INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A CONTEMPORARY SURVEY ON INTELLIGENT HUMAN-ROBOT INTERFACES FOCUSED ON NATURAL LANGUAGE PROCESSING

Ioannis Giachos[1], Dimitrios Piromalis[2], Michail Papoutsidakis[3], Stavros Kaminaris[4], Evangelos C. Papakitsos[5]

[1]University of West Attica, giachosg@uniwa.gr
[2] University of West Attica, piromali@uniwa.gr
[3] University of West Attica, mipapou@uniwa.gr
[4] University of West Attica, skamin@uniwa.gr
[5] University of West Attica, papakitsev@uniwa.gr
Author Correspondence: UNIWA, Thivon 250, 12244 Egaleo, Greece, +30 2105381810, papakitsev@uniwa.gr

Abstract: - In the evolution of robotic technology, a machine must be able to hear, understand, think, act and speak. This makes it much easier for people to work with robots, both as partners and by using them to cover their own weaknesses. In this paper, through the study of a series of works in Human-Machine Interface/Interaction (HMI) or Human-Robot Interface/Interaction (HRI), information is gathered about the state of the art in this field that uses natural language with an ability in understanding and taking action. Robots are met in ambient assisted living environments, in the role of tour guide, partners, researchers in difficult environments, etc., in which all modern techniques are applied. A gap has been observed in their ability to manage commands according to the implied time of action. In other words, there can be a number of commands without a time sequence, where a machine should be able to place these commands in the correct order of time, using every means to obtain further information, even by asking questions or observing its space.

Keywords*:* Human-Machine Interface, Human-Machine Interaction, Human-Robot Interface, Human-Robot Interaction.

1. Introduction

It was in March 1942 that science fiction author Isaac Asimov, in his short story "Runaround" (Asimov, 1942), first formulated the three laws (Asimov, 1942; Barthelmess & Furbach, 2014) that governed the behavior of robots which are:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the first law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second law.

In the decades that followed, more and more members of the scientific community began to research into robotics at the communication level. The above three laws were adopted by scientists and formed the basis of

research into the creation of robot ethics. Many years later and at a more mature moment of robotic intelligence, on July 31st, 2017, the electronic version of the Independent, entitled "Facebook's artificial intelligence robots shut down after they started talking to each other in their own language" (Griffin, 2017), describes an issue that was came to trouble public opinion at the time. It was a Facebook experiment where two bots had to communicate and negotiate with each other as two people would. The result, as recorded by those in charge, was that the robots appeared to chat to each other in a language that they each understood, but which appears mostly incomprehensible to humans and so the experiment came to an end. In these two extremes of the development of both ethics and independent communication, using speech between machines, but also in many other cases, the development of machine's intelligence is required. So a machine should be able to hear, understand, think, act and speak. With this in mind and on the basis of the communication of machines with humans, what we call the Human Machine Interface/Interaction (HMI) or Human Robot Interface/Interaction (HRI), it will be herein inquired into what has been done so far in this field of research.

2. Related works

In this paper, the work that has been done in HMI and HRI is surveyed, for the period 2015 – to present. The total results received for all this time in scientific articles and bibliographic reviews were 1128 publications. All publications cover all fields of research and science, from biology, astronomy, mathematics, engineering and more. After a first sorting, both in relation to the scientific field and in relation to the substance of the research, a total of 413 publications had been kept for further study. Out of these, 119 were rated for final evaluation, comprising essentially 101 articles and 18 literature reviews. During the study, all papers were distinguished and categorized into two major categories, which are:

- Brain Machine Interface/Interaction (BMI) .
- Human Machine Interface/Interaction (HMI).

Both groups have a relative equivalence, but with the first group having a more compact presence in the field of research, while the second is divided into other smaller ones. They were also ranked in two other groups posts about interactive robots (mainly in the industry), Collaborative Robots (CR) and Human Robot Environment Interaction (HREI). More specifically, the classification rates in all 119 publications are as follows:

- HMI: 46.2% with 55 papers
- BMI: 38.5% with 46 papers
- CR: 14.6% with 17 papers
- HREI: 0.9% with 1 papers

The HMI ones have been categorized as follows:

1. Speech Recognition (SR), Natural Language Processing (NLP), Natural Language Understanding (NLU), Natural Language Generation (NLG) and Grounding. The ranking was made according to the combination of the above parameters, so that the whole structure includes at least speech as an input and the output is at least a response.
2. Regarding Medicine, recognition of manipulations and expressions to assist patients with skeletal problems or doctors for surgery or psychologists.
3. For the recognition of a movement language with image processing or kinesthetic perception - response using special gloves.
4. Regarding the Social Robots, which, however, do not fall under the conditions of category 1.
5. On the use of techniques for smart development of various types of robots.

The 55 articles related to the above categories, and more broadly HMI, are classified such as:

- Category 1: 29 papers.
- Category 2: 10 papers.
- Category 3: 8 papers.
- Category 4: 5 papers.
- Category 5: 3 papers.

According to the above search results of category 1 that is the herein research scope, the question is answered of what is the interest of the scientific community in HRI/HMI and where research has been in recent years. So they are researching:

- What it means for humans to understand what the robots do and think (Hellström & Bensch, 2018), which is mainly about the evolution of social robots that work closely with them. In this area, and especially in the Ambient Assisted Living category, there is interest in the combination of a humanoid robot with a smart home environment (Bui & Chong, 2018; Bui *et al.*, 2017).

