improving the current behavior pattern using real-time human NL instructions. Typical tasks include correcting the robot's wrong behavior with an instruction like "stop, and do something else", where robots terminate the current wrong action and perform other tasks. This system has an aim of using human NL feedback to directly tell the robot "the unobservable human behaviors". With human NL feedback, robot behaviors in human-guided execution were logically modified by adding and removing some operation steps [69,109,110] or subjectively emphasizing on executions [131–133]. Compared with training using human demonstrations, training using human feedback proactively and explicitly indicates a robot with operation logic and decision-making mechanisms that its human users desire. A better robot understanding is enabled by enriching existing robot decision-making models with execution methods, such as execution logics and execution conditions, and execution details, such as potential action and tool usages, supporting better robot understanding of the human cognitive process in task execution. Based on both human cognition, understanding, and environment perception, a robot's surrounding environments in NLexe were interpreted as a human-centered situation. In this human-centered situation, task execution was interpreted from a human perspective, improving a robot's reasoning capability in cooperating with a human. However, feedback-based learning requires frequent human involvement, imposing a heavy cognitive burden on a human. Moreover, the knowledge learned from human feedback was given by a human without considering the robot's actual knowledge needs, limiting the robot's adaptation to new environments where its knowledge shortage was waiting to be compensated for successful NLexe.

Typical execution training systems using human feedback include the following. (1) Object arrangement robots with consideration of human-desired object clusters, such as "yellow objects, rectangle objects" [110]. (2) Real-time robot behavior correction with real-time human NL feedback on hand pose and object selections [69]. (3) Self-improved robots with human-guided failure learning during industrial assembly [109]. (4) Human-sentiment-considered robot executions with subjective NL rewards and punishments, such as "joy, anger" [132]. (5) Human NL feedback, such as "yes, I think so, Okay, let's do that", was used to confirm the robot's understanding, such as "We should send UAV to the upper region", to achieve a consistent understanding between humans and robots [134].

4.1.4. NL-Based Execution Training Systems Using Proactive Querying

To solve new-situation adaptation problems for further improving a robot's reasoning ability, execution training systems using proactive querying were designed, in which robots proactively learn from a human according to knowledge needs in specific task execution scenarios and robots decide what knowledge to learn. In this system, speech recognition, NL understanding, and generation are required, which brings the difficulty of being aware

of the current situation and generating a concise description of the knowledge a robot needs to learn. This system is designed for actively acquiring knowledge in uncertain conditions or physical help from a human. Typical tasks include a robot asking for the user's confirmation if the instruction is ambiguous and then performing a task with consent. In the querying process, a robot used NL to proactively query its human users about its missing knowledge related to human-intention disambiguation, environment interpretations, and knowledge-to-world mapping. After the training, a robot was endowed with more targeted knowledge to adapt to previously-encountered situations, thereby improving a robot's environmental adaptability. With a proactive querying manner, robots were endowed with an advanced self-improving capability during human-guided task execution. Supported by a never-ending learning mechanism, robot performances in NLexe were improved in the long term by continuous knowledge acquiring and refining [70].

For developing NL-based execution training systems using proactive querying, a challenging research problem is robot NL generation, which uses NL to generate questions to appropriately express a robot's knowledge needs to a human. Robot NL generation is challenging for the following reasons. (1) Human decision-making in human-guided execution is uncertain due to limited robot observations on the human cognitive process [135]. It is difficult for a robot to accurately infer a human's execution intentions [136]. (2) Self-evaluation mechanism for reasoning about a robot's knowledge proficiency is missing. It is difficult for a robot to reason about its missing knowledge by itself [137]. (3) NL questions for expressing the robot's task failures and knowledge needs are hard to organize in a concise and accurate way [138]. To address these problems, several solutions were proposed. (1) Use environmental context to reduce uncertainties in human execution request understanding, improving recognition accuracy of human intentions during human-instructed robot executions [71]. (2) Hierarchical knowledge structure representations were used for detecting missing knowledge in execution [6]. (3) Concise NL questions were generated by involving both human execution request understanding and missing knowledge filling [139].

Typical NL-based execution training systems using proactive robot querying include the following. (1) Ask for cognitive decisions on trajectory, action, and pose selections in tasks, such as "human-robot jointly carrying a bulky bumper" [140]. (2) Ask for knowledge disambiguation of human commands, such as confirming the human-attended object "the blue cup" [141,142]. (3) Ask for human physical assistance to deliver a missing object or to execute robot-incapable actions, such as "deliver a table leg for a robot" [109,139]. (4) Ask for additional information, such as "the object is yellow and rectangle", from a human to help with robot perception [46,110].

4.2. Open Problems

During the development of training methods starting from instruction training to querying training, the human cognition burden was gradually decreased, and the robot cognition level was gradually improved. Robot training using the above-mentioned methods is suffering the shortcomings of the robot knowledge scalability and adaptability. The knowledge scalability problem, which is caused by limited knowledge sources and limited knowledge learning methods, has been solved to some degree, improving a robot's execution intuitiveness and naturalness. However, some open problems for current NL-based execution training systems are still existing. (1) Learning from a human is time-consuming and labor-intensive. It is challenging to largely scale up robot knowledge in an economical manner. (2) Human knowledge is not always reliable, or at least different types of human knowledge have different degrees of reliabilities. It is really challenging for a robot to assess knowledge reliability and use knowledge differently in manners, such as different knowledge types, different knowledge amounts, and different knowledge implementation scenarios. Detailed comparisons among NL-based execution training systems including verbal instruction, demonstration, human feedback, and proactive query, and their instruction methods, roles, applications, and advantages are shown in Table 3.

Table 3. Summary of NL-based execution training systems.

	Verbal Instruction	Demonstration	Human Feedback	Proactive Query
Manner	spoken description	physical demonstration	verb/physical feedback	robot queries
Format	speech	speech, motion	speech, motion	speech
Human Role	instructor	demonstrator	leader	assistant
Robot Role	follower	follower	assistant	leader
Applications	object recognition, grasping	NL-features (force/color /size/visual) association, gesture learning, environment understanding	human-preferred execution learning, human-like manipulation learning, robot learns from failures	human-in-the-loop decision-making, human physical assistance, intention disambiguation
Advantages	completed property/process definition	make abstract and ambiguous NL instructions explicit and machine-executable.	human preference consideration, initial human-like cognitive process modeling	robot get knowledge in need, relatively strong environment adaptability
References	[107,111–114,143]	[108,117,125,126,128]	[69,110,132,144]	[135,139–142]

5. NL-Based Interactive Execution Systems

To fully cooperate with robots in both interactive information sharing and interactive physical executions, instead of mainly information sharing (NL-based execution training systems) or physical execution (NL-based execution control systems), NL-based interactive execution systems were developed. An NL-based interactive execution system used NL in the stage of task execution and combined NL instructions and other sensor information to understand the optimal approach to complete a task. In NLexe interactive execution systems, speech recognition, natural language understanding, and dialog management are needed for establishing effective communication between a human and a robot. NL generation and speech synthesis are used to have a conversation with users. The instruction format is complex and based on interactive descriptions, which contain logic correspondence and the use of pronouns to refer to a previous element already mentioned in the conversation/dialog. Simple instructions are provided by robots to remind a user to help the execution process in robot-centered interactive execution systems. Instructions with multiple sentences about the execution target, logic, and environment are provided by users while guiding a robot. Different from NL-based execution training systems, in which human NL was helping a robot with its task understanding, in NL-based execution systems, human NL was helping a robot with its task executions, in which the understanding of execution requests, working statuses of both robots and humans and the task progress is focused for physically integrating both robots and humans. In NL-based training, a robot created a structure-completed and execution-specified knowledge representation. But in NL-based task executions, including understanding the task, the robot was also required to understand the surrounding environments, predict human intentions, and make optimal decisions satisfying the constraints from the environment, task executions, robot capabilities, and human requirements. Typical interactive execution systems using NL-based task execution are shown in Figure 6. Given that the reasoning was strictly requested in NL-based task execution, robot cognition levels in NL-based task execution were higher than that in NL-based execution training.

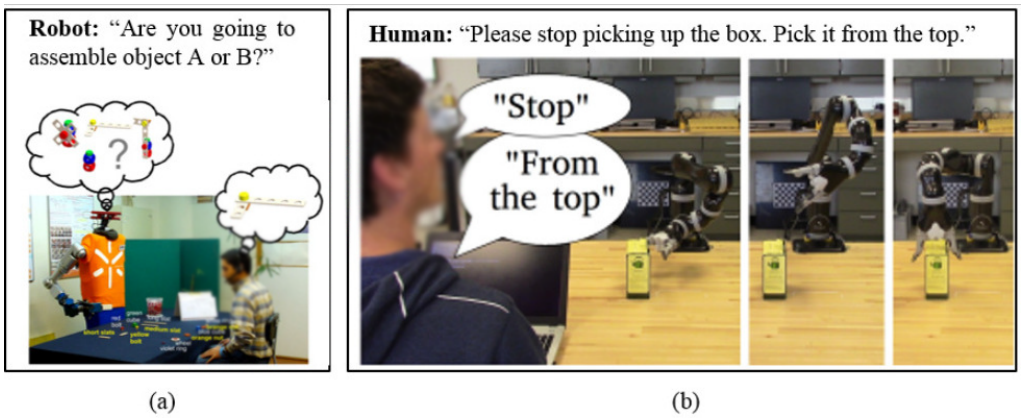


Figure 6. Typical NL-based interactive execution systems. (**a**) A human-centered execution system. A human was performing tasks, such as "assemble a toy". A robot was standing by and meanwhile prepared to provide help, ensuring the success and smoothness of the human's task executions. A robot was expected to infer the human's ongoing activities, detect human needs timely and proactively provide the appropriate help, such as "a toy part" [145]. (**b**) Robot-centered execution system. A robot was autonomously performing a task. A human was standing by to monitor the robot executions. If abnormal executions or execution failures occurred, the human-provided timely verbal corrections, such as "stop, grasp the top" or physical assistance, such as "delivering the robot-needed object" [146].

5.1. Typical NLexe Systems

With respect to who is leading the execution, NL-based interactive execution systems are categorized into human-centered task execution systems and robot-centered task execution systems.

5.1.1. Human-Centered Task Execution Systems

To integrate human mental intelligence and robot physical executions, human-centered task execution systems were designed, in which a human mentally leads the task executions, and a robot mainly provides appropriate physical assistance for facilitating human executions. In this system, speech recognition, NL understanding, and multi-sensor fusion are required, which brings the difficulty of closely monitoring human task processes by combining both NL and multi-sensor data to provide help when needed. This system is mainly designed to provide appropriate physical help to human users when asked. Typical tasks include assembling a device and asking for help like "give me a wrench" or "what is the next step" where robots provide assistance accordingly. NL expressions in task execution deliver information, such as explanations of a human's execution requests, descriptions of a human's execution plan, and indications of a human's urgent needs. With this information, a robot provides appropriate assistance timely. Correspondingly, a robot took on only physical responsibilities, such as "grasping and transferring the fragile objects" [147,148]. Both the human and the robot performed independent sub-tasks by sharing the same high-level task goal. The robot received fewer instructions for its tasks and meanwhile was expected to monitor the human's task processes so that the robot provided appropriate assistance when the human needs it. This execution proposed a relatively high standard towards the robot cognition on providing appropriate assistance at the right location and time. Overall, in the human-centered NL-based task execution, a human was leading the execution at the cognition level, and a robot provided the appropriate assistance for saving the human's time and energy, thereby enhancing the human's physical capability.

Typical NLexe systems with a human-centered task execution manner include the following. (1) Performing tasks, such as "table assembly", during which the human-made task goals (assembly of a specific part) and plans (action steps, pose and tool usages), and partially executes tasks (assemble the parts together), and the robot provides human-desired

assistances (tool delivery, part delivery, part holding) which were required verbally by a human user [149,150]. During the executions, the human took both cognitive and physical responsibilities, and the robot took partial physical responsibilities. (2) Comprehensive human-centered execution was developed so that a human user was only burdened with cognitive responsibilities, such as "explaining the navigation routine" [151,152], "describing the needed objects, location and pose" [153,154], and "guiding the fine and rough processing" [6,155].

5.1.2. Robot-Centered Task Execution Systems

To further improve execution intuitiveness and decrease a human's mental and physical burdens, robot-centered task execution systems were developed, in which a robot mentally leads the task executions, and the human mainly provides physical assistance for facilitating robot executions. In this system, speech recognition, NL understanding and generation, and multi-sensor fusion are required, which brings the difficulty of understanding the whole execution process and having the current situation considered to ask for human assistance when needed. This system is mainly designed for acquiring human help when robots are incapable of some type of task process. Typical tasks include robots moving heavy objects and asking human users for help with NL sentences/utterances like "not enough room to pass". Different from human-centered NLexe systems in which a human mainly took the cognitive and physical burdens while a robot gave human-needed assistances to facilitate human execution, in robot-centered systems, a robot mainly took the cognitive and physical burdens, while a human gives robot-needed assistances physically to facilitate the robot executions. NL expressions in the robot-centered systems were used for a robot to ask for assistance from a human. Compared with robots in human-centered NLexe systems, where a robot was required to comprehensively understand human behaviors, robots in robot-centered applications were required to comprehensively understand the limitations on robot knowledge, real-world conditions, and both humans' and robots' physical capabilities. The advantage was that the human was less involved and their hands and mind were partially set free.

Typical robot-centered NLexe systems include the following. (1) A Robot led the industrial assembly, in which a human enhanced a robot's physical capability by providing a robot with physical assistance, such as grasping [156] and fetching [157] (2) A robot executed tasks, such as object moving and elderly navigation in unstructured outdoor environments, in which a human analyzed and conquered the environment limitations, such as objects and space availability [158–160]. (3) By considering both human NL instructions, such as "go up, ascend, come back towards the tables and chairs", and indoor environment conditions, such as "door is open that path is available, a large circular pillar is in front that it is an obstacle", Micro-air vehicle (MAV) intuitively planed the path in an uncertain indoor environment with assistances of human NL instructions [161].