- Remaining on the home environment, the interaction with the robots is upgraded as giving importance to the lexical entrainment (Iio *et al*., 2015), while there is a new proposed method to avoid the inevitable occurrence of errors of incoming speech in a service robot (Tada *et al.*, 2020) and the correction of a collaboration robot's behavior in real time, using natural language commands (Broad *et al*., 2017).
- At the dialogue level, there is interest in the ability of a robot to detect errors in the words of its interlocutor (Schütte *et al*., 2017), a cloud-based dialog navigation agent, placed in a moving robot on a campus (Chen *et al*., 2019), an Augmented Robotic Dialog System, which is developed for a high analysis Natural Language Processing (Alonso-Martín *et al.*, 2015), as well as a research framework presented for handling miscommunication between people and robots in task-oriented spoken dialogue (Marge & Rudnicky, 2019).
- There is interest in the combination of movements and speech for the training of a robot (Du *et al*., 2018), for the identification of movements with conceptual terms (Almagro-Cádiza *et al.*, 2018) and for the fusion of speech and gesture (Yongda *et al*., 2018).
- There is some research efforts in Natural Language Processing on online e-commerce applications (Kandhari *et al*., 2018), in Natural Language Processing and Understanding, where the robot with dialogue ability can request clarifications (Thomason *et al.*, 2019), and a research in grammatical formalisms for developing a flexible and efficient model of Natural Language Processing (Kovács & Barabás, 2015).
- There is a proposal for a general architecture to support tools for speech recognition and synthesis, using the Wizard of Oz method (Schlögl *et al.*, 2015), a study for the usage of GUIs that are introduced to teach a machine (Igarashi & Inami, 2015), a study that presents an approach to the design and implementation of a Spoken Language Understanding (SLU) workflow (Bastianelli *et al*., 2017), and one more for pinpointing the relations of new ASR techniques to properties of the human speech communication process (Hermansky, 2019). Two of the first researches of the considered period are a telecommuting robot via voice commands that uses a Map Metric system (Poncela & Gallardo-Estrella, 2015), as well as a RVDK (Robot Voice Development Kit) for easy implementation of voice interfaces (Bastianelli *et al*., 2016).
- Another interesting work is Human-robot communication from a distance, where the robot explains at every step what it does (Garcia *et al*., 2018) and another one where the robot can describe in natural language what it sees in its environment (Cascianelli *et al*., 2018). There is an initial attempt to use a speech recognition system, performing a point-to-point control of a robot, to accomplish specific tasks (Saravanan & Sivaramakrishnan, 2019) and one more for the understanding of natural expressions by a robot, according to instructions given to it and its understanding ability if the required movements can be made or are absolutely impossible (Recupero & Spiga, 2019).
- Finally, the last researched papers are those describing, in the first case, a method of how to acquire knowledge from a robot through multiple sources that can be devices manuals or information from the web (Xie *et al*., 2015), another case is in expanding the understanding of natural language through real knowledge and environment's observation (Paul *et al*., 2017) and in the last case there are information about the construction of a method so that a robot can complete omitted information in human speech, by using information from its environment (Chen *et al*., 2019).

All of the above are part of a unified research into the development of an HRI or HMI system that will bring a robotic system closer to the human hypostasis, in order to offer cooperation and services to humanity. In this context, the development of intelligence has the first role to build skills and ethics!

3. Research on understanding

One of the key features in HRI is understanding, which includes not only the intentions but also desires, knowledge, beliefs, emotions, perceptions, capabilities, and limitations of and by a robot. The related herein studied papers appear in the table that follows (Table 1), sorted by the year of publication; they will be presented and discussed in the two following subsections, respectively.

3.1 Presentation

There is the study No. 1 of Alonso-Martín *et al.* (2015) on the creation of an Augmented Robot Dialogue System (ARDS), for more realistic human interaction. The robotic research platform has been the social robot Maggie. Research is based on the development of systems capable of achieving a high level of understanding

of natural language (NLU). According to the author and through his bibliographic references, in order to understand natural languages, it is important to distinguish between seven interdependent levels which are Symbolic, Morphological, Lexical, Syntactic, Semantic, Discourse and Pragmatic. One of the main advantages of ARDS is the ability to communicate with the robot, with or without a grammar, which uses natural language. Because grammar significantly limits the interpretable input, the use of a grammarless ASR (Automatic Speech Recognition) was chosen. The NLU Unit is complemented by an Optical Character Recognition (OCR), an Information Extraction (IEx) module and an Information Enrichment (IEn) module. This research aims to enable the robot to gather information from any medium (e.g., from the network) about all the parts of speech that can be made; for example, regarding a place, a famous person or a process. It must also understand the timing of the dialogue. These parameters make the robot capable of expanding a discussion more specifically on this subject. All robotic software has been implemented in Maggie, through the Robotic Operative System (ROS) (URL 30), while the Google ASR web service has been used to convert speech into text.

Table 1: Classification of publications based on publication year

No.	Reference	No.	Reference
1	Alonso-Martín *et al.*, 2015	16	Cascianelli *et al*., 2018
2	Igarashi & Inami, 2015	17	Du *et al*., 2018
3	Iio *et al*., 2015	18	Garcia *et al*., 2018
4	Kovács & Barabás, 2015	19	Hellström & Bensch, 2018
5	Poncela & Gallardo-Estrella, 2015	20	Kandhari *et al*., 2018
6	Schlögl *et al.*, 2015	21	Yongda *et al*., 2018
7	Xie *et al*., 2015	22	Chen *et al.*, 2019a
8	Bastianelli *et al*., 2016	23	Chen *et al.*, 2019b
9	Bastianelli *et al*., 2017	24	Hermansky, 2019
10	Broad *et al*., 2017	25	Marge & Rudnicky, 2019
11	Bui *et al.*, 2017	26	Recupero & Spiga, 2019
12	Paul *et al*., 2017	27	Saravanan & Sivaramakrishnan, 2019
13	Schütte *et al*., 2017	28	Thomason *et al.*, 2019
14	Almagro-Cádiza *et al.*, 2018	29	Tada *et al.*, 2020
15	Bui & Chong, 2018		

The study No. 2 of Igarashi & Inami (2015) begins with the fact that, at the time of this particular research, the human-robot interaction (HRI) approaches usually fall into two categories. The first category is a practical approach, where the user provides simple abstract instructions to the robot, with voice or gesture commands, and the other is a direct operation approach, where the user sends detailed control commands to the robot, using a joystick or control pad. As part of the work on Japan's ERATO Igarashi Design Interface Project (URL 66), alternative approaches were explored that fall between these two extremes, leveraging the knowledge and methodologies developed in the human-computer interaction (HCI) field. So, according to the thinking that GUIs, because they use a display and mouse, permit the user to interact with graphical representations of the problem, thus facilitating efficient solutions, the exploration to teach mobile robots to perform physical tasks was initiated, by implementing GUIs. Another HCI feature used is Augmented Reality (AR), whose purpose is for the user to manipulate the target objects in this environment. One more feature is the Tangible User Interfaces (TUI) that employ graspable physical objects as a means for interacting with virtual information. Here, they are used as a means to provide in-situ instructions to robotic systems that perform physical tasks in the real world. The Last HCI feature is the I/O developing. This is a trend that is motivated by the need for interaction with the real world, having the result that they develop enabling technologies, such as Arduino, that allow the rapid development of such novel hardware devices. The authors conclude that it is necessary to establish correspondences between real-world entities and information in the robotic system.