5.2. Open Problems

NL-based robot execution enables a robot to practically implement its knowledge in complex execution situations. A robot becomes situation-aware that robots' capability limits, human capability limits, and environmental conditions' constraints are analyzed by a robot to facilitate the execution with a human. However, limited by current techniques in both robotics and artificial intelligence, some open problems in NLexe still exist. (1) Robot cognitive levels are still low and human involvements are still intensive, bringing heavy cognitive burdens for a human [146]. (2) For a robot, it is difficult to understand human NL requests. This is because human NL expressions are abstract and implicit [162]. It is difficult to interpret high-level abstract human execution requests, such as "bring me a cup of water" into low-level robot-executable plans, such as "bring action: grasp; speed: 0.1 m/s; ....". It is also challenging to understand indicated meanings, such as "water type: iced water; placement location: table surface (x, y, z); ...", from implicit NL expressions, such as "bring me a cup of water" [44]. (3) Human intentions are difficult to infer for that human

intention is dynamic and implicit [163]. (4) It is challenging for a robot to generate NL execution requests. Execution issues related to task execution progress, human status, and robot knowledge gaps are hard to identify; appropriate NL questions covering execution issue description and solution querying are hard to generate [136]. Therefore, an NLexe system involving human cognitive process modeling, intelligent robot decision-making, autonomous robotic task execution, and human and robot physical capability consideration is in urgent need. As shown in Table 4, human-centered task execution and robot-centered task execution focus on different centers in NL-based interactive execution systems.

Table 4. Summary of NL-based interactive execution systems.

	Human-Centered Task Execution	**Robot-Centered Task Execution**
Human Involvements	high	low
Human's Job	cognitive burden	physical burden
Robot's Job	physical burden	cognitive burden
Human Involvement	leader	assistant
Robot Involvements	assistant	leader
Applications	assembly, object grasping	daily assistance, industrial assembly, outdoor navigation, heavy object delivering
Advantages	accurate assistance	environment adaptation
Disadvantages	high requirements on robot' reasoning capability on recognizing human activity	human cognitive burden is high
Typical References	[150–152,154,155]	[156–160,164,165]

6. NL-Based Social Execution Systems

To make the NLexe system socially natural and acceptable, NL-based social execution systems were developed by involving human social norms in both communication and task executions, as shown in Figure 7. An NL-based social execution system uses NL not only in task execution but also appropriately communicates with human users as social norms are under consideration. In NLexe social execution systems, speech recognition, natural language understanding with extra knowledge, dialog management, natural language generation, and speech synthesis are all required to understand both user's intention and social norms. The instruction format used in NLexe social execution systems is relatively more complex and based on an interactive conversation that contains logic correspondence and ambiguous descriptions that have implied referents. A robot needs to be accepted by a human, and it needs to understand what users say and react accordingly. Various users have different expressions which further increase the complexity. Different from NL-based interactive execution systems which merely consider task execution details, such as robot and human capabilities, working status, and task statuses—NL-based social execution systems consider social norms as well as execution details.

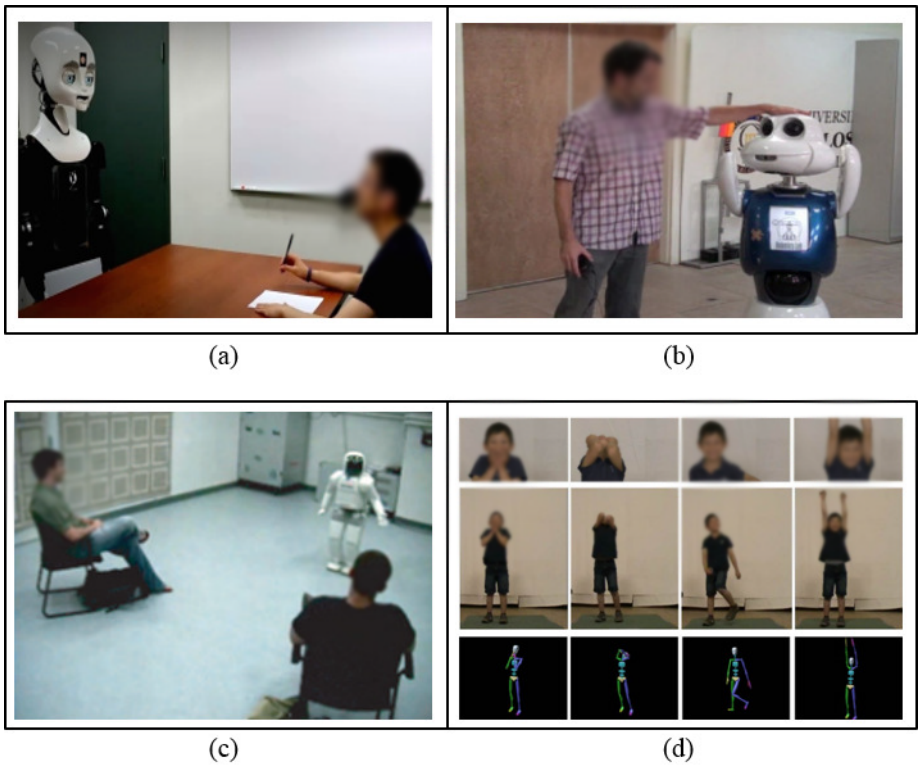


Figure 7. Typical NL-based social execution systems. (**a**,**b**) are NL-based social communication. A robot learned to nicely response to a human's request, such as "drawing a picture on the paper" [166] and "stop until I touch you" [167]. (**c**,**d**) are NL-based social execution. A robot learned to use appropriate body language during its speaking, such as storytelling [168,169].

6.1. Typical NLexe System

According to application scenarios of using social norms, NL-based social execution systems are categorized into social communication systems and social execution systems.

6.1.1. NL-Based Social Communication Systems

To make human-robot communication information correct and social-norm appropriate for natural NLexe, NL-based social communication systems were designed, in which social NL expressions are used for facilitating communication. In this system, speech recognition, highly capable NL understanding, and generation are required, which brings the difficulties of exploiting commonsense knowledge to understand both human instructions and environmental conditions and also generating reasonable NL for a response. This system is mainly designed for socially acceptable robot execution with consideration of human emotions. Typical tasks include robots using friendly expressions like "please" to have a conversation with users. Capturing social norms from humans' NL expressions was helpful in aspects, such as detecting human preferences in execution, specifying execution roles, such as "leader, follower, cooperator", and increasing social acceptance. NL in social communications served as an information source, from which both the objective execution methods and the subjective human preferences were extracted.

Typical NL-based social communication systems include the following. (1) A receptionist robot increased its social acceptance in conference arrangements by using social dialogs with pleasant prosodic contours [170]. (2) Cooperative machine operations used social cues, such as a subtle head nod, perk ears, friendly NL feedback "let's learn . . . , can you . . . , do task x . . . ," to indicate human execution preferences, such as "please press the red button", in task executions [171,172]. (3) Health-caregiving robots searched and delivered objects by considering user speech confidences, user safety, and user roles, such as "primary user, bystander" [173]. (4) Adapted unfamiliar users by using NL expressions with

fuzzy emotion statuses, such as "fuzzy happiness, sadness, or anger" [174]. (5) Modeled social NL communications in NLexe by defining human-robot relations, such as "love", "friendship", and "marriage" [175]. (6) A robotic doctor used friendly NL conversations to lead physiotherapy [43,176,177].

6.1.2. NL-Based Social Execution Systems

To make task executions social-norm appropriate for natural NLexe, NL-based social execution systems were designed, in which social NL expressions are used for facilitating executions. In this system, speech recognition, highly capable NL understanding, and environment understanding are required, which brings the difficulty of understanding implied user preferences in human NL and acting accordingly. This system is mainly designed to adapt a robot's behaviors to the current environment based on human NL and improve their social acceptance. Typical tasks include robots slowing their speed when the environment is crowded with the reminding of NL. NL was used to indicate socially preferred executions for robots, enhancing robots' understanding of social motivations behind task executions and further making robot executions socially acceptable.

With the NLexe systems, typical applications using NL-based social executions include the following. (1) A navigation robot autonomously modified its motion behaviors (stop, slower, faster) by considering human density (crowded, dense) with the reminding of human NL instructions ("go ahead to move", "stop") [178]. (2) A companion robot moved its head towards the human speaker according to a human's NL tones [179]. (3) A storytelling robot told stories by mapping NL expressions with a human's body motion behaviors to catch human attention [168,180]. (4) A bartender robot appropriately adjusted its serving orders in social manners, such as "serve customers one by one, notify the new-coming customers, proactively greet and serve, keep appropriate distance with customers", according to business situations [181]. (5) By detecting emotion statuses, such as "angry, confused, happy, and impatient", in human NL instructions, such as "left, right, forward, backward, down, and up", a surgical robot autonomously adjusted camera positions to prevent harms to patients during surgical operation [182,183].

6.2. Open Problems

NL-based social communication and NL-based social execution focused on two different aspects of NLexe. To develop socially-intuitive NLexe systems, the two aspects need to be developed simultaneously. Even though the introduction of social norms in NLexe systems increased systems' social acceptance in human-guided task execution, some open problems still exist, impeding performance improvements of NL-based social execution systems. (1) Social norms are too variable to be summarized. Different regions, countries, cultures, and races have different social norms. It is difficult to summarize representative norms from various social norms for supporting robot executions [168]. (2) Social norms are too implicit to technically learn. It is technically challenging to learn social norms from human behaviors [173]. (3) Last, social norms are currently non-evaluable. It is challenging to assess the correctness of social norms because there are no clear standards to judge the correctness of social norms. Different persons have different levels of social behavior acceptance and tolerance [171]. Detailed comparisons including the instruction manner and format, applications, and respective advantages among NL-based social execution systems are shown in Table 5.

Table 5. Summary of NL-based social execution systems.

	NL-Based Social Communication	NL-Based Social Execution
Instruction Manner	verbal	physical
Instruction Format	NL expressions	physical execution
Applications	restaurant receptionist, health caregiving, industrial execution, human-robot cooperation	social distance maintenance, storytelling, social expressions
Advantages	more social acceptance, customized	respect human's social preferences
Disadvantages	social expression development	user safety consideration
References	[170–174,184,185]	[178–182,186,187]

7. Methods and Realizations for Human Instruction Understanding

To support natural communications between robots and humans during task executions, the key technical challenges are the methods of human instruction understanding and system realization. With instruction understanding, robots extract task-related information from human instructions; with system realization, speech recognition systems are integrated into robotic systems to design NLexe systems.

7.1. Models for Human Instruction Understanding

An accurate understanding of human execution requests influences robot executions' accuracy and intuitiveness. With recent decades' developments, human instruction understanding has been developed from shallow-level literally request understanding to comprehensive-level interpretive request understanding. According to linguistic features involved in analyzing execution requests' semantic meanings, the understanding models are mainly categorized into two types: literal understanding model and interpreted understanding model.

7.1.1. Literal Understanding Model

Literal understanding models for execution request understanding use literal linguistic features, such as words, Part-of-Speech (PoS), word dependencies, word references, and sentence syntax structures, which are explicitly mentioned in human NL expression modality. PoS is a common NLP technique that categorizes each word in the text as corresponding to a particular part of speech based on the definition of the word and the context around it. Word dependency indicates the semantic relation between words in a sentence, which is used to identify the inner logic of the sentence. Word reference is introduced to resolute co-references for understanding pronoun references. Sentence syntax structures indicate the word order in a sentence like "subject + verb + object", and are used to analyze the causal relations. According to literal linguistic features' usage manners, literal understanding models are mainly categorized into predefined models, grammar models, and association models. For predefined models, symbolic literal linguistic features, such as keywords and PoS, are manually defined to model NL requests' meanings, realizing an initial interactive information sharing and execution between humans and robots.

Feature usage manners of predefinition models include the following. (1) Used keywords as triggers to identify task-related targets, such as 'book, person' [188]. (2) Used PoS tags, such as 'verb, the noun' to disambiguate word meanings, such as book (verb: reading materials; v: buying tickets) in polysemy situations [189]. (3) Used features, such as 'red color' to improve robot understanding accuracy in identifying human-desired objects, such as 'apple' [190]. For grammar models, grammar patterns, such as execution procedure 'V(go) + NN(Hallway), V(grasp) + NN(cup)', logic relation 'if(door open), then(turn right)', and spatial relation 'cup IN room', were manually defined to model NL requests' meanings, realizing a relatively improved robot adaptability towards various human NL expressions.

Feature usage manners of grammar models include the following. (1) Using action-object relations to define robot grasping methods, such as 'grasp(cup)' [191]. (2) Using action-location relations to describe robot navigation behaviors, such as 'go(Hallway), open(door), turn right (stairs)' [22]. (3) Using logic relations to define execution preconditions, such as 'if (door open), then (turn right)' [192]. (4) Using spatial relation to give execution suggestions, such as '(cup)CloseTo(plate)' for a robot [39].

For association models, semantic meanings, such as empirical explanations 'beverage: juice', quantitative dynamic values, such as '1 m/s' for NL expressions, such as 'quickly, slowly', and execution parameters, such as 'driller, upper-left corner' were manually defined for specifying NL expressions, such as 'drill a hole' [6], scaling up robot knowledge and further improving robot cognition levels. Typical feature usage manners of association models include the following. (1) Using probabilistic correlations, such as 'beverage-juice (probability 0.7)' to recommend empirical explanation for disambiguating human NL requests, such as 'delivery a beverage' [151]. (2) Using quantitative dynamic values to translate subjective NL expressions 'quickly, slowly' to sensor measurable values, such as '0.5–1 m/s' [193].

7.1.2. Interpreted Understanding Model

Interpreted understanding models for execution request understanding use implicitly interpreted linguistic features, which are not explicitly mentioned in human NL modality while are implicitly indicated by NL expressions. According to interpreted features' usage manners, interpreted understanding models are mainly categorized into single-modality models and multi-modality models.

In a single-modality model, interpreted linguistic features are only from the NL information modality. Implicit expressions, such as indicated objects/locations and commonsense-based logic, are inferred from explicit NL expressions for enriching information in human NL commands. Typical interpreted linguistic features include object function interpretations, such as 'cup: containing liquid', human fact interpretations, such as 'preferred action, head motion', and environment context interpretations, such as 'elevator (at right side)'. Typical feature usage manners of single-modality models include the following. (1) Combined explicitly-mentioned execution goals, such as 'drill, clean', with implicitly-mentioned execution details, such as 'drill: tool: driller. Action: move down, sweep, move up. Precondition: hole does not exist', to specify human-instructed abstract plans as robot-executable plans [44]. (2) Combined explicitly-mentioned human NL requests, such as 'take', with human physical statuses, such as 'human torso pose', to understand human visual perspective, such as 'gazed cup' [122].