In the base that the using consistent linguistic schemes, during a conversation with another person, decreases the cost of resolving the ambiguities of each utterance, that might have more than one interpretation, the lexical entrainment is an important element for successful communication. Lexical entrainment or alignment is called the phenomenon that when a person is talking with an addressee tends to repeat the same verbal expressions as him/her. The paper No. 3 of Iio *et al*. (2015) is focused on lexical entrainment in human robot interaction. More specifically, a situation is examined where a speaker and her addressee refer to objects

around them in a physical space. They share the same space and can see the behaviors of each other. This situation is essential for social robots, which interact with people in real environment, including guide, or porter robots. For the experiments, a Robovie R ver.2 (URL 13) was used that is a humanoid robot, which has a human like upper body, designed for communication with humans. Due to its three degrees of freedom (DoF) for its neck and the four DoF for each arm, the robot can perform such human-like gestures. For precisely recognizing speech or pointing gestures in real time, the Wizard of Oz method (Bella & Hanington, 2012) was used.

On the fact that the grammar parsing and transformation of the incoming sentence into a corresponding semantic model are the kernel modules in natural language interface engines, in the paper No. 4 of Kovács & Barabás (2015), a grammar formalism is proposed, which is a combination of regular grammar and dependency grammar, extended with predicate oriented annotations, to fulfill this purpose. The goal here is to implement a Natural Language Interface (NLI) engine for the Hungarian language, and to develop a flexible and efficient language processing model. For ASR module construction, reference is made to Nuance, which has the products: Nuance Recognizer (URL 58) and Dragon (Sarnataro, 2012). The Nuance Recognizer is used to convert speech into written text, but it needs a well-defined grammar XML files that should be built, which describes the word sequences that the robot is able to recognize. The robot that was used is the applied humanoid Nao (URL 46) with the API NaoQi application (URL 59) that is embedded in it.

In the work No. 5 of Poncela & Gallardo-Estrella (2015), a tele-operated robot is presented, controlled through voice commands. It consists of an HRI and a Map Metric system that creates a geometric map of the environment, in which the robot moves. HRI is structured with three subsystems, being State, which collects information from the path and location to build the geometric model, Speech Recognition, and Command Generation. For the Speech Recognition System (SRS), because the work required open-source software (which was able to manage Spanish language), Julius/Julian (URL 18) was selected that is an open-source large vocabulary continuous speech recognition (LVCSR) engine. This SRS includes a dictionary in Spanish with a grammar and an acoustic model that contains a statistical representation of phonemes, which make up each of the words of the grammar and is implemented with the HTK (Hidden Markov Model) Toolkit (URL 71). To build the acoustic model, a set of training sentences is needed to obtain the model after a training process. Each user has to train his/her own acoustic model with his/her data. Training sentences have to be recorded with a microphone and next, recorded data are parameterized to convert the raw speech waveforms into a sequence of feature vectors. Linear prediction coefficients (LPC) and Mel frequency cepstral coefficients (MFCC) are the most widely used in speech recognition, because with this method, the characteristics of the vocal tract are taken into account. Once speech data have been parameterized, the training stage starts. The system has been tested with a Pioneer P2AT (URL 39) on Mobile Sim. The speech interface has been successfully tested with different users. It has been evaluated with the WRR (word recognition rate) and a new proposed metric, the CSR (command success rate), more suitable for command-base systems. They have yielded very high recognition rates, better than 98.8% for the WRR and better than 95% for the CSR.

The paper No. 6 of Schlögl *et al.* (2015) presents a systematic investigation and analysis of the design space for the Wizard of Oz (WOZ), for language technology applications. WOZ is considered a well-established method for simulating the functionality and user experience of future systems. WOZ is used to simulate ASR, MT (machine translation), NLU and natural language generation (NLG) as well as TTS. In this paper, a generic architecture is proposed for tool support that supports the integration of components for speech recognition and synthesis, as well as for MT. Two categories of applications and frameworks may support WOZ exploration. In the first category, these tools belong, whose primary application lies in the area of NLP and machine learning, and comprise the dialogue management (DM) tools. The second category relies completely on human simulation, and a reference is made in it, as pure WOZ tools. The authors give as the most referred DM tools, the CSLU toolkit (URL 60) and the Olympus dialogue framework (URL 32). These units include all the required modules like ASR, TTS, NLU etc. The CSLU toolkit can be seen as a straightforward prototyping tool that requires little experience, which makes it suitable for both designers and NLP researchers. In contrast, the Olympus dialogue framework, which includes the Rosetta language generator (URL 36) and the Kalliope speech synthesizer (URL 37), constitutes a powerful client–server environment for implementing and running spoken dialogue systems; so it requires a high level of technical know-how to set up and use. The Jaspis (URL 61) DM system and the EPFL (URL 62) dialogue platform are also referred to. The Jaspis dialogue manager provides high adaptability, while the EPFL dialogue platform is based on the Rapid Dialogue Prototyping Methodology that makes it interesting for researchers, who have little technical knowledge. Pure WOZ tools are, however, scarce and tend to be only suitable for the one experiment, which they were constructed for. Two publicly available tools that allow for more generic experimentation include SUEDE (URL 63) and Richard Breuer's WOZ tool (URL 64). An alternative, yet not publicly available, solution may be found in the NEIMO (URL 65) platform. SUEDE allows a designer to rapidly create prompts,

and supports the graphical design of a dialogue flow. Following a similar goal, Richard Breuer's WOZ tool puts a stronger focus on more complex dialogue designs. The NEIMO platform was mainly used in the 1990s to study the potential of multi-modal user interfaces. In this paper, a set of tools are also mentioned that have been developed for various other tasks.

The paper No. 7 of Xie *et al*. (2015) presents a proposal on the effort to reach a solution to the notorious problem of knowledge acquisition. In the context of this proposal, a major challenge that arises is the translation of the open knowledge, organized in multiple modes, unstructured or semi-structured, into the internal representations of agents. In this direction, it presents a set of multi-mode NLP techniques, for formalizing multiple modes of open knowledge into the internal representations of agents. According to the authors, these techniques, implemented in the robot OK-KeJia (URL 80), have successfully enabled it to acquire knowledge (to accomplish user tasks) from open knowledge sources, like a device manual or the OMICS (Open Mind Indoor Common Sense) database (URL 81). In addition, more knowledge are required, especially from users (including their desires), because human users can provide endless amount of knowledge and help for the system. So, the human-system collaboration is a promising approach to the development of autonomous agents, where knowledge can be obtained through dialogue. All knowledge, extracted from open knowledge resources, is transformed into an internal representation. For this task, the authors use a logic programming language that is called ASP (Answer Set Programming), as the underlying language, with a Prolog-like syntax. The extracted knowledge is transformed into ASP-rules. The goal of the multi-mode NLP is to formalize knowledge of multiple modes into the internal representation, i.e., the ASP-rules. The tool used as the underlying formalism, for semantic parsing over English sentences/phrases, is the CCG (URL 52). This is a linguistic formalism that provides a tight interface between natural language syntax and it's semantics. A semantic lexicon constitutes the core of any CCG and for each of them, each lexeme consists of a word, a syntactic category, and a semantic form (represented in λ-calculus expression). Additionally, the attempt for an efficient calculation of a parse is applied by a CKY algorithm with beam search.

In paper No. 8, the aim of Bastianelli *et al*. (2016) was to design a methodology for the easy implementation of human-robot voice interfaces and to implement an industrial tool, called the Robot Voice Development Kit (RVDK), for the construction of such resources. The overall architecture of the RVDK system consists of two subsystems. The first one implements the Spoken Language Understanding (SLU) process and includes ASR, and the second one, which is installed on the robotic platform, includes the Grounding unit that extracts the actual commands from the recognized and analyzed sentence, converting them into robot's action. As an ASR, the Loquendo ASR and Microsoft Speech Platform were used. The semantic frameworks used were FrameNet and RoboFrameNet (URL 72). A Dialog Manager (DM), that keeps track of the state of the interaction and implements the corresponding actions, is modeled as a finite-state machine (FSM) that decides whether to send the command to be executed to the platform or to extend the interaction with the user, to clarify the command. As part of Grounding user commands and for building semantic maps, the Human-Augmented Mapping (HAM) way had been followed. The paper concludes that many aspects of SLU for HRI require further investigation. For example, robots must have dialogic capabilities to fully exchange information with the user. The request for clarification and confirmation is a key point in achieving a satisfactory level of interaction. Another important feature is the ability of robots to start a conversation, when necessary. Using a simple DM, a minimal approach to dialogue can be achieved. More complex systems and examples should be considered, such as the involvement of more contexts, arising from the interaction and the field of speech, or the use of a more complex and dynamic computational model. The Grounding problem is another key aspect of SLU. Once semantic structures are extracted from a sentence, a mapping must be performed between the concepts expressed using symbols (e.g. words) and representations of the real world, as provided by the robot's perception system. This mapping is usually called Grounding or Anchoring, when the process expands over time. The interpretation of the meaning of an expression that contains references to the real world cannot be completed, until all these references are resolved in the right context and thus an approach is proposed that exploits knowledge bases that contain semantic information about the environment, called semantic maps. Obviously, a mechanism is needed to build or acquire these semantic maps in a flexible way. One possible approach is the usage of HAM, where the semantic map is built interactively with the help of a human, who guides the robot within the environment, points to the objects using a laser pointer, and narrates them accordingly.

In paper No. 9 of Bastianelli *et al*. (2017), an approach is presented to the design and implementation of a spoken language understanding (SLU) workflow. A little more specific in the field of interest herein, a statistical semantic parser has been built that works: (i) on Speech and morpho-syntactic analysis (Speech re-ranking); (ii) on Semantic parsing; (iii) and Grounding. In a few words, the re-ranking module aims to improve the recognition of the free-form ASR engine, by applying post-processing on the n-best candidate transcriptions. Here we meet the Support Vector Machine (SVMA), which is the learning algorithm that is used