In a multi-modality model, interpreted linguistic features are from other modalities, such as vision modality, motion modality, tactile modality as well as NL modality. Mutual correlations among modalities enrich and confirm the information embedded in human NL commands. Typical interpreted linguistic features include human tactile indication (tactile modality), human hand/body pose (vision modality and motion dynamics modality), and environmental conditions (environment context modality). Typical feature usage manners include the following. (1) Combined human NL expression 'person' with real-time RFID sensor values to identify individual identities [194]. (2) Combined NL expressions, such as 'hand over a glass of water', with a tactile event, such as 'handshaking', to confirm object exchange between a robot and a human [156]. (3) Combined NL requests, such as 'very good', with facial expressions, such as 'happy', to enable a social human-robot interaction [170]. (4) Combined NL commands, such as 'manipulate the screw on the table', with visual cues, such as 'screw color', to perform cooperative object manipulation tasks [195].

7.1.3. Model Discussion

For literal models, they are good at scenarios with simple execution procedures and clear work assignments, such as robot arm control and robot pose control. For interpreted

models, they are good at scenarios with involvements of daily commonsense, human cognitive logics, and rich domain information, such as object physical property assisted object searching, intuitive machine-executable plan generation, vision-verbal-motion-supported object delivery, etc. However, a literal model relies on a large amount of training data to define correlations between NL expressions and practical execution details. Learning these correlations is time-consuming and labor-intensive. An interpreted model's performance is limited by robot cognition levels and data fusion algorithms. Ref. [196] discussed the synergies between human knowledge and learnable knowledge, providing insights into the natural language model learning process.

7.2. System Realizations

7.2.1. NLP Techniques

Natural instruction understanding during NLexe is supported by NLP methods which build computational models for modeling semantic meanings of execution requests, NLP software libraries which are software tools for realizing NLP methods, and NLP dictionaries which provide NL vocabulary for supporting NLP methods. Usually, it has two phases: Preprocessing and modeling, which brings difficulties with noise reduction and semantic understanding. An NLP system aims to analyze texts' structure and understand natural language in multiple scenarios.

To be passed to an NLP model, data must be preprocessed to a standardized form. Commonly used preprocessing methods include (1) lemmatization, which groups together different inflected-form words, such as "belief, believing, believe, beliefs" into one word "belief", (2) stemming, which removes affixes, such as "es" from a word, such as "believes", and remains only the stem "believe", (3) noise removal, which replaces incorrect words, such as "gooooooood", to a correct word, such as "good", indexed in a dictionary, and (4) tokenization, which is then used to divide NL requests into separate sentences and words. Preprocessing techniques should be carefully applied for datasets with different topics or noisy levels.

Multiple traditional machine learning models are used for NLP tasks, and the current trending method is a deep neural network model. Since Recurrent Neural Networks (RNN) and more particularly long-short-term memory (LSTM) have been introduced [197], such neural architectures perform well for sequential data like text. More recently, a novel transformer network based on an attention mechanism was proposed with better performance in various tasks and faster training speed [198]. Pretrained word embedding models based on the transformer, like BERT (Bidirectional Encoder Representations from Transformers) have been developed for general NLP tasks and are fine-tuned to get state-of-the-art results in a wide range of tasks [199].

Typical NLP applications involved in NL request understanding include the following [200–202]. (1) Part-of-Speech (PoS) tagging and Chunk extraction, which labels words by PoS tags and extracts short phrases, such as [('the', 'DT'), ('book', 'NN')], from a part-of-speech tagged sentence. (2) Named entity extraction, which extracts the words referring to real-world objects, such as "person, organization, location". (3) Relation extraction, which extracts semantic relations such a "Beijing in China" among named entities. (4) Sentiment analysis, which analyzes positive and negative sentiments contained in human NL expressions, such as "Positive: your grasping is really good, robot. Negative: This grasping is really terrible".

Other techniques are used to examine the similarity of texts to check if two pieces of texts have the same meaning, including synset level similarity calculation, which measures similarities between two synsets [203], such as "cookbook, instruction book". and sentence level similarity calculation, which measures semantic meaning similarities between two sentences, such as "Deliver me a cup of water" and "please prepare me a cup of water.

Typical NL software libraries for analyzing NL's meaning in NLexe include Natural Language Toolkit (NLTK) [200], Standford CoreNLP [204], Apache OpenNLP [205], Apache Lucene [206], GATE [207], and spaCy [208]. Typical NLP dictionaries for supporting NLP

in NLexe include WordNet [203], ConceptNet5 [209], CMU Pronouncing Dictionary of American English [210], MRC Psycholinguistic Database [211], Word Frequency Data: Corpus of contemporary American English [212], CSLI The Verb Semantics Ontology Project [213], Leiden Weibo Corpus [214], and Spanish FrameNet [215].

7.2.2. Speech Recognition Systems

To finally realize human NL request understanding in practical human-guided task execution, speech recognition systems, such as DARE recognizer [216], HTK [217], Loquendo-ASR [167], CMU Sphinx [218], Kaldi [219], Julius [220], iATROS [221], and AWTH ASR [222], are involved in translating speech into text for further NLP analysis.

A typical speech recognition system receives audio signals from sensors and outputs a series of words that were spoken. A conventional system usually consists of several independently learned components: (1) A lexicon (pronunciation model) describes how each word is pronounced phonetically. Constructing a lexicon involves a selection of desired words that cover most of the scenarios. This linguistic resource will be used in the lexical decoding step of the speech recognition process for the mapping between words and phoneme models. Recent research also achieved lexicon-free speech recognition by mapping acoustic input to characters via neural networks [223]. (2) A previously trained acoustic model represents mappings between audio signals and phonemes and predicts which phoneme is being spoken. Many acoustic models were built based on the Hidden Markov Model (HMM), Deep Neural Networks (DNNs), or Convolutional Neural Network (CNN). (3) A language model describes word sequences. Statistical language models like N-grams make predictions of a word given the N-1 previous ones, while neural language models are based on neural networks and each word context (previous and next words). Both are trained on large text corpora. Neural networks use vector representation of words and their context called embeddings. A novel end-to-end speech recognition method without requiring an intermediate phonetic representation has also been proposed, which directly maps a sequence of audio signals into a sequence of words [224]. DNN-based speech recognition systems like [225–227] have dramatically improved the recognition accuracy and provided great prospects in the NLexe scenario.

To enable robots to the surrounding environment perceiving during NL communications, typical sensors were used as follows. (1) Human-speech-receiving sensors, such as microphones and Arduino-supported sound-detecting sensors, were used to translate human speech into text [228]. (2) Robot-speech-sending sensors, such as speakers, were used to speak out robot requests [228]. (3) Human-activity-monitoring sensors, such as RGB camera [229], RGB-depth camera [230], RFID sensors [14], tactile sensors [79], and motion tracking systems [231], were used to recognize human activities and to predict human intentions. (4) Environment perceiving sensors, such as temperature sensor [232], humidity sensor [232], sound intensity sensor [71], and distance sensor [71], were used to monitor a robot's surrounding environments.

In a real-world scenario, a speech recognition system will not have an ideal condition to handle various situations. Some aspects of the system are affected by factors such as noisy environment, sensor malfunction, and recognition error. Thus, the robust design of a system is essential. Recent speech recognition models utilizing neural networks are sensitive to unknown noisy conditions, and researches were conducted to solve this problem. The aurora-2 dataset is one of the speech recognition datasets with artificially added noise signals designed to help systems handle noisy environments [233]. Another research introduced a novel method using electroencephalography (EEG) to overcome performance loss in the presence of noise [234]. Because of the noise signal and sensor malfunction, recognition error is inevitable, so techniques were applied to handle data uncertainty. A novel model referred to as the Neutrosophic Convolutional Neural Network (NCNN) [235] was proposed to solve uncertainty handling. Another method is to detect and correct such errors automatically, and typical techniques using word error rate as metric were summarized in [236]. A novel approach explicitly corrects those errors by training a

spelling correction model [237]. From our perspective, future research could lead to a more advanced error tolerance system for promising development in human-robot interaction.

To realize a human-robot interaction via NL, a dialog management system takes input from speech recognition and human instruction understanding and exploits an external knowledge base to control the dialog flow. The process of dialog management includes two major challenges: Dialog modeling which keeps track of the current state of the dialog and Dialog Control which decides the next action of a system. A conversation objective is typically confirmed for a robot with a specific target and grounding is important to establish a shared understanding of the conversation scope. The user intent is then identified using the current user input and dialog state. A traditional dialog management method is handcrafted to have finite states where a rigid set of rules and transitions between states are predefined [238]. Furthermore, probabilistic models learn what to say next in a conversation by modeling real conversation data, and they are commonly used to realize a novel dialog management system [239]. In the NLexe scenario, a dialog management system is used to further improve the robustness of such a system in multiple aspects, such as (1) A robot learns simple board games like Connect Four from demonstrations using dialog management [240] and (2) services robots utilizing a friendly user interface and dialog management system to help older people [241].

7.2.3. Evaluations

To testify to the ability of an NLexe system, the same evaluation metric and baseline should be used. All the systems must deal with the same problem to ensure a fair comparison. The systems are evaluated in multiple phases including speech recognition, intent recognition, and task execution.

For speech recognition, current standard metrics are word error rate (WER) and sentence error rate (SER). Novel metrics for specific tasks are also proposed, such as using intuitive labels from human subjects and replacing human annotations with a machine-learning algorithm to evaluate voice search tasks [242]. Common datasets for speech recognition tasks include Switchboard Hub 500 [243], Fisher [244], and LibriSpeech [245], which consist of read English speech and text. Recent approaches combine multiple corpora to build a more robust system [226,246].

Intent recognition is vital to any task-oriented system. In the NLexe system, it mainly focuses on using human NL obtained from speech recognition to analyze the intention of a user. Datasets like SNIPS Natural Language Understanding benchmark (SNIPS-NLU) [247] and Spoken Language Understanding Resource Package (SLURP) [248] are used for benchmarking voice assistants and human-robot conversations. Simple datasets like the Airline Travel Information System (ATIS) are also used to build compact models for method verification.

Task execution evaluation methods vary according to the task a robot is designed to do. Ref. [249] provided a comprehensive review of HRI metrics in various task setups including navigation, management, manipulation, and social interaction. In addition, a user's subjective evaluation of NLP accuracy and naturalness is used to judge a system's performance. Objective robot performance improvement is another metric to decide whether a system is improved after introducing NL instructions.

For a human-robot interaction system, user studies should be conducted to provide training/test data or to evaluate a system. Multiple approaches are used to conduct a user study, including a crowdsourcing platform, online questionnaire, in-person volunteer interview, or interaction. Crowdsourcing currently involves using the internet to attract participants to achieve a cumulative result, and it has been widely used for human intrinsic tasks in human-robot interaction. The popular platform includes Amazon's Mechanical Turk [250] and gMission [251]. In-person volunteer interaction or public interaction is usually useful for observing the interaction process between robot and human, which is critical for social robot development as human users' reactions to a robot's behavior are collected and then analyzed. User study also may have drawbacks since individual

behavior is unpredictable and have a negative effect on a system, so the data from user study should be carefully preprocessed to exploit the result.

8. Emerging Trends of NLexe

NLexe has been developed to improve the effectiveness and naturalness of human-guided task executions. Due to the limitations of NLexe-related techniques, such as NLP, machine learning and robot design, NLexe performances in dealing with complex tasks, various users, and dynamic/unstructured environments still need to be improved. Recent transformer-based models for NLP have high flexibility in representing connections between words in a sequence and sentences in a document. Different pre-trained models were proposed to be fine-tuned and get state-of-the-art results in major NLP tasks [199,252]. A model figures out user intent with higher accuracy by fine-tuning these pre-trained models with specific task targets. For example, by fine-tuning a BERT model for named entity recognition (NER) task combined with a DCNN model, a robot understands what object a user is referred to [253]. Hence the comprehensive understanding of human instruction in different situations becomes possible [254]. And smaller models with the same level of performance are developed, so a robot with low energy consumption can use these novel techniques [255]. GPT-3 model shows great performance in natural language generation tasks as it is used to generate human-like language to improve social acceptance of a robot [256]. Based on our comprehensive review, the future trends for future NLexe research are summarized as follows.

8.1. Robot Knowledge Scalability

Scaling up robot knowledge to support robot decision-making is a critical issue in NLexe. On the one hand, to understand human NL instructions, plan tasks, or fill up the knowledge gaps, the effective knowledge scaling-up capability is needed to accurately learn a large amount of knowledge. On the other hand, the time and labor costs are expected to be reduced. Currently, the knowledge-scaling-up research goes in two directions: existed-knowledge exploitation, and new-knowledge exploration. In existed-knowledge exploitation, the abstract meanings of existing knowledge are summarized at a high level to increase knowledge interchangeability, making one type of knowledge useful in other similar scenarios. Pretrained models showed an ability to learn from a long article and catch the key point; thus are used to generate a large amount of knowledge for robot execution in real-time [257].

The new-knowledge exploration includes human-based methods, which query human users for new knowledge, and the big-data-based method, which is an automatic and low-cost information retrieval method that extracts knowledge from information sources, such as the World Wide Web [258], books [259], machine operation log files [260], and videos [261]. Typically, robot applications related to learning knowledge from the web include: (1) kitchen cooking robots by learning ontologies, such as manipulation poses, object involvements, and sequential action involvements, from Wikipedia and wikiHow [151,262,263]; (2) indoor object searching and delivery by learning statistical object-location correlations from general Word Wide Web pages [258]; (3) Human-centered assistive robots by learning human-involved 3D indoor scenes from 3D modeling websites, such as 3D warehouse [264].

It is promising to teach a robot to learn NLexe from the web since the information is rich and knowledgeable making the learning method efficient and cheap. In the meantime, the information is required to be reliable for robots to learn otherwise issues may arise in sensible domains. From our perspective, future research could be developing state-of-the-art NLP techniques, advanced learning algorithms, and information retrieval methods for a robot to learn related knowledge from information sources, which could be the World Wide Web (WWW), human verbal communication, other robot instructions, or tutorial handbooks. As high-quality knowledge sources are available with low access costs, robot commonsense development becomes feasible and low-cost. Knowledge learning

is a critically needed research direction for NL-supported human-robot interaction. The potential applications could be commonsense-supported robot assistance in the surgical operation room, manufacturing workshop, and indoor places.