to acquire the classification function, the kernels which compute the similarity between two sentences that used syntactic kernels, such as TKs and the lexical kernel Klex, and the Smoothed Partial Tree kernels (SPTKs) (URL 67). The SPTK estimates a high similarity with sentences characterized by a strong syntactic analogy and, at the same time, a strong lexical relatedness. The extraction of linguistic information for individual candidates is carried out through the Stanford CoreNLP suite2 (URL 68), which is one of the largely used off-the-shelf morpho-syntactic parsers and includes tokenization, morphological processing and grammatical analysis. About the Semantic parsing, it relies on frame semantics and is divided in three stages. First, The Frame Semantics is not only a sound linguistic theory, but it also allows the adoption of large scale annotated resources, that extend (or integrate) the background knowledge of an interactive robot. The FrameNet annotated corpus (URL 69) represents a collection of more than 150,000 annotated sentences. This belongs to the Semantic role labeling for command semantic parsing. Secondly, on the action detection, in order to disambiguate all possible semantic frames that represent predicates, a classification function follows a multi-classification scheme, based on the ulticla SVM mss. Each instance is modeled as a set of manually engineered features that reflects different linguistic observations, which are classified as Lexical features and Shallow syntactic features. The last stage is the Full command recognition. It is used as a classification scheme, based on a Markovian formulation of a structured SVMs (SVMhmm) (URL 70). Given an input sequence, the SVMhmm algorithm learns a model, isomorphic to a k-order hidden Markov model. The authors claim that this modeling provides the best possible labeling of a sentence. The Viterbi algorithm computes the output labeling. In the training phase, the SVMhmm solves the optimization problem described. In the discriminative perspective of SVMhmm, each word is represented by a feature vector, describing its different observable properties, which are the Position (its relative distance from the target predicate), the Lexical features and the Distributional features. The last step of the processing chain takes as input a fully instantiated semantic frame and turns it into a robot action. This process is usually referred to as grounding. Although the proposed solution allows for the development of SLU chains that do not require specific formalisms for expressing the robot's plans, in the experimental evaluation, the robot actions are modeled using Petri-net plans (PNPs). A PNP is basically composed of the Places, the Transitions and the Edges. In this work, the perceptual representation of the world, provided by the robot, is represented through a semantic map that contains, in addition to spatial information about the environment, assignments of mapped features to entities of known classes. In particular, a semantic map may include representations of objects and locations, their categories, their positions with respect to the metric map used for navigation, the perception that the robot has acquired, and possibly other features. In this sense, the semantic map is a resource that enables grounding, since it represents the environment acquired through perception.

In paper No. 10 of Broad *et al*. (2017), a proposed natural language interface can be studied that allows a human user to correct the behavior of a collaborative robot system at run-time. This work is motivated by the desire to improve the collaboration between humans and robots, in a home environment. This interface is based on the Distributed Correspondence Graph (DCG), a probabilistic graphical model that assigns semantic meaning to user utterances, in the context of the robot's environment and current behavior. The problem of how to best map linguistic elements to their corresponding symbolic representations in the real world, which is known as the Symbol Grounding Problem, is proposed to be solved by limiting the size and range of the language that the robot must understand in the correction space. A MICO robotic arm was employed from Kinova Robotics (URL 19), with six DoF. During development, SR was incorporated through using a javascript library (URL 20), which was connected to the rest of the system through roslibjs (URL 21). The implementation of a language model is based on an adaptation of the open-source Human-to-Structured Language (H2SL) software package (URL 22), which allows converting free-form text input to its corresponding structured language grounding in the robot's environment. Symbolic representation for natural language corrections is sufficiently small for real-time performance with the DCG; however, the Adaptive Distributed Correspondence Graph (ADCG) or Hierarchical Distributed Correspondence Graph (HDCG) could also be employed, for scenarios in which the symbolic representation becomes sufficiently complex. Training data for the language model was gathered using Amazon Mechanical Turk (AMT) (URL 23), an online crowd-sourcing platform. Finally, the authors claim that this method improves the user's experience, while simultaneously improving the communication and execution of a user's desired trajectory, when compared to high-level goal driven approaches.

The research No. 11 of Bui *et al*. (2017) uses the same robot and the same smart environment for the same purpose. Between several software frameworks and middleware (MW) (URL 05) for AAL that are based on advanced information and communication technologies (ICT), universAAL (uAAL) (URL 06) seems to be the most holistic platform. It integrates KNX (URL 07), Zigbee (URL 08), and Z-wave (Kaven, 2005) partially into its MW. This ECHONET-based smart home environment can provide AAL services by designing and

implementing the integration of uAAL and ECHONET (URL 09) standards. uAAL allows the seamless integration of heterogeneous devices within a network environment, through two basic concepts:

- the usage of three communication buses for topic-based communication among components and
- the usage of ontologies for information and services, shared between these components.

The above referred buses are a Context Bus, a Service Bus, and a User Interface Bus, being the heart of the uAAL MW that is the core component of the uAAL platform. All devices that run the MW are nodes which can share knowledge and functionalities with other nodes in the form of ontology. The Api.ai (URL 10) uses a natural language understanding platform and also uses the NAOqi framework (URL 03) for speech-to-text transformation. The integration between the uAAL and ECHONET standards has been done by embedding the uAAL platform into the HomeGateWay (HGW) (URL 11) that includes the ECHONET Interface, the uAAL Adaptation and uAAL MW. To cover all attributes of ECHONET standard, the SAREF (URL 12) ontologies were used as alternative vs. the uAAL ontologies. By the authors, we know that the robot could communicate with a user in an intuitive way, by employing a free natural language API. Moreover, they confirm that the proposed speech-activated interface enables the robot to gain access to the iHouse network and provide user-requested data and services accordingly.

In the work No. 12, Paul *et al*. (2017) propose a model, called Temporal Grounding Graphs, which significantly expands the range of language that a robot can understand, by incorporating factual knowledge and observations of its workspace in its inference about the meaning and grounding of natural-language utterances. In fact, is a probabilistic model that enables incremental grounding of natural language utterances, using learned knowledge from accrued visual observations and language utterances. The model allows efficient inference over the constraints for future actions, in the context of a rich space of perceptual concepts, such as object class, shape, color, relative position, mutual actions, etc., as well as factual concepts ("does an object belong to an agent such as a human?") that grow over time. The system was deployed on the Baxter Research Robot (URL 82), operating on a tabletop workspace. Regarding the Spoken commands from the human operator, they were converted to text using an Amazon Echo Dot (URL 83). The parsing model is realized by using a dependency parser, named START (SynTactic Analysis using Reversible Transformation) (URL 84), that transforms the input language into a set of constituents, related to each other, using syntactic dependency relationships. It is only mentioned that the method identifies a factor, which models the likelihood of a factual grounding and the joint likelihood over perceptual groundings and correspondences. This factor is realized using a vision-language framework, the Sentence Tracker (URL 85), which models the correspondences between visual observations and an event, described by a sentence as a factorial hidden Markov model.

In the context (No. 13) of Schütte *et al.* (2017), an experiment was performed, in which, problems were artificially induced in a dialogue based human-robot interaction. The concept was to investigate the problem of errors' perception in human-robot dialogue. According to the experiment's conditions, there were 20 scenes that in 14 of them perception errors were designed. There were three types of errors: the missing object error, the wrong colour error and the wrong type error. The experiment, that was called "The Toy Block", consisted of five phases: "The No Error phase", "The Error phase", "The Description phase", "The Markup phase" and "The Querying phase". The results showed that if errors are present in the perception of the robot, this leads to an increase of problems in the interactions, and makes the successful completion of the tasks more difficult. Giving the human dialogue partners access to information about the robot's understanding of the scene allows them to align with it, such as to avoid and resolve problems and complete these tasks more efficiently. The Toy Block system enables users to interact through a dialogue system with a robot, similar to the SHRDLU system (Winograd, 1971). The robot itself is abstract, and not physically represented in the simulation. When the user sends an instruction, the robot analyses the instruction and attempts to perform it, in the simulation world. In another case, it replies through the user interface and explains its problem. The basic natural language processing (NLP) pipeline of the system involves the NLTK parser (URL 24), an analyzer for the resulting parse similar to the Spatial Description Clause structures, the grounding of the referring expressions and a unit that if the grounding succeeds then it proceeds to the action's executing, else it generates an appropriate response.

The paper No. 14 of Almagro-Cádiza *et al.* (2018) focuses on a specific type of co-verbal gesture, related to the content of speech, also known as iconic gestures. According to existing surveys so far, the fact that a gesture is initiated only when some defined word is detected does not seem to simulate natural behavior. The main idea that is proposed in a preliminary study is to define the meaning of body expressions through relevant terms, giving the robot the ability to execute a motion by finding any word semantically related to those terms. The purpose of this paper is to extend this study by introducing and evaluating language models, generated from lexical learning, based on distributed semantics with them that are based on semantic schemes, prepared by expert linguists. Because of a more in-depth examination of results, which shows that both types of models can fit together, the semantic analysis component that is included in the proposed gestural interaction

architecture is determined with this combination. The components of the proposed architecture have been developed on a Nao robot (URL 46). FreeLing (URL 47) has been used as a morphosyntactic analyzer, isolating words and categorizing them, and semantic comparisons have been applied, using the models: Word2Vec model (Mikolov *et al*., 2013) and MultiWordNet (URL 48). For developing a list of sixty gestures, concepts have been used that have been identified as the most frequent words in language, for each grammatical category, from the Corpus of Contemporary American English (COCA) (URL 49).

In a proposal (No. 15) to create a system for Ambient Assisted Living (AAL), the humanoid Pepper (URL 01) is combined with a smart home environment (Bui & Chong, 2018). Pepper receives a command and then acts on the home systems by applying whatever was asked of it and finally confirms this action with speech. The exchange of proposals is achieved with DialogFlow (URL 02), a natural language editing and comprehension platform by Google. The humanoid itself has a programming framework, called the NAOqi framework (URL 03), which undertakes to convert speech into text and vice versa through two APIs (Application Programming Interface). The procedure is as follows: The incoming speech first goes to the ALTextToSpeech (URL 04), API of the NAOqi framework and becomes text. It is then transformed into Structured Data structures in DialogFlow's NLP Library. Then, keywords are identified so that the machine understands what is required, and acts accordingly. Finally, a text is created with the act it did and announces it through the ALTextToSpeech API of the NAOqi framework. A major problem is that in noisy environments the system cannot clear incoming commands.

In the case of a robot, called to interact with a human by analyzing visual stimuli, the question arises of describing the environment. Can the robot obtain the ability to provide a description of the scene, in a form that every user can easily understand? The answer is a keystone for the success of effective and user-friendly service robotic products. In fact, a natural language description offers an interpretable manifestation of the robot's inner representation of the scene and is also a good basis for natural language question, answering about what is happening in the environment. Hence, this functionality would provide a friendly interface also for non-expert people, who would then be able to easily interact with their home robot, in the near future. The problem that the robot can describe in natural language, of what it sees in its space, is formalized as a Machine Translation (MT), from "visual language" to natural language (English, in the current study). In the project No. 16 of Cascianelli *et al*. (2018), the robot side of the interface has been investigated, in particular, the ability to generate natural language descriptions for the scene it observes. The describing of a scene in natural language, is usually referred to as Natural Language Video Description (NLVD). In recent years, many researchers, from both Computer Vision and Natural Language Processing communities, are studying the problem of describing generic videos, by using natural language phrases. Some popular approaches are based on filling in predefined template sentences, with the subject-verb-object concepts detected in the video. In this study, a full-GRU (Generalized Recurrent Unit) encoder-decoder architecture is presented to address it. The authors argue that τhe proposed approach was faster to train and less memory consuming than other State-of-the-Art algorithms; in addition, this method is also competitive in terms of performance on the public datasets, which were partially used also for training.

In paper No. 17 of Du *et al*. (2018), an online robot teaching system is presented. As it is known, natural human-computer interaction is an important interface to realize friendly collaboration of intelligent robots and humans. Most of the communication between humans is conducted through speech and gesture, and the interaction of this is natural and intuitive. Here, a teaching method is presented, based on this relationship. In fact, the human trainer gives instructions (to the robotic system), using speech and gestures. The right sensors receive both sources, which feeds an Audio-Visual Fusion, based on Text (AVFT) unit. This unit consists of a proposed method that has the advantage of converting sources' information into texts. A GOOGOL GRB3016 robot (URL 41) and a KUKA KR 6 R700 sixx robot (URL 42) were used. A Kinect depth camera (URL 43) and an inertial measurement unit (IMU) (URL 44) were also used to capture the speech and gesture of the humans. The Microsoft speech SDK (URL 45) was used in the system to collect the speech of the operator. AVFT was proposed to fuse the speech and the hand gesture. The maximum entropy algorithm was employed to transform the text into the instructions. According to the authors, the feasibility of manipulation for high-precision tasks was validated in the experiment. The proposed interface could be extended to the actual industrial scene. By using gesture and speech, operators can control the robot without complex operations. Besides, the system can teach two collaborative manipulators, using two hands easily.