8.2. Robot Adaptability

Weak robot adaptability is typically caused by the ignorance of execution importance, based on which the execution priority is made, and execution interchangeability modeling, based on which the execution flexibility is made. Different users with the same intention may have different language expressions, so word embedding like BERT provides a solution that similar words are resolved and identified correctly [265]. To increase robot adaptability, new research was launched to model the human cognition process [44,266], which aimed to explore humans' decision-making mechanisms for modeling robot execution priority and flexibility. For execution priority, not all executions are essential for execution success. For example, in the task "assembly", the procedure "clean the place" is much less important than the procedure "install the screw". For achieving interchangeability, a tool request "deliver me a brush" does not necessarily mean the involvement of a specific tool "brush", but instead means a practical purpose "cleaning the surface" [6]. By knowing these meanings, the execution plans are flexibly changed by ignoring the trivial execution procedures, focusing on the important procedures, and replacing the unavailable tools with the other available similar-function tools.

Current methods focus on exploring object affordances (object-action correlation) [267], and lacking in-depth interpretations of task execution. In the future, NLexe research could be methods that interpret robot executions from a human perspective, improving robot adaptability in unstructured environments and unfamiliar human users. The most important research is enabling robot adaptability in new situations based on available sensing and motion capability. The potential research could be knowledge transfer supported environment and task adaptation, in scenarios such as manipulating different-geometry objects and navigation in different building clusters.

8.3. Failure Learning

Sub-step execution failures cause unnatural task executions or even task failures. The learning-from-failure mechanism has been implemented in computer science for algorithm efficiency improvement [268], and in material science for new material discovery [269]. NLG model produced a description of problems or predefined error messages telling users what failures have happened and asking for help or clarification. Then by combining the error message and human advice, the solution is provided by users through natural language understanding [270]. By exploring useful information from failure experiences, robot capability could be improved to avoid similar failures in the future. In NLexe, learning from failure could be done by using NL to query humans and instruct a robot in both vogue NL question "help me" or detailed NL question "Please hand me the black table leg that is on the white table" [271].

Therefore, in NLexe, learning from failure is also a promising research direction. From our perspective, the potential research problems could be in-depth failure cause analysis, concise NL failure explaining to a human, and proactive knowledge updating methods for recovering from the failures. Explaining the failure and self-diagnosis of the failure could be a critically needed research direction. The research can enable quick robot performance improvement; it also can enhance the trust between robots and humans, where the robot status needs to be monitored and corrected in a timely manner.

8.4. Versatility and Transparency

Robot behavior under NL instructions should be versatile and transparent in NLexe. Being versatile brings the ability to perform multiple tasks in a dynamic environment. Being transparent brings the ability to explain why an action is performed, which increases human trust. Currently, understanding user intent and generating human-like language

are two main directions for HRI. The mainstream method in NLP for both tasks is transfer learning, which is to train a relatively sizeable generic model and fine-tune it to fit in tasks. State-of-the-art generative model GPT-3 trains a large generic model using numerous amounts of data. With no hint given, the model still performs well in most scenarios, which proves that generic models have the same or better performance than specified models citepbrown2020language. Current methods usually train a specified model for each task environment and lacking flexibility. From our perspective, developing a general model for language understanding and generation using a pre-trained model and reinforcement learning with augmented data is a promising research direction.

In NLexe, robot executions are based on NL instructions, and robots combine task information and execution progress to reply to the user. Therefore, an intermediate level is required, which allows the fusion of multi-source information and transparency of execution logic. Unlike a chatbot using a sequence-to-sequence model, separating understanding and generation models and using reinforcement learning for better transparency is promising in NLexe. Research using reinforcement learning was launched to improve both the language understanding and generation process when interacting with human users [272,273]. With a dedicated reward set, RL enables a robot with better performance and versatility. And methods were proposed to achieve interpretable reinforcement learning for generic robotic planning [274]. Babyai++ [275,276] presented reinforcement learning used in grounded-language learning. Therefore, it is promising to use reinforcement learning in understanding and generation models to improve versatility and transparency since a multi-purpose robot reduces the cost and improves the execution process. Future research could be general language models for instruction understanding, task execution in various environments, interpretable planning, and execution process, which significantly increase human trust in robot execution.

9. Conclusions

This review summarized the state-of-the-art natural-language-instructed robot execution (NLexe) systems. With an in-depth analysis of research motivations, system categories, technique supports, and emerging research trends, NLexe systems, which use NL to improve the accuracy and intuitiveness of human-guided robot executions, were summarized comprehensively.

For development motivations, NLexe's rapid development was summarized as the pushing results of both NLP developments and robot execution developments. NLP was for interactively sharing information during task execution. Robot execution was for physically executing tasks instructed by human NL.

For system design, according to robot cognition levels, NLexe systems were categorized into four main types: NL-based execution control systems, NL-based execution training systems, NL-based interactive execution systems, and NL-based social execution systems. NL-based execution control systems were mainly used to replace hands in the operations of robotic arms and navigation robots. NL-based execution training systems were mainly used to scale up robot knowledge by using NL to transfer human execution experiences in object manipulation, outdoor navigation, and industrial assembly, to robots. NL-based interactive execution systems were designed to practically guide robot executions with both interactive knowledge sharing and interactive physical executions in real-world situations. NL-based social execution systems were designed to make NLexe socially natural and acceptable. Humans' social norms in both communication and task executions are considered in human-guided task executions.

In the future, NLexe systems could be developed in directions introduced in "Emerging trends of NLexe". At a high level, robot knowledge scale, learning capability, environment understanding levels, and self-proficiency-evaluation capability will be improved in future NLexe research. NLexe systems will be developed toward human-guided task execution with more natural communications and more intuitive executions.

Author Contributions: Conceptualization, R.L., Y.G., R.J. and X.Z.; methodology, R.L., Y.G., R.J. and X.Z.; software, R.L., Y.G., R.J. and X.Z.; validation, R.L., Y.G., R.J. and X.Z.; formal analysis, R.L., Y.G., R.J. and X.Z.; investigation, R.L., Y.G., R.J. and X.Z.; resources, R.L., Y.G., R.J. and X.Z.; data curation, R.L., Y.G., R.J. and X.Z.; writing—original draft preparation, R.L., Y.G., R.J. and X.Z.; writing—review and editing, R.L., Y.G., R.J. and X.Z.; visualization, R.L., Y.G., R.J. and X.Z.; supervision, R.L.; project administration, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NL	Natural Language
NLexe	NL-instructed Robot Execution
NLP	Natural Language Processing
VAE	Variational Autoencoders
GAN	Generative Adversarial Network
HRI	Human-Robot Interaction
WWW	World Wide Web

References

1. Baraglia, J.; Cakmak, M.; Nagai, Y.; Rao, R.; Asada, M. Initiative in robot assistance during collaborative task execution. In Proceedings of the 11th IEEE International Conference on Human Robot Interaction, Christchurch, New Zealand, 7–10 March 2016; pp. 67–74. [CrossRef]
2. Gemignani, G.; Bastianelli, E.; Nardi, D. Teaching robots parametrized executable plans through spoken interaction. In Proceedings of the 2015 International Conference on Autonomous Agents and Multi-Agent Systems, Istanbul, Turkey, 4–8 May 2015; pp. 851–859.
3. Brooks, D.J.; Lignos, C.; Finucane, C.; Medvedev, M.S.; Perera, I.; Raman, V.; Kress-Gazit, H.; Marcus, M.; Yanco, H.A. Make it so: Continuous, flexible natural language interaction with an autonomous robot. In Proceedings of the AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012.
4. Fong, T.; Thorpe, C.; Baur, C. Collaboration, Dialogue, Human-Robot Interaction. *Robotics Research*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 255–266. [CrossRef]
5. Krüger, J.; Surdilovic, D. Robust control of force-coupled human–robot-interaction in assembly processes. *CIRP Ann.-Manuf. Technol.* **2008**, *57*, 41–44. [CrossRef]
6. Liu, R.; Webb, J.; Zhang, X. Natural-language-instructed industrial task execution. In Proceedings of the 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Charlotte, NC, USA, 21–24 August 2016; p. V01BT02A043. [CrossRef]
7. Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M.R.; Banerjee, A.G.; Teller, S.J.; Roy, N. Understanding natural language commands for robotic navigation and mobile manipulation. *Assoc. Adv. Artif. Intell.* **2011**, *1*, 2. [CrossRef]
8. Iwata, H.; Sugano, S. Human-robot-contact-state identification based on tactile recognition. *IEEE Trans. Ind. Electron.* **2005**, *52*, 1468–1477. [CrossRef]
9. Kjellström, H.; Romero, J.; Kragić, D. Visual object-action recognition: Inferring object affordances from human demonstration. *Comput. Vis. Image Underst.* **2011**, *115*, 81–90. [CrossRef]
10. Kim, S.; Jung, J.; Kavuri, S.; Lee, M. Intention estimation and recommendation system based on attention sharing. In Proceedings of the 26th International Conference on Neural Information Processing, Red Hook, NY, USA, 5–10 December 2013; pp. 395–402. [CrossRef]
11. Hu, N.; Englebienne, G.; Lou, Z.; Kröse, B. Latent hierarchical model for activity recognition. *IEEE Trans. Robot.* **2015**, *31*, 1472–1482. [CrossRef]
12. Barattini, P.; Morand, C.; Robertson, N.M. A proposed gesture set for the control of industrial collaborative robots. In Proceedings of the 21st International Symposium on Robot and Human Interactive Communication (RO-MAN), Paris, France, 9–13 September 2012; pp. 132–137. [CrossRef]
13. Jain, A.; Sharma, S.; Joachims, T.; Saxena, A. Learning preferences for manipulation tasks from online coactive feedback. *Int. J. Robot. Res.* **2015**, *34*, 1296–1313. [CrossRef]

14. Liu, R.; Zhang, X. Understanding human behaviors with an object functional role perspective for robotics. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *8*, 115–127. [CrossRef]
15. Ramirez-Amaro, K.; Beetz, M.; Cheng, G. Transferring skills to humanoid robots by extracting semantic representations from observations of human activities. *Artif. Intell.* **2017**, *247*, 95–118. [CrossRef]
16. Zampogiannis, K.; Yang, Y.; Fermüller, C.; Aloimonos, Y. Learning the spatial semantics of manipulation actions through preposition grounding. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation, Seattle, WA, USA, 26–30 May 2015; pp. 1389–1396. [CrossRef]
17. Takano, W.; Nakamura, Y. Action database for categorizing and inferring human poses from video sequences. *Robot. Auton. Syst.* **2015**, *70*, 116–125. [CrossRef]
18. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137. [CrossRef]
19. Raman, V.; Lignos, C.; Finucane, C.; Lee, K.C.; Marcus, M.; Kress-Gazit, H. *Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language*; Technical Report; University of Pennsylvania: Philadelphia, PA, USA, 2013. [CrossRef]
20. Hemachandra, S.; Walter, M.; Tellex, S.; Teller, S. Learning semantic maps from natural language descriptions. In Proceedings of the 2013 Robotics: Science and Systems IX Conference, Berlin, Germany, 24–28 June 2013. [CrossRef]
21. Duvallet, F.; Walter, M.R.; Howard, T.; Hemachandra, S.; Oh, J.; Teller, S.; Roy, N.; Stentz, A. Inferring maps and behaviors from natural language instructions. In *Experimental Robotics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 373–388. [CrossRef]
22. Matuszek, C.; Herbst, E.; Zettlemoyer, L.; Fox, D. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 403–415. [CrossRef]
23. Ott, C.; Lee, D.; Nakamura, Y. Motion capture based human motion recognition and imitation by direct marker control. In Proceedings of the IEEE International Conference on Humanoid Robots, Daejeon, Republic of Korea, 1–3 December 2008; pp. 399–405. [CrossRef]
24. Waldherr, S.; Romero, R.; Thrun, S. A gesture based interface for human-robot interaction. *Auton. Robot.* **2000**, *9*, 151–173. [CrossRef]
25. Dillmann, R. Teaching and learning of robot tasks via observation of human performance. *Robot. Auton. Syst.* **2004**, *47*, 109–116. [CrossRef]
26. Medina, J.R.; Shelley, M.; Lee, D.; Takano, W.; Hirche, S. Towards interactive physical robotic assistance: Parameterizing motion primitives through natural language. In Proceedings of the 21st International Symposium on Robot and Human Interactive Communication (RO-MAN), Paris, France, 9–13 September 2012; pp. 1097–1102. [CrossRef]
27. Hemachandra, S.; Walter, M.R. Information-theoretic dialog to improve spatial-semantic representations. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 5115–5121. [CrossRef]
28. Hunston, S.; Francis, G. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*; No. 4; John Benjamins Publishing: Amsterdam, The Netherlands, 2000. [CrossRef]
29. Bybee, J.L.; Hopper, P.J. *Frequency and the Emergence of Linguistic Structure*; John Benjamins Publishing: Amsterdam, The Netherlands, 2001; Volume 45. [CrossRef]
30. Yang, Y.; Li, Y.; Fermüller, C.; Aloimonos, Y. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 3686–3693. [CrossRef]
31. Cheng, Y.; Jia, Y.; Fang, R.; She, L.; Xi, N.; Chai, J. Modelling and analysis of natural language controlled robotic systems. *Int. Fed. Autom. Control.* **2014**, *47*, 11767–11772. [CrossRef]
32. Wu, C.; Lenz, I.; Saxena, A. Hierarchical semantic labeling for task-relevant rgb-d perception. In Proceedings of the 2014 Robotics: Science and Systems X Conference, Berkeley, CA, USA, 12–16 July 2014. [CrossRef]
33. Hemachandra, S.; Duvallet, F.; Howard, T.M.; Roy, N.; Stentz, A.; Walter, M.R. Learning models for following natural language directions in unknown environments. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 5608–5615. [CrossRef]
34. Tenorth, M.; Perzylo, A.C.; Lafrenz, R.; Beetz, M. The roboearth language: Representing and exchanging knowledge about actions, objects, and environments. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA), St. Paul, MN, USA, 14–18 May 2012; pp. 1284–1289. [CrossRef]
35. Pineau, J.; West, R.; Atrash, A.; Villemure, J.; Routhier, F. On the feasibility of using a standardized test for evaluating a speech-controlled smart wheelchair. *Int. J. Intell. Control. Syst.* **2011**, *16*, 124–131.
36. Granata, C.; Chetouani, M.; Tapus, A.; Bidaud, P.; Dupourqué, V. Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders. In Proceedings of the 19th International Symposium on Robot and Human Interactive Communication (RO-MAN), Viareggio, Italy, 13–15 September 2010; pp. 785–790. [CrossRef]
37. Stenmark, M.; Malec, J. A helping hand: Industrial robotics, knowledge and user-oriented services. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems Workshop: AI-based Robotics, Tokyo, Japan, 3–7 November 2013.