The work No. 18 of Garcia *et al*. (2018) aims at a robot human interface, where the robot is away from its operator and will have to explain what it is doing, thus creating a relationship of trust between the two of them. More specifically, autonomous systems (AxV-Advanced eXperimental Vehicle) that usually operate in areas that are dangerous or impossible for humans to reach (such as a submarine), should be able to maintain a constant communication as to what they are doing and increase capturing the situation they are in, by explaining their actions and behaviors. Explanations can help to formulate clear and accurate mental models of

autonomous systems and robots. Mental models, in cognitive theory, provide an idea of how people think, either functionally (understanding what the robot does) or structurally (understanding how it works). Mental models are important, as they strongly influence whether and how robots and systems are used. In this presented work, the MIRIAM robot (Multimodal Intelligent interaction for Autonomous systems) allows 'on-demand' queries for status and explanations of behavior. MIRIAM is connected with the Neptune autonomy software (URL 73) and runs alongside their SeeTrack interface (URL 74). Here the 'speak-aloud' method is adopted, where an expert provides rationalization of AxV behaviors, while watching the videos of the missions in SeeTrack software. This is how an interpretive model of autonomy emerged. If a justification is requested, then the decision tree checks in the current state and the history of the mission, and then the possible reasons are determined together with a probability. There can be many reasons, with different levels of certainty, depending on the information available at a given point in the mission. Therefore, when the same question is asked at another point later, only the highest confidence answer is given. The surface representations of the explanations are generated using template-based NLG. The wording of the output reflects the certainty on three levels: above 80% (high), 80% to 40% (medium) and below 40% (low).

In paper No. 19 of Hellström & Bensch (2018), it is investigated what it means for humans to understand how the robots act and how to be designed to support this understanding and also what this includes. In particular, social robots that work closely with humans should be designed so that the humans can understand how the robots think and act, because otherwise this may negatively affect efficiency and safety. Under this condition, a novel model of interaction for understanding is proposed, applicable to cases in which both human and robot utilize a first or higher-order Theory of Mind (ToM) (Gweon & Saxe, 2013), also known as mindreading or mentalizing, to understand each other or also to simpler cases in which the robot performs static communicative actions, in order to support the human's understanding of the robot. This model builds on and extends Shannon's model of general communication (Ash, 1966). The term State-of-Mind (SoM) was used to refer to all relevant information in a robot's cognitive system, for example actions, intentions, desires, knowledge, beliefs, emotions, perceptions, capabilities and limitations of the robot, task uncertainty, and task progress. The term communicative actions were introduced to refer to robot actions, aiming at increasing the interacting human's knowledge of the robot's SoM, i.e., to increase the human's understanding of the robot. A specific valuable insight provided by the model is the conceptual separation of information to be communicated, from the means to communicate, i.e., the communicative action.

In a study (No. 20) of online voice-controlled e-commerce applications (Kandhari *et al*., 2018), a group of SRS (Speech Recognition System) teams are presented, who undertake (NLP) techniques and methods of machine learning for the conversion of speech into text. Initially, SRSs are categorized according to the following criteria:

(a) Type of speech;
(b) Dictionary;
(c) Style of Speech;
(d) Speaker Model.

Two SRS teams in the global community, those of "Cloud-based Proprietary" and "Open source", were compared by the authors and ended up as following:

- Cloud-based Proprietary: (i) Google, (ii) Nuance, (iii) IBM, (iv) AT&T, (v) WIT.
- Open source: (i) Kaldi, (ii) CMU Sphinx, (iii) HTK (Hidden Markov Model Toolkit).

The quality of the results was determined by the phrase recognition rate (PRR) and the word error rate (WER). In terms of comparison between the two teams, cloud-based performance is better than Open source, in terms of WER and PRR. Greater accuracy ensures that users don't have to give the same voice commands repeatedly to the app. Open source tools can also support high-precision speeds, but a significant amount of time is required to train and customize the system for a particular application.

In the work No. 21 of Yongda *et al*. (2018), after the advantages and disadvantages of previous researches have being comprehensively considered, complementing each other, an improved human-robot interaction is proposed that explores the possibility of multimodal human-robot interaction, based on fusion of speech and gesture. The authors argue that the advantage of the proposed method is that the combination of speech and gesture makes the human-robot interaction more convenient and direct. In a series of experiments that were carried out to validate this method, they argue that they proved that this is better than the other proposed methods. The experimental platform consists of the GOOGOL GRB3016 robot. Speech and gesture commands are sent for controlling the robot. These commands are passed to the human-computer interaction system, through which the natural language will be transferred to the corresponding execution instructions. The tools that are used to measure the position, velocity and acceleration of hands are the Leap Motion method (URL 50). On the procedure to reduce the noise of the sensor, which increases over time, an Interval Kalman Filter (IKF) is used (Zarchan & Musoff, 2000). IKF can deal with this situation via statistical parameters and

inaccurate dynamics, compared to a standard Kalman Filter (UKF). To collect the speech of the operator, the Microsoft speech SDK (URL 45) is used. Then, a corpus-based algorithm of maximum entropy classification for NLU (URL 51) is employed to generate commands. For text analysis, it is important to extract text information automatically. Here, the TF-IDF (Term Frequency-Inverse Document Frequency) (Rajaraman & Ullman, 2011) has been selected to complete this task. Finally, it is stated that some defects exist for this method, if it is to be used in a real environment. Except that the experiments were simple and give a poor analysis, compared to more complex instructions, there is not an assessment about the complexity of this method. In a real environment, robots do face more complex tasks and instructions, which inevitably require a lower complexity of the entire algorithm.