38. Schulz, R.; Talbot, B.; Lam, O.; Dayoub, F.; Corke, P.; Upcroft, B.; Wyeth, G. Robot navigation using human cues: A robot navigation system for symbolic goal-directed exploration. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 1100–1105. [CrossRef]
39. Boularias, A.; Duvallet, F.; Oh, J.; Stentz, A. Grounding spatial relations for outdoor robot navigation. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 1976–1982. [CrossRef]
40. Kory, J.; Breazeal, C. Storytelling with robots: Learning companions for preschool children's language development. In Proceedings of the 23rd International Symposium on Robot and Human Interactive Communication (RO-MAN), Edinburgh, UK, 25–29 August 2014; pp. 643–648. [CrossRef]
41. Salvador, M.J.; Silver, S.; Mahoor, M.H. An emotion recognition comparative study of autistic and typically-developing children using the zeno robot. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 6128–6133. [CrossRef]
42. Breazeal, C. Social interactions in hri: The robot view. *IEEE Trans. Syst. Man Cybern.* **2004**, *34*, 81–186. [CrossRef]
43. Belpaeme, T.; Baxter, P.; Greeff, J.D.; Kennedy, J.; Read, R.; Looije, R.; Neerincx, M.; Baroni, I.; Zelati, M.C. Child-robot interaction: Perspectives and challenges. In Proceedings of the International Conference on Social Robotics, Bristol, UK, 27–29 October 2013; pp. 452–459. [CrossRef]
44. Liu, R.; Zhang, X. Generating machine-executable plans from end-user's natural-language instructions. *Knowl.-Based Syst.* **2018**, *140*, 15–26. [CrossRef]
45. Alterovitz, R.; Sven, K.; Likhachev, M. Robot planning in the real world: Research challenges and opportunities. *Ai Mag.* **2016**, *37*, 76–84. [CrossRef]
46. Misra, D.K.; Sung, J.; Lee, K.; Saxena, A. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *Int. J. Robot. Res.* **2016**, *35*, 281–300. [CrossRef]
47. Twiefel, J.; Hinaut, X.; Borghetti, M.; Strahl, E.; Wermter, S. Using natural language feedback in a neuro-inspired integrated multimodal robotic architecture. In Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, NY, USA, 26–31 August 2016. [CrossRef]
48. Ranjan, N.; Mundada, K.; Phaltane, K.; Ahmad, S. A survey on techniques in nlp. *Int. J. Comput. Appl.* **2016**, *134*, 6–9. [CrossRef]
49. Kulić, D.; Croft, E.A. Safe planning for human-robot interaction. *J. Field Robot.* **2005**, *22*, 383–396. [CrossRef]
50. Tuffield, P.; Elias, H. The shadow robot mimics human actions. *Ind. Robot. Int. J.* **2003**, *30*, 56–60. [CrossRef]
51. He, J.; Spokoyny, D.; Neubig, G.; Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. In Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, 6–9 May 2019. Available online: https://openreview.net/forum?id=rylDfnCqF7 (accessed on 7 May 2024).
52. Guo, J.; Lu, S.; Cai, H.; Zhang, W.; Yu, Y.; Wang, J. Long text generation via adversarial training with leaked information. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
53. Ferreira, T.C.; Lee, C.v.; Miltenburg, E.v.; Krahmer, E. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 552–562. [CrossRef]
54. McColl, D.; Louie, W.-Y.G.; Nejat, G. Brian 2.1: A socially assistive robot for the elderly and cognitively impaired. *IEEE Robot. Autom. Mag.* **2013**, *20*, 74–83. [CrossRef]
55. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. In Proceedings of the 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13–15 September 2016; p. 125. Available online: https://dblp.org/rec/journals/corr/OordDZSVGKSK16.bib (accessed on 7 May 2024).
56. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.v.d.; Dieleman, S.; Kavukcuoglu, K. Efficient neural audio synthesis. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 2410–2419. Available online: https://dblp.org/rec/journals/corr/abs-1802-08435.bib (accessed on 7 May 2024).
57. Cid, F.; Moreno, J.; Bustos, P.; Núnez, P. Muecas: A multi-sensor robotic head for affective human robot interaction and imitation. *Sensors* **2014**, *14*, 7711. [CrossRef] [PubMed]
58. Ke, X.; Cao, B.; Bai, J.; Zhang, W.; Zhu, Y. An interactive system for humanoid robot shfr-iii. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420913787. [CrossRef]
59. Zhao, X.; Luo, Q.; Han, B. Survey on robot multi-sensor information fusion technology. In Proceedings of the 2008 7th World Congress on Intelligent Control and Automation, Chongqing, China, 25–27 June 2008; pp. 5019–5023. [CrossRef]
60. Denoyer, L.; Zaragoza, H.; Gallinari, P. Hmm-based passage models for document classification and ranking. In Proceedings of the European Conference on Information Retrieval, Darmstadt, Germany, 2001; pp. 126–135. Available online: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/hugoz_ecir01.pdf (accessed on 7 May 2024).
61. Busch, J.E.; Lin, A.D.; Graydon, P.J.; Caudill, M. Ontology-Based Parser for Natural Language Processing. U.S. Patent 7,027,974, 11 April 2006. Available online: https://aclanthology.org/J15-2006.pdf (accessed on 7 May 2024).
62. Alani, H.; Kim, S.; Millard, D.E.; Weal, M.J.; Hall, W.; Lewis, P.H.; Shadbolt, N.R. Automatic ontology-based knowledge extraction from web documents. *IEEE Intell. Syst.* **2003**, *18*, 14–21. [CrossRef]

63. Cambria, E.; Hussain, A. *Sentic Computing: Techniques, Tools, and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 2. [CrossRef]
64. Young, R.M. Story and discourse: A bipartite model of narrative generation in virtual worlds. *Interact. Stud.* **2007**, *8*, 177–208. [CrossRef]
65. Bex, F.J.; Prakken, H.; Verheij, B. Formalising argumentative story-based analysis of evidence. In Proceedings of the International Conference on Artificial Intelligence and Law, Stanford, CA, USA, 4–8 June 2007; pp. 1–10. [CrossRef]
66. Stenzel, A.; Chinellato, E.; Bou, M.A.T.; del Pobil, P.; Lappe, M.; Liepelt, R. When humanoid robots become human-like interaction partners: Corepresentation of robotic actions. *J. Exp. Psychol. Hum. Percept. Perform.* **2012**, *38*, 1073. [CrossRef] [PubMed]
67. Mitsunaga, N.; Smith, C.; Kanda, T.; Ishiguro, H.; Hagita, N. Adapting robot behavior for human-robot interaction. *IEEE Trans. Robot.* **2008**, *24*, 911–916. [CrossRef]
68. Bruce, A.; Nourbakhsh, I.; Simmons, R. The role of expressiveness and attention in human-robot interaction. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Washington, DC, USA, 11–15 May 2002; Volume 4, pp. 4138–4142. [CrossRef]
69. Staudte, M.; Crocker, M.W. Investigating joint attention mechanisms through spoken human–Robot interaction. *Cognition* **2011**, *120*, 268–291. [CrossRef] [PubMed]
70. Liu, R.; Zhang, X.; Webb, J.; Li, S. Context-specific intention awareness through web query in robotic caregiving. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 1962–1967. [CrossRef]
71. Liu, R.; Zhang, X.; Li, S. Use context to understand user's implicit intentions in activities of daily living. In Proceedings of the IEEE International Conference on Mechatronics and Automation (ICMA), Tianjin, China, 3–6 August 2014; pp. 1214–1219. [CrossRef]
72. Selman, B. Nri: Collaborative Research: Jointly Learning Language and Affordances. 2014. Available online: https://www.degruyter.com/document/doi/10.1515/9783110787719/html?lang=en (accessed on 7 May 2024).
73. Mooney, R. Nri: Robots that Learn to Communicate Through Natural Human Dialog. 2016. Available online: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1637736&HistoricalAwards=false (accessed on 7 May 2024).
74. Roy, N. Nri: Collaborative Research: Modeling and Verification of Language-Based Interaction. 2014. Available online: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1427030&HistoricalAwards=false (accessed on 7 May 2024).
75. University of Washington. Robotics and State Estimation Lab. 2017. Available online: http://rse-lab.cs.washington.edu/projects/language-grounding/ (accessed on 5 January 2017).
76. Lund University. Robotics and State Estimation Lab. 2017. Available online: http://rss.cs.lth.se/ (accessed on 5 January 2017).
77. Argall, B.D.; Chernova, S.; Veloso, M.; Browning, B. A survey of robot learning from demonstration. *Robot. Auton. Syst.* **2009**, *57*, 469–483. [CrossRef]
78. Bethel, C.L.; Salomon, K.; Murphy, R.R.; Burke, J.L. Survey of psychophysiology measurements applied to human-robot interaction. In Proceedings of the 16th International Symposium on Robot and Human Interactive Communication (RO-MAN), Jeju Island, Republic of Korea, 26–29 August 2007; pp. 732–737. [CrossRef]
79. Argall, B.D.; Billard, A.G. Survey of tactile human–robot interactions. *Robot. Auton. Syst.* **2010**, *58*, 1159–1176. [CrossRef]
80. House, B.; Malkin, J.; Bilmes, J. The voicebot: A voice controlled robot arm. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 4–9 April 2009; pp. 183–192. [CrossRef]
81. Stenmark, M.; Nugues, P. Natural language programming of industrial robots. In Proceedings of the International Symposium on Robotics (ISR), Seoul, Republic of Korea, 24–26 October 2013; pp. 1–5. [CrossRef]
82. Jain, D.; Mosenlechner, L.; Beetz, M. Equipping robot control programs with first-order probabilistic reasoning capabilities. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, 12–17 May 2009; pp. 3626–3631. [CrossRef]
83. Zelek, J.S. Human-robot interaction with minimal spanning natural language template for autonomous and tele-operated control. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Grenoble, France, 7–11 September 1997; Volume 1, pp. 299–305. [CrossRef]
84. Romano, J.M.G.; Camacho, E.F.; Ortega, J.G.; Bonilla, M.T. A generic natural language interface for task planning—Application to a mobile robot. *Control Eng. Pract.* **2000**, *8*, 1119–1133. [CrossRef]
85. Wang, B.; Li, Z.; Ding, N. Speech control of a teleoperated mobile humanoid robot. In Proceedings of the IEEE International Conference on Automation and Logistics (ICAL), Chongqing, China, 15–16 August 2011; pp. 339–344. [CrossRef]
86. Gosavi, S.D.; Khot, U.P.; Shah, S. Speech recognition for robotic control. *Int. J. Eng. Res. Appl.* **2013**, *3*, 408–413. Available online: https://www.ijera.com/papers/Vol3_issue5/BT35408413.pdf (accessed on 7 May 2024).
87. Tellex, S.; Roy, D. Spatial routines for a simulated speech-controlled vehicle. In Proceedings of the ACM SIGCHI/SIGART Conference on Human-Robot Interaction, Salt Lake City, UT, USA, 2–3 March 2006; pp. 156–163. [CrossRef]
88. Stiefelhagen, R.; Fugen, C.; Gieselmann, R.; Holzapfel, H.; Nickel, K.; Waibel, A. Natural human-robot interaction using speech, head pose and gestures. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sendai, Japan, 28 September–2 October 2004; Volume 3, pp. 2422–2427. [CrossRef]