In the study No. 22 of Chen *et al*. (2019a), a cloud-based dialog navigation agent (CDNA) system was designed for a campus navigation robot (CNR) that provides navigation services to students, school visitors, and people with impaired vision. Dialog interaction is a comfortable approach to human–computer interactions. According the researches' results, a dialog system with a large amount of unclassified training data has some uncoordinated dialogs in practical applications. An unrestricted dialog system is the source of these uncoordinated dialogs, and the scope involved is extremely broad. The meaning of the same sentence in different situations will not be the same, thereby generating different responses. Therefore, limiting dialog systems to situations is crucial. In this study, the CellS-based CDNA software was developed and semantic analysis and dialog were used to eliminate this problem. The CellS is characterized by high performance and high scalability; it can autonomously perceive an external environment to actively change the system behavior. Each Cell is a Java class. CellS programming is based on the structure programming and object-oriented programming, and focuses on the design of the classes Cell and Plan. Each Cell is an independent individual, depending on the ligand message to decide whether to be triggered or not. When rapidly sensing a large amount of environmental information, all Cells may be triggered simultaneously to speed up the processing of external messages. The CDNA uses an ASR and a GPS. To extract also the notion of words, it uses a part-of-speech tagger (POST) (ConceptNet) (URL 25) and question–answer services through a web service (E-HowNet) (URL 26). The CellS is a framework, based on a belief–desire–intention (BDI) and has a flow parallel mechanism. A BDI software model based on human practical reasoning is a software model developed for programming intelligent agents. The model provides a mechanism for distinguishing flow selection from flow execution. The autonomy of BDI is suitable for the development of perceptual software systems, such as the Internet of Things (IoT) and robot systems. In addition, the Stanford Parser (URL 27), an Android platform, and the Jersey RESTful web service (URL 28; URL 29) are also used. After acquiring the user sentence through ASR, the dialog engine divides the sentence into words and determines the POST and notion for each word. The dialog engine, which analyzes the semantics and determines the intention of users, consists of five stages: CSR (class-switch recombination) stage, Preprocessing stage, NLP stage, Inference stage and Action stage. To achieve the user goal, the navigation engine autonomously leads the robot to the destination. The navigation model continually and autonomously updates GPS information and checks whether the robot has reached its destination. If the robot has not reached the destination, the navigation engine transfers the relevant information through the ligand and sends a TTS (Text-to-Speech) response to the user. In the navigation engine, the location in the sentence is used to explain the part that eliminates ambiguity.

Because natural language is inherently unstructured and often reliant on environmental context, in paper No. 23 of Chen *et al*. (2019b), a Language-Model-based Commonsense Reasoning (LMCR) is introduced, being a method which enables a robot to listen to a natural language instruction from a human and automatically fill in missing information, using environmental context around it and a new commonsense reasoning approach. This happens because people are more likely to omit information from an instruction, if it is obvious to the listener. Using this premise, LMCR fills in missing arguments in the parsed verb frame, with objects detected in the environment. A robot with LMCR takes as input a spoken instruction and a raw RGB-D image of the environment. The robot first parses the input instruction and detects objects in the environment. Then, it resolves incompleteness in the instructions and outputs a complete verb frame. In the end, it computes a sequence of movements, to execute the task specified in the complete verb frame. To achieve the above, the LMCR includes an SR module, an object detector module, a Predicate-Argument Parsing module and a Motion Planning module. The LMCR was deployed on a Baxter robot (URL 82), a research robotic platform with two arms, each with seven degrees of freedom. The several tools that were used are the Google Cloud API (URL 15) as SR module, an object detector such as Mask R-CNN (URL 86), a Recurrent Neural Network (RNN), which was used by the language model and the motion planning toolkit MoveIt (URL 87). The method was compared against other methods, like Co-occur (Bordag, 2008), Word2Vec (Mikolov et al., 2013), and ConceptNet (URL 25). Results show that the neural network language model is able to learn domain-specific knowledge, by using raw textual data from the web, and further that the learned knowledge matches well with human common sense.

Speech signals, aside from the message, also contain information about other particular idiosyncrasies of the speaker, as well as information about noise introduced during speech production, perception and transmission. Speech evolved for human communication in realistic noisy environments. Since message decoding is done using human perception, it makes sense that this redundant coding of the message in the signal evolved so that relevant spectral and temporal properties of human hearing can extract this message from the noisy speech signal. The focus in the work No. 24 of Hermansky (2019) is on a particular information rate increase, due to redundancies introduced in speech production, in order to protect the message during its transmission through a realistic noisy acoustic environment. Therefore, through this research, an alternative view of human speech communication is provided, in which the vocal tract generates surplus information. This alternative view of human communication is reflected in new ASR techniques, often without a conscious effort to mimic the properties of hearing, but as a result of system optimization in speech data. One of the purposes of this paper is to identify the relationships of new ASR techniques with the properties of human speech, during the communication process.

In the article No. 25 of Marge & Rudnicky (2019), a research framework was presented for handling miscommunication between people and robots, under situated dialogue. During the procedure, the infrastructure and representation was described for detecting and recovering grounding problems that consist of incompatible factors for robots environment. A conversational interface with robots, the TeamTalk (URL 31), implements the core capabilities that robots must have, in order to handle situated grounding problems. An extension of TeamTalk, The TeamTalk Situated Reasoner (TSR), provides robot skills, useful for engaging in physically situated dialogue, and detect miscommunication events in situated human–robot dialogue. Finally, a method is present to improve recovery strategy selection, using nearest-neighbor learning (Altman, 1992). TeamTalk is a platform for spoken-dialogue interaction in human–robot teams, situated in real world and virtual spaces. It provides a spoken-language interface between one human operator and a team of one or more robots. Commands may be spoken or typed into the GUI. Robots respond with synthesized speech. The human interface component is built, using the Olympus Spoken Dialogue Framework (URL 32). Among the components are the RavenClaw dialogue manage, the PocketSphinx SR engine (URL 16), the Phoenix parser (URL 34), the annotated speech and parsed input for confidence Helios (URL 35), the generated language Rosetta (URL 36), and the produced synthesized speech Kalliope (URL 37). TeamTalk supports spoken dialogue with virtual robots, using USARSim (URL 38). A four-wheeled ground robot, the Pioneer P2AT (URL 39), in a virtual environment, has been used in this study. About TSR, it connects spoken language with robot mapping and path planner information. TSR receives language context as user input and processes it for relevance in the current situation. Features from the language context include information about recognized speech, semantics obtained