89. Chen, S.; Kazi, Z.; Beitler, M.; Salganicoff, M.; Chester, D.; Foulds, R. Gesture-speech based hmi for a rehabilitation robot. In Proceedings of the IEEE Southeastcon'96: Bringing Together Education, Science and Technology, Tampa, FL, USA, 11–14 April 1996; pp. 29–36. [CrossRef]
90. Bischoff, R.; Graefe, V. Integrating vision, touch and natural language in the control of a situation-oriented behavior-based humanoid robot. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Tokyo, Japan, 12–15 October 1999; Volume 2, pp. 999–1004. [CrossRef]
91. Landau, B.; Jackendoff, R. Whence and whither in spatial language and spatial cognition? *Behav. Brain Sci.* **1993**, *16*, 255–265. [CrossRef]
92. Ferre, M.; Macias-Guarasa, J.; Aracil, R.; Barrientos, A. Voice command generation for teleoperated robot systems. In Proceedings of the 7th International Symposium on Robot and Human Interactive Communication (RO-MAN), Kagawa, Japan, 30 September–2 October 1998; p. 679685. Available online: https://www.academia.edu/65732196/Voice_command_generation_for_teleoperated_robot_systems (accessed on 7 May 2024).
93. Savage, J.; Hernández, E.; Vázquez, G.; Hernandez, A.; Ronzhin, A.L. Control of a Mobile Robot Using Spoken Commands. In Proceedings of the Conference Speech and Computer, St. Petersburg, Russia, 20–22 September 2004. Available online: https://workshops.aapr.at/wp-content/uploads/2019/05/ARW-OAGM19_24.pdf (accessed on 7 May 2024).
94. Jayawardena, C.; Watanabe, K.; Izumi, K. Posture control of robot manipulators with fuzzy voice commands using a fuzzy coach–player system. *Adv. Robot.* **2007**, *21*, 293–328. [CrossRef]
95. Antoniol, G.; Cattoni, R.; Cettolo, M.; Federico, M. Robust speech understanding for robot telecontrol. In Proceedings of the International Conference on Advanced Robotics, Tokyo, Japan, 8–9 November 1993; pp. 205–209. Available online: https://www.researchgate.net/publication/2771643_Robust_Speech_Understanding_for_Robot_Telecontrol (accessed on 7 May 2024).
96. Levinson, S.; Zhu, W.; Li, D.; Squire, K.; Lin, R.-s.; Kleffner, M.; McClain, M.; Lee, J. Automatic language acquisition by an autonomous robot. In Proceedings of the International Joint Conference on Neural Networks, Portland, OR, USA, 20–24 July 2003; Volume 4, pp. 2716–2721. [CrossRef]
97. Scioni, E.; Borghesan, G.; Bruyninckx, H.; Bonfè, M. Bridging the gap between discrete symbolic planning and optimization-based robot control. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 5075–5081. [CrossRef]
98. Lallée, S.; Yoshida, E.; Mallet, A.; Nori, F.; Natale, L.; Metta, G.; Warneken, F.; Dominey, P.F. Human-robot cooperation based on interaction learning. In *From Motor Learning to Interaction Learning in Robots*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 491–536. [CrossRef]
99. Allen, J.; Duong, Q.; Thompson, C. Natural language service for controlling robots and other agents. In Proceedings of the International Conference on Integration of Knowledge Intensive Multi-Agent Systems, Waltham, MA, USA, 18–21 April 2005; pp. 592–595. [CrossRef]
100. Fainekos, G.E.; Kress-Gazit, H.; Pappas, G.J. Temporal logic motion planning for mobile robots. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Barcelona, Spain, 18–22 April 2005; pp. 2020–2025. [CrossRef]
101. Thomason, J.; Zhang, S.; Mooney, R.J.; Stone, P. Learning to interpret natural language commands through human-robot dialog. In Proceedings of the International Joint Conferences on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 1923–1929. Available online: https://dblp.org/rec/conf/ijcai/ThomasonZMS15.bib (accessed on 7 May 2024).
102. Oates, T.; Eyler-Walker, Z.; Cohen, P. *Using Syntax to Learn Semantics: An Experiment in Language Acquisition with a Mobile Robot*; Technical Report; University of Massachusetts Computer Science Department: Amherst, MA, USA, 1999. Available online: https://www.researchgate.net/publication/2302747_Using_Syntax_to_Learn_Semantics_An_Experiment_in_Language_Acquisition_with_a_Mobile_Robot (accessed on 7 May 2024).
103. Stenmark, M.; Malec, J.; Nilsson, K.; Robertsson, A. On distributed knowledge bases for robotized small-batch assembly. *IEEE Trans. Autom. Sci. Eng.* **2015**, *12*, 519–528. [CrossRef]
104. Vogel, A.; Raghunathan, K.; Krawczyk, S. A Situated, Embodied Spoken Language System for Household Robotics. 2009. Available online: https://cs.stanford.edu/~rkarthik/Spoken%20Language%20System%20for%20Household%20Robotics.pdf (accessed on 7 May 2024).
105. Nordmann, A.; Wrede, S.; Steil, J. Modeling of movement control architectures based on motion primitives using domain-specific languages. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 5032–5039. [CrossRef]
106. Bollini, M.; Tellex, S.; Thompson, T.; Roy, N.; Rus, D. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 481–495. [CrossRef]
107. Kruijff, G.-J.M.; Kelleher, J.D.; Berginc, G.; Leonardis, A. Structural descriptions in human-assisted robot visual learning. In Proceedings of the ACM SIGCHI/SIGART Conference on Human-Robot Interaction, Salt Lake City, UT, USA, 2–3 March 2006; pp. 343–344. [CrossRef]
108. Salem, M.; Kopp, S.; Wachsmuth, I.; Joublin, F. Towards an integrated model of speech and gesture production for multi-modal robot behavior. In Proceedings of the 19th International Symposium on Robot and Human Interactive Communication (RO-MAN), Viareggio, Italy, 13–15 September 2010; pp. 614–619. [CrossRef]
109. Knepper, R.A.; Tellex, S.; Li, A.; Roy, N.; Rus, D. ecovering from failure by asking for help. *Auton. Robot.* **2015**, *39*, 347–362. [CrossRef]

110. Dindo, H.; Zambuto, D. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010; pp. 790–796. [CrossRef]
111. Cuayáhuitl, H. Robot learning from verbal interaction: A brief survey. In Proceedings of the New Frontiers in Human-Robot Interaction, Canterbury, UK, 21–22 April 2015. Available online: https://www.cs.kent.ac.uk/events/2015/AISB2015/proceedings/hri/14-Cuayahuitl-robotlearningfrom.pdf (accessed on 7 May 2024).
112. Yu, C.; Ballard, D.H. On the integration of grounding language and learning objects. In Proceedings of the 19th National Conference on Artificial Intelligence, San Jose, CA, USA, 25–29 July 2004; Volume 4, pp. 488–493. Available online: https://dl.acm.org/doi/abs/10.5555/1597148.1597228 (accessed on 7 May 2024).
113. Nicolescu, M.; Mataric, M.J. Task learning through imitation and human-robot interaction. In *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*; Cambridge University Press: Cambridge, UK, 2007; pp. 407–424. [CrossRef]
114. Roy, D. Learning visually grounded words and syntax of natural spoken language. *Evol. Commun.* **2000**, *4*, 33–56. [CrossRef]
115. Lauria, S.; Bugmann, G.; Kyriacou, T.; Bos, J.; Klein, A. Training personal robots using natural language instruction. *IEEE Intell. Syst.* **2001**, *16*, 38–45. [CrossRef]
116. Nicolescu, M.N.; Mataric, M.J. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia, 14–18 July 2003; pp. 241–248. [CrossRef]
117. Sugiura, K.; Iwahashi, N. Learning object-manipulation verbs for human-robot communication. In Proceedings of the 2007 Workshop on Multimodal Interfaces in Semantic Interaction, Nagoya, Japan, 15 November 2007; pp. 32–38. [CrossRef]
118. Kordjamshidi, P.; Hois, J.; van Otterlo, M.; Moens, M.-F. Learning to interpret spatial natural language in terms of qualitative spatial relations. In *Representing Space in Cognition: Interrelations of Behavior, Language, and Formal Models*; Oxford University Press: Oxford, UK, 2013; pp. 115–146. [CrossRef]
119. Iwahashi, N. Robots that learn language: A developmental approach to situated human-robot conversations. In *Human-Robot Interaction*; IntechOpen: London, UK, 2007; pp. 95–118. [CrossRef]
120. Yi, D.; Howard, T.M.; Goodrich, M.A.; Seppi, K.D. Expressing homotopic requirements for mobile robot navigation through natural language instructions. In Proceedings of the International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; pp. 1462–1468. [CrossRef]
121. Paul, R.; Arkin, J.; Roy, N.; Howard, T.M. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In Proceedings of the 2016 Robotics: Science and Systems XII Conference, Ann Arbor, MI, USA, 18–22 June 2016. [CrossRef]
122. Uyanik, K.F.; Calskan, Y.; Bozcuoglu, A.K.; Yuruten, O.; Kalkan, S.; Sahin, E. Learning social affordances and using them for planning. In Proceedings of the Annual Meeting of the Cognitive Science Society, Berlin, Germany, 31 July–3 August 2013; Volume 35, No. 35. Available online: https://escholarship.org/uc/item/9cj412wg (accessed on 7 May 2024).
123. Holroyd, A.; Rich, C. Using the behavior markup language for human-robot interaction. In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, Boston, MA, USA, 5–8 March 2012; pp. 147–148. [CrossRef]
124. Arumugam, D.; Karamcheti, S.; Gopalan, N.; Wong, L.L.; Tellex, S. Accurately and efficiently interpreting human-robot instructions of varying granularities. In Proceedings of the 2017 Robotics: Science and Systems XIII Conference, Cambridge, MA, USA, 12–16 July 2017. [CrossRef]
125. Montesano, L.; Lopes, M.; Bernardino, A.; Santos-Victor, J. Modeling affordances using bayesian networks. In Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Diego, CA, USA, 29 October–2 November 2007; pp. 4102–4107. [CrossRef]
126. Matuszek, C.; Bo, L.; Zettlemoyer, L.; Fox, D. Learning from unscripted deictic gesture and language for human-robot interactions. In Proceedings of the AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 2556–2563. [CrossRef]
127. Forbes, M.; Chung, M.J.-Y.; Cakmak, M.; Zettlemoyer, L.; Rao, R.P. Grounding antonym adjective pairs through interaction. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction—Workshop on Humans and Robots in Asymmetric Interactions, Bielefeld, Germany, 3–6 March 2014; pp. 1–4. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7016190/ (accessed on 7 May 2024).
128. Krause, E.A.; Zillich, M.; Williams, T.E.; Scheutz, M. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In Proceedings of the AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 2796–2802. [CrossRef]
129. Chai, J.Y.; Fang, R.; Liu, C.; She, L. Collaborative language grounding toward situated human-robot dialogue. *AI Mag.* **2016**, *37*, pp. 32–45. [CrossRef]
130. Liu, C.; Yang, S.; Saba-Sadiya, S.; Shukla, N.; He, Y.; Zhu, S.-C.; Chai, J. Jointly learning grounded task structures from language instruction and visual demonstration. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1482–1492. [CrossRef]

131. Williams, T.; Briggs, G.; Oosterveld, B.; Scheutz, M. Going beyond literal command-based instructions: Extending robotic natural language interaction capabilities. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 1387–1393. [CrossRef]
132. Bannat, A.; Blume, J.; Geiger, J.T.; Rehrl, T.; Wallhoff, F.; Mayer, C.; Radig, B.; Sosnowski, S.; Kühnlenz, K. A multimodal human-robot-dialog applying emotional feedbacks. In Proceedings of the International Conference on Social Robotics, Singapore, 23–24 November 2010; pp. 1–10. [CrossRef]
133. Thomaz, A.L.; Breazeal, C. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artif. Intell.* **2008**, *172*, 716–737. [CrossRef]
134. Savage, J.; Rosenblueth, D.A.; Matamoros, M.; Negrete, M.; Contreras, L.; Cruz, J.; Martell, R.; Estrada, H.; Okada, H. Semantic reasoning in service robots using expert systems. *Robot. Auton. Syst.* **2019**, *114*, 77–92. [CrossRef]
135. Brick, T.; Scheutz, M. Incremental natural language processing for hri. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Arlington, VA, USA, 10–12 March 2007; pp. 263–270. [CrossRef]
136. Gkatzia, D.; Lemon, O.; Rieser, V. Natural language generation enhances human decision-making with uncertain information. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 264–268. [CrossRef]
137. Hough, J. Incremental semantics driven natural language generation with self-repairing capability. In Proceedings of the Student Research Workshop Associated with RANLP, Hissar, Bulgaria, 13 September 2011; pp. 79–84. Available online: https://aclanthology.org/R11-2012/ (accessed on 7 May 2024).
138. Koller, A.; Petrick, R.P. Experiences with planning for natural language generation. *Comput. Intell.* **2011**, *27*, 23–40. [CrossRef]
139. Tellex, S.; Knepper, R.; Li, A.; Rus, D.; Roy, N. Asking for help using inverse semantics. In Proceedings of the 2014 Robotics: Science and Systems X Conference, Berkeley, CA, USA, 12–16 July 2014. [CrossRef]
140. Medina, J.R.; Lawitzky, M.; Mörtl, A.; Lee, D.; Hirche, S. An experience-driven robotic assistant acquiring human knowledge to improve haptic cooperation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 2416–2422. [CrossRef]
141. Sugiura, K.; Iwahashi, N.; Kawai, H.; Nakamura, S. Situated spoken dialogue with robots using active learning. *Adv. Robot.* **2011**, *25*, 2207–2232. [CrossRef]
142. Whitney, D.; Rosen, E.; MacGlashan, J.; Wong, L.L.; Tellex, S. Reducing errors in object-fetching interactions through social feedback. In Proceedings of the International Conference on Robotics and Automation, Singapore, 29 May–3 June 2017. [CrossRef]
143. Thomason, J.; Padmakumar, A.; Sinapov, J.; Walker, N.; Jiang, Y.; Yedidsion, H.; Hart, J.; Stone, P.; Mooney, R.J. Improving grounded natural language understanding through human-robot dialog. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 6934–6941. [CrossRef]
144. Alok, A.; Gupta, R.; Ananthakrishnan, S. Design considerations for hypothesis rejection modules in spoken language understanding systems. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 8049–8053. [CrossRef]
145. Bicho, E.; Louro, L.; Erlhagen, W. Integrating verbal and nonverbal communication in a dynamic neural field architecture for human-robot interaction. *Front. Neurorobot.* **2010**, *4*, 5. [CrossRef] [PubMed]
146. Broad, A.; Arkin, J.; Ratliff, N.; Howard, T.; Argall, B.; Graph, D.C. Towards real-time natural language corrections for assistive robots. In Proceedings of the Robotics: Science and Systems Workshop on Model Learning for Human-Robot Communication, Ann Arbor, MI, USA, 18–22 June 2016. Available online: https://journals.sagepub.com/doi/full/10.1177/0278364917706418 (accessed on 7 May 2024).
147. Deits, R.; Tellex, S.; Thaker, P.; Simeonov, D.; Kollar, T.; Roy, N. Clarifying commands with information-theoretic human-robot dialog. *J. Hum.-Robot. Interact.* **2013**, *2*, 58–79. [CrossRef]
148. Rybski, P.E.; Stolarz, J.; Yoon, K.; Veloso, M. Using dialog and human observations to dictate tasks to a learning robot assistant. *Intell. Serv. Robot.* **2008**, *1*, 159–167. [CrossRef]
149. Dominey, P.F.; Mallet, A.; Yoshida, E. Progress in programming the hrp-2 humanoid using spoken language. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Roma, Italy, 10–14 April 2007; pp. 2169–2174. [CrossRef]
150. Profanter, S.; Perzylo, A.; Somani, N.; Rickert, M.; Knoll, A. Analysis and semantic modeling of modality preferences in industrial human-robot interaction. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 1812–1818. [CrossRef]
151. Lu, D.; Chen, X. Interpreting and extracting open knowledge for human-robot interaction. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 686–695. [CrossRef]
152. Thomas, B.J.; Jenkins, O.C. Roboframenet: Verb-centric semantics for actions in robot middleware. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), St. Paul, MN, USA, 14–18 May 2012; pp. 4750–4755. [CrossRef]
153. Ovchinnikova, E.; Wachter, M.; Wittenbeck, V.; Asfour, T. Multi-purpose natural language understanding linked to sensorimotor experience in humanoid robots. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids), Seoul, Republic of Korea, 3–5 November 2015; pp. 365–372. [CrossRef]
154. Burger, B.; Ferrané, I.; Lerasle, F.; Infantes, G. Two-handed gesture recognition and fusion with speech to command a robot. *Auton. Robot.* **2012**, *32*, 129–147. [CrossRef]

155. Fong, T.; Nourbakhsh, I.; Kunz, C.; Fluckiger, L.; Schreiner, J.; Ambrose, R.; Burridge, R.; Simmons, R.; Hiatt, L.; Schultz, A.; et al. The peer-to-peer human-robot interaction project. *Space* **2005**, 6750. [CrossRef]
156. Bischoff, R.; Graefe, V. Dependable multimodal communication and interaction with robotic assistants. In Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, Berlin, Germany, 27 September 2002; pp. 300–305. [CrossRef]
157. Clodic, A.; Alami, R.; Montreuil, V.; Li, S.; Wrede, B.; Swadzba, A. A study of interaction between dialog and decision for human-robot collaborative task achievement. In Proceedings of the 16th International Symposium on Robot and Human Interactive Communication (RO-MAN), Jeju Island, Republic of Korea, 26–29 August 2007; pp. 913–918. [CrossRef]
158. Ghidary, S.S.; Nakata, Y.; Saito, H.; Hattori, M.; Takamori, T. Multi-modal human robot interaction for map generation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Maui, HI, USA, 29 October–3 November 2001; Volume 4, pp. 2246–2251. [CrossRef]
159. Kollar, T.; Tellex, S.; Roy, D.; Roy, N. Grounding verbs of motion in natural language commands to robots. In *Experimental Robotics*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 31–47. [CrossRef]
160. Bos, J. Applying automated deduction to natural language understanding. *J. Appl. Log.* **2009**, *7*, 100–112. [CrossRef]
161. Huang, A.S.; Tellex, S.; Bachrach, A.; Kollar, T.; Roy, D.; Roy, N. Natural language command of an autonomous micro-air vehicle. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010; pp. 2663–2669. [CrossRef]
162. Moore, R.K. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 281–291. [CrossRef]
163. Sakita, K.; Ogawara, K.; Murakami, S.; Kawamura, K.; Ikeuchi, K. Flexible cooperation between human and robot by interpreting human intention from gaze information. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sendai, Japan, 28 September–2 October 2004; Volume 1, pp. 846–851. [CrossRef]
164. Abioye, A.O.; Prior, S.D.; Thomas, G.T.; Saddington, P.; Ramchurn, S.D. The multimodal speech and visual gesture (msvg) control model for a practical patrol, search, and rescue aerobot. In Proceedings of the Annual Conference towards Autonomous Robotic Systems, Bristol, UK, 25–27 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 423–437. [CrossRef]
165. Schiffer, S.; Hoppe, N.; Lakemeyer, G. Natural language interpretation for an interactive service robot in domestic domains. In Proceedings of the International Conference on Agents and Artificial Intelligence, Algarve, Portugal, 6–8 February 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 39–53. [CrossRef]
166. Strait, M.; Briggs, P.; Scheutz, M. Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. In Proceedings of the International Symposium on New Frontiers in Human Robot Interaction, Canterbury, UK, 21–22 April 2015. Available online: https://hrilab.tufts.edu/publications/straitetal15aisb/ (accessed on 7 May 2024).
167. Gorostiza, J.F.; Salichs, M.A. Natural programming of a social robot by dialogs. In Proceedings of the Association for the Advancement of Artificial Intelligence Fall Symposium: Dialog with Robots, Arlington, VA, USA, 11–13 November 2010. Available online: https://dblp.org/rec/conf/aaaifs/GorostizaS10.bib (accessed on 7 May 2024).
168. Mutlu, B.; Forlizzi, J.; Hodgins, J. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots, Genova, Italy, 4–6 December 2006; pp. 518–523. [CrossRef]
169. Wang, W.; Athanasopoulos, G.; Yilmazyildiz, S.; Patsis, G.; Enescu, V.; Sahli, H.; Verhelst, W.; Hiolle, A.; Lewis, M.; Cañamero, L.C. Natural emotion elicitation for emotion modeling in child-robot interactions. In Proceedings of the WOCCI, Singapore, 19 September 2014; pp. 51–56. Available online: http://www.isca-speech.org/archive/wocci_2014/wc14_051.html (accessed on 7 May 2024).
170. Breazeal, C.; Aryananda, L. Recognition of affective communicative intent in robot-directed speech. *Auton. Robot.* **2002**, *12*, 83–104. [CrossRef]
171. Lockerd, A.; Breazeal, C. Tutelage and socially guided robot learning. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sendai, Japan, 28 September–2 October 2004; Volume 4, pp. 3475–3480. [CrossRef]
172. Breazeal, C. Toward sociable robots. *Robot. Auton. Syst.* **2003**, *42*, 167–175. [CrossRef]
173. Severinson-Eklundh, K.; Green, A.; Hüttenrauch, H. Social and collaborative aspects of interaction with a service robot. *Robot. Auton. Syst.* **2003**, *42*, 223–234. [CrossRef]
174. Austermann, A.; Esau, N.; Kleinjohann, L.; Kleinjohann, B. Fuzzy emotion recognition in natural speech dialogue. In Proceedings of the 24th International Symposium on Robot and Human Interactive Communication (RO-MAN), Kobe, Japan, 31 August–4 September 2005; pp. 317–322. [CrossRef]
175. Coeckelbergh, M. You, robot: On the linguistic construction of artificial others. *AI Soc.* **2011**, *26*, 61–69. [CrossRef]
176. Read, R.; Belpaeme, T. How to use non-linguistic utterances to convey emotion in child-robot interaction. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Boston, MA, USA, 5–8 March 2012; pp. 219–220. [CrossRef]
177. Kruijff-Korbayová, I.; Baroni, I.; Nalin, M.; Cuayáhuitl, H.; Kiefer, B.; Sanna, A. Children's turn-taking behavior adaptation in multi-session interactions with a humanoid robot. *Int. J. Humanoid Robot.* **2013**, *11*, 1–27. Available online: https://schiaffonati.faculty.polimi.it/TFI/ijhr.pdf (accessed on 7 May 2024).
178. Sabanovic, S.; Michalowski, M.P.; Simmons, R. Robots in the wild: Observing human-robot social interaction outside the lab. In Proceedings of the IEEE International Workshop on Advanced Motion Control, Auckland, New Zealand, 22–24 April 2016; pp. 596–601. [CrossRef]

179. Okuno, H.G.; Nakadai, K.; Kitano, H. Social interaction of humanoid robot based on audio-visual tracking. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Cairns, Australia, 17–20 June 2002; pp. 725–735. [CrossRef]
180. Chella, A.; Barone, R.E.; Pilato, G.; Sorbello, R. An emotional storyteller robot. In Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium: Emotion, Personality, and Social Behavior, Stanford, CA, USA, 26–28 March 2008; pp. 17–22. Available online: https://dblp.org/rec/conf/aaaiss/ChellaBPS08.bib (accessed on 7 May 2024).
181. Petrick, R. Extending the knowledge-level approach to planning for social interaction. In Proceedings of the 31st Workshop of the UK Planning and Scheduling Special Interest Group, Edinburgh, Scotland, UK, 29–30 January 2014; p. 2. Available online: http://plansig2013.org/ (accessed on 7 May 2024).
182. Schuller, B.; Rigoll, G.; Can, S.; Feussner, H. Emotion sensitive speech control for human-robot interaction in minimal invasive surgery. In Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (RO-MAN), Munich, Germany, 1–3 August 2008; pp. 453–458. [CrossRef]
183. Schuller, B.; Eyben, F.; Can, S.; Feussner, H. Speech in minimal invasive surgery-towards an affective language resource of real-life medical operations. In Proceedings of the 3rd Intern. Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, Valletta, Malta, 17–23 May 2010; pp. 5–9. Available online: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W24.pdf (accessed on 7 May 2024).
184. Romero-González, C.; Martínez-Gómez, J.; García-Varea, I. Spoken language understanding for social robotics. In Proceedings of the 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Ponta Delgada, Portugal, 15–17 April 2020; pp. 152–157. [CrossRef]
185. Logan, D.E.; Breazeal, C.; Goodwin, M.S.; Jeong, S.; O'Connell, B.; Smith-Freedman, D.; Heathers, J.; Weinstock, P. Social robots for hospitalized children. *Pediatrics* **2019**, *144*. [CrossRef] [PubMed]
186. Hong, J.H.; Taylor, J.; Matson, E.T. Natural multi-language interaction between firefighters and fire fighting robots. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, 11–14 August 2014; Volume 3, pp. 183–189. [CrossRef]
187. Fernández-Llamas, C.; Conde, M.A.; Rodríguez-Lera, F.J.; Rodríguez-Sedano, F.J.; García, F. May i teach you? Students' behavior when lectured by robotic vs. human teachers. *Comput. Hum. Behav.* **2018**, *80*, 460–469. [CrossRef]
188. Fry, J.; Asoh, H.; Matsui, T. Natural dialogue with the jijo-2 office robot. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Victoria, BC, Canada, 13–17 October 1998; Volume 2, pp. 1278–1283. [CrossRef]
189. Lee, K.W.; Kim, H.-R.; Yoon, W.C.; Yoon, Y.-S.; Kwon, D.-S. Designing a human-robot interaction framework for home service robot. In Proceedings of the 14th International Symposium on Robot and Human Interactive Communication (RO-MAN), Nashville, Tennessee, 13–15 August 2005; pp. 286–293. [CrossRef]
190. Hsiao, K.-y.; Vosoughi, S.; Tellex, S.; Kubat, R.; Roy, D. Object schemas for responsive robotic language use. In Proceedings of the ACM/IEEE International Conference on Human Robot Interaction, Amsterdam, The Netherlands, 12–15 March 2008; pp. 233–240. [CrossRef]
191. Motallebipour, H.; Bering, A. A Spoken Dialogue System to Control Robots. 2002. Available online: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=3129332&fileOId=3129339 (accessed on 7 May 2024).
192. McGuire, P.; Fritsch, J.; Steil, J.J.; Rothling, F.; Fink, G.A.; Wachsmuth, S.; Sagerer, G.; Ritter, H. Multi-modal human-machine communication for instructing robot grasping tasks. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Lausanne, Switzerland, 30 September–4 October 2002; Volume 2, pp. 1082–1088. [CrossRef]
193. Zender, H.; Jensfelt, P.; Mozos, O.M.; Kruijff, G.-J.M.; Burgard, W. An integrated robotic system for spatial understanding and situated interaction in indoor environments. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22–26 July 2007; Volume 7, pp. 1584–1589. Available online: https://dblp.org/rec/conf/aaai/ZenderJMKB07.bib (accessed on 7 May 2024).
194. Foster, M.E.; By, T.; Rickert, M.; Knoll, A. Human-robot dialogue for joint construction tasks. In Proceedings of the International Conference on Multimodal Interfaces, Banff, AB, Canada, 2–4 November 2006; pp. 68–71. [CrossRef]
195. Dominey, P.F. Spoken language and vision for adaptive human-robot cooperation. In *Humanoid Robots: New Developments*; IntechOpen: London, UK, 2007. [CrossRef]
196. Ranaldi, L.; Pucci, G. Knowing knowledge: Epistemological study of knowledge in transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]
197. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
198. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: New York, NY, USA; pp. 5998–6008. Available online: https://dblp.org/rec/conf/nips/VaswaniSPUJGKP17.bib (accessed on 7 May 2024).
199. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. 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), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
200. Perkins, J. *Python Text Processing with NLTK 2.0 Cookbook*; Packt Publishing Ltd.: Birmingham, UK, 2010. Available online: https://dl.acm.org/doi/10.5555/1952104 (accessed on 7 May 2024).

201. Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V. Gate: An architecture for development of robust hlt applications. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 168–175. [CrossRef]
202. Jurafsky, D.; Martin, J.H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. In *Prentice Hall Series in Artificial Intelligence*; Prentice Hall: Saddle River, NJ, USA, 2009; pp. 1–1024. Available online: https://dblp.org/rec/books/lib/JurafskyM09.bib (accessed on 7 May 2024).
203. Fellbaum, C. Wordnet. 2010. Available online: https://link.springer.com/chapter/10.1007/978-90-481-8847-5_10#citeas (accessed on 7 May 2024).
204. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The stanford corenlp natural language processing toolkit. In Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 22–27 June 2014; pp. 55–60. [CrossRef]
205. Foundation, A.S. Opennlp Natural Language Processing Library. 2017. Available online: http://opennlp.apache.org (accessed on 5 January 2017).
206. McCandless, M.; Hatcher, E.; Gospodnetic, O. *Lucene in Action: Covers Apache Lucene 3.0*; Manning Publications Co.: Shelter Island, NY, USA, 2010.
207. Cunningham, H. Gate, a general architecture for text engineering. *Comput. Humanit.* **2002**, *36*, 223–254. [CrossRef]
208. Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. 2017. Available online: https://spacy.io (accessed on 5 January 2017).
209. Speer, R.; Chin, J.; Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31. [CrossRef]
210. Weide, R. The Carnegie Mellon Pronouncing Dictionary of American English. 2014. Available online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict (accessed on 5 January 2017).
211. Wilson, M. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behav. Res. Methods Instrum. Comput.* **1988**, *20*, 6–10. [CrossRef]
212. Davies, M. Word Frequency Data: Most Frequent 100,000 Word Forms in English (Based on Data from the Coca Corpus). 2011. Available online: http://www.wordfrequency.info/ (accessed on 5 January 2017).
213. Beth, L.; John, S.; Bonnie, D.; Martha, P.; Timothy, C.; Charles, F. Verb Semantics Ontology Project. 2011. Available online: http://lingo.stanford.edu/vso/ (accessed on 5 January 2017).
214. Daan, V.E. Leiden Weibo Corpus. 2012. Available online: http://lwc.daanvanesch.nl/ (accessed on 5 January 2017).
215. Carlos, S.-R. *Spanish Framenet: A Frame-Semantic Analysis of the Spanish Lexicon.(w:) Multilingual Framenets in Computational Lexicography: Methods and Applications.(red.) Hans Boas*; Mouton de Gruyter: Berlin, Germany; New York, NY, USA, 2009; pp. 135–162. Available online: https://www.researchgate.net/publication/230876727_Spanish_Framenet_A_frame-semantic_analysis_of_the_Spanish_lexicon (accessed on 5 January 2017).
216. Lee, S.; Kim, C.; Lee, J.; Noh, H.; Lee, K.; Lee, G.G. Affective effects of speech-enabled robots for language learning. In Proceedings of the Spoken Language Technology Workshop (SLT), Berkeley, CA, USA, 12–15 December 2010; pp. 145–150. [CrossRef]
217. Majdalawieh, O.; Gu, J.; Meng, M. An htk-developed hidden markov model (hmm) for a voice-controlled robotic system. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sendai, Japan, 28 September–2 October 2004; Volume 4, pp. 4050–4055. [CrossRef]
218. Tikhanoff, V.; Cangelosi, A.; Metta, G. Integration of speech and action in humanoid robots: Icub simulation experiments. *IEEE Trans. Auton. Ment. Dev.* **2011**, *3*, 17–29. [CrossRef]
219. Linssen, J.; Theune, M. R3d3: The rolling receptionist robot with double dutch dialogue. In Proceedings of the Companion of the ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017; pp. 189–190. [CrossRef]
220. Mitsunaga, N.; Miyashita, T.; Ishiguro, H.; Kogure, K.; Hagita, N. Robovie-iv: A communication robot interacting with people daily in an office. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; pp. 5066–5072. [CrossRef]
221. Sinyukov, D.A.; Li, R.; Otero, N.W.; Gao, R.; Padir, T. Augmenting a voice and facial expression control of a robotic wheelchair with assistive navigation. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), San Diego, CA, USA, 5–8 October 2014; pp. 1088–1094. [CrossRef]
222. Nikalaenka, K.; Hetsevich, Y. Training Algorithm for Speaker-Independent Voice Recognition Systems Using Htk. 2016. Available online: https://elib.bsu.by/bitstream/123456789/158753/1/Nikalaenka_Hetsevich.pdf (accessed on 7 May 2024).
223. Maas, A.; Xie, Z.; Jurafsky, D.; Ng, A.Y. Lexicon-free conversational speech recognition with neural networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 345–354. [CrossRef]
224. Graves, A.; Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1764–1772. Available online: https://dblp.org/rec/conf/icml/GravesJ14.bib (accessed on 7 May 2024).
225. Xiong, W.; Wu, L.; Alleva, F.; Droppo, J.; Huang, X.; Stolcke, A. The microsoft 2017 conversational speech recognition system. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5934–5938. [CrossRef]

226. Saon, G.; Kurata, G.; Sercu, T.; Audhkhasi, K.; Thomas, S.; Dimitriadis, D.; Cui, X.; Ramabhadran, B.; Picheny, M.; Lim, L.-L.; et al. English conversational telephone speech recognition by humans and machines. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 132–136. [CrossRef]
227. Synnaeve, G.; Xu, Q.; Kahn, J.; Grave, E.; Likhomanenko, T.; Pratap, V.; Sriram, A.; Liptchinsky, V.; Collobert, R. End-to-end asr: From supervised to semi-supervised learning with modern architectures. In Proceedings of the Workshop on Self-Supervision in Audio and Speech (SAS) at the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020. Available online: https://dblp.org/rec/journals/corr/abs-1911-08460.bib (accessed on 7 May 2024).
228. Graciarena, M.; Franco, H.; Sonmez, K.; Bratt, H. Combining standard and throat microphones for robust speech recognition. *IEEE Signal Process. Lett.* **2003**, *10*, 72–74. [CrossRef]
229. Lauria, S.; Bugmann, G.; Kyriacou, T.; Klein, E. Mobile robot programming using natural language. *Robot. Auton. Syst.* **2002**, *38*, 171–181. [CrossRef]
230. Sung, J.; Ponce, C.; Selman, B.; Saxena, A. Unstructured human activity detection from rgbd images. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA), St. Paul, MN, USA, 14–18 May 2012; pp. 842–849. [CrossRef]
231. Tenorth, M.; Bandouch, J.; Beetz, M. The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In Proceedings of the International Conference on Computer Vision Workshops (ICCV), Kyoto, Japan, 27 September–4 October 2009; pp. 1089–1096. [CrossRef]
232. Nehmzow, U.; Walker, K. Quantitative description of robot–environment interaction using chaos theory. *Robot. Auton. Syst.* **2005**, *53*, 177–193. [CrossRef]
233. Hirsch, H.-G.; Pearce, D. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In Proceedings of the ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop (ITRW), Pairs, France, 18–20 September 2000; pp. 181–188. [CrossRef]
234. Krishna, G.; Tran, C.; Yu, J.; Tewfik, A.H. Speech recognition with no speech or with noisy speech. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 1090–1094. [CrossRef]
235. Rashno, E.; Akbari, A.; Nasersharif, B. A convolutional neural network model based on neutrosophy for noisy speech recognition. In Proceedings of the 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), Tehran, Iran, 6–7 March 2019. [CrossRef]
236. Errattahi, R.; Hannani, A.E.; Ouahmane, H. Automatic speech recognition errors detection and correction: A review. *Procedia Comput. Sci.* **2018**, *128*, 32–37. [CrossRef]
237. Guo, J.; Sainath, T.N.; Weiss, R.J. A spelling correction model for end-to-end speech recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5651–5655. [CrossRef]
238. Abella, A.; Gorin, A.L. Method for Dialog Management. U.S. Patent 8,600,747, 3 December 2013. Available online: https://patentimages.storage.googleapis.com/05/ba/43/94a73309a3c9ef/US8600747.pdf (accessed on 7 May 2024).
239. Lu, D.; Zhang, S.; Stone, P.; Chen, X. Leveraging commonsense reasoning and multimodal perception for robot spoken dialog systems. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 6582–6588. [CrossRef]
240. Zare, M.; Ayub, A.; Wagner, A.R.; Passonneau, R.J. Show me how to win: A robot that uses dialog management to learn from demonstrations. In Proceedings of the 14th International Conference on the Foundations of Digital Games, San Luis Obispo, CA, USA, 26–30 August 2019; pp. 1–7. [CrossRef]
241. Jayawardena, C.; Kuo, I.H.; Unger, U.; Igic, A.; Wong, R.; Watson, C.I.; Stafford, R.; Broadbent, E.; Tiwari, P.; Warren, J.; et al. Deployment of a service robot to help older people. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010; pp. 5990–5995. [CrossRef]
242. Levit, M.; Chang, S.; Buntschuh, B.; Kibre, N. End-to-end speech recognition accuracy metric for voice-search tasks. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 5141–5144. [CrossRef]
243. Godfrey, J.J.; Holliman, E. Switchboard-1 release 2 ldc97s62. In *Philadelphia: Linguistic Data Consortium*; The Trustees of the University of Pennsylvania: Philadelphia, PA, USA, 1993. [CrossRef]
244. Cieri, C.; Graff, D.; Kimball, O.; Miller, D.; Walker, K. Fisher english training speech part 1 transcripts ldc2004t19. In *Philadelphia: Linguistic Data Consortium*; The Trustees of the University of Pennsylvania: Philadelphia, PA, USA, 2004. [CrossRef]
245. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210. [CrossRef]
246. Xiong, W.; Droppo, J.; Huang, X.; Seide, F.; Seltzer, M.; Stolcke, A.; Yu, D.; Zweig, G. Toward human parity in conversational speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2410–2423. [CrossRef]
247. Coucke, A.; Saade, A.; Ball, A.; Bluche, T.; Caulier, A.; Leroy, D.; Doumouro, C.; Gisselbrecht, T.; Caltagirone, F.; Lavril, T.; et al. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *arXiv* **2018**, arXiv:1805.10190. Available online: https://dblp.org/rec/journals/corr/abs-1805-10190.bib (accessed on 7 May 2024).

248. Bastianelli, E.; Vanzo, A.; Swietojanski, P.; Rieser, V. SLURP: A Spoken Language Understanding Resource Package. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7252–7262, Association for Computational Linguistics. [CrossRef]
249. Steinfeld, A.; Fong, T.; Kaber, D.; Lewis, M.; Scholtz, J.; Schultz, A.; Goodrich, M. Common metrics for human-robot interaction. In Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, Salt Lake City, UT, USA, 2–3 March 2006; pp. 33–40. [CrossRef]
250. Buhrmester, M.; Kwang, T.; Gosling, S.D. Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? In *Methodological Issues and Strategies in Clinical Research*; American Psychological Association: Washington, DC, USA, 2016; pp. 133–139. [CrossRef]
251. Chen, Z.; Fu, R.; Zhao, Z.; Liu, Z.; Xia, L.; Chen, L.; Cheng, P.; Cao, C.C.; Tong, Y.; Zhang, C.J. Gmission: A general spatial crowdsourcing platform. In Proceedings of the VLDB Endowment, Hangzhou, China, 1–5 September 2014; Volume 7, pp. 1629–1632. [CrossRef]
252. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9. Available online: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 7 May 2024).
253. Hatori, J.; Kikuchi, Y.; Kobayashi, S.; Takahashi, K.; Tsuboi, Y.; Unno, Y.; Ko, W.; Tan, J. Interactively picking real-world objects with unconstrained spoken language instructions. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 3774–3781. [CrossRef]
254. Patki, S.; Daniele, A.F.; Walter, M.R.; Howard, T.M. Inferring compact representations for efficient natural language understanding of robot instructions. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 6926–6933. [CrossRef]
255. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020. Available online: https://openreview.net/forum?id=H1eA7AEtvS (accessed on 7 May 2024).
256. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Glasgow, UK, 2020; Volume 33, pp. 1877–1901. Available online: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (accessed on 7 May 2024).
257. Dai, Z.; Callan, J. Deeper text understanding for ir with contextual neural language modeling. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 985–988. [CrossRef]
258. Massouh, N.; Babiloni, F.; Tommasi, T.; Young, J.; Hawes, N.; Caputo, B. Learning deep visual object models from noisy web data: How to make it work. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5564–5571. [CrossRef]
259. Ronzano, F.; Saggion, H. Knowledge extraction and modeling from scientific publications. In *Semantics, Analytics, Visualization. Enhancing Scholarly Data*; González-Beltrán, A., Osborne, F., Peroni, S., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 11–25. [CrossRef]
260. Liu, C. Automatic discovery of behavioral models from software execution data. *IEEE Trans. Autom. Sci. Eng.* **2018**, *15*, 1897–1908. [CrossRef]
261. Liu, R.; Zhang, X.; Zhang, H. Web-video-mining-supported workflow modeling for laparoscopic surgeries. *Artif. Intell. Med.* **2016**, *74*, 9–20. [CrossRef] [PubMed]
262. Kawakami, T.; Morita, T.; Yamaguchi, T. Building wikipedia ontology with more semi-structured information resources. In *Semantic Technology*; Wang, Z., Turhan, A.-Y., Wang, K., Zhang, X., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 3–18. [CrossRef]
263. Liu, R.; Zhang, X. Context-specific grounding of web natural descriptions to human-centered situations. *Knowl.-Based Syst.* **2016**, *111*, 1–16. [CrossRef]
264. Chaudhuri, S.; Ritchie, D.; Wu, J.; Xu, K.; Zhang, H. Learning generative models of 3d structures. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2020; Volume 39, pp. 643–666. [CrossRef]
265. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [CrossRef]
266. Tanevska, A.; Rea, F.; Sandini, G.; Cañamero, L.; Sciutti, A. A cognitive architecture for socially adaptable robots. In Proceedings of the 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), Oslo, Norway, 19–22 August 2019; pp. 195–200. [CrossRef]
267. Koppula, H.S.; Saxena, A. Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 14–29. [CrossRef] [PubMed]
268. MacGlashan, J.; Ho, M.K.; Loftin, R.; Peng, B.; Wang, G.; Roberts, D.L.; Taylor, M.E.; Littman, M.L. Interactive learning from policy-dependent human feedback. In Proceedings of the 34th International Conference on Machine Learning—Volume 70, ICML'17, Sydney, NSW, Australia, 6–11 August 2017; pp. 2285–2294, JMLR.org. Available online: https://dblp.org/rec/conf/icml/MacGlashanHLPWR17.bib (accessed on 7 May 2024).

269. Raccuglia, P.; Elbert, K.C.; Adler, P.D.; Falk, C.; Wenny, M.B.; Mollo, A.; Zeller, M.; Friedler, S.A.; Schrier, J.; Norquist, A.J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73. [CrossRef] [PubMed]
270. Ling, H.; Fidler, S. Teaching machines to describe images with natural language feedback. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Glasgow, UK, 2017; Volume 30.
271. Honig, S.; Oron-Gilad, T. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Front. Psychol.* **2018**, *9*, 861. [CrossRef]
272. Ritschel, H.; André, E. Shaping a social robot's humor with natural language generation and socially-aware reinforcement learning. In Proceedings of the Workshop on NLG for Human—Robot Interaction, Tilburg, The Netherlands, 31 December 2018; pp. 12–16. [CrossRef]
273. Shah, P.; Fiser, M.; Faust, A.; Kew, C.; Hakkani-Tur, D. Followmet: Robot navigation by following natural language directions with deep reinforcement learning. In Proceedings of the Third Machine Learning in Planning and Control of Robot Motion Workshop at ICRA, Brisbane, Australia, 21–25 May 2018.
274. Li, X.; Serlin, Z.; Yang, G.; Belta, C. A formal methods approach to interpretable reinforcement learning for robotic planning. *Sci. Robot.* **2019**, *4*, eaay6276. [CrossRef]
275. Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T.H.; Bengio, Y. Babyai: A platform to study the sample efficiency of grounded language learning. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://dblp.org/rec/conf/iclr/Chevalier-Boisvert19.bib (accessed on 7 May 2024).
276. Cao, T.; Wang, J.; Zhang, Y.; Manivasagam, S. Babyai++: Towards grounded-language learning beyond memorization. In Proceedings of the ICLR 2020 Workshop: Beyond Tabula Rasa in RL, Addis Ababa, Ethiopia, 26–30 April 2020. Available online: https://dblp.org/rec/journals/corr/abs-2004-07200.bib (accessed on 7 May 2024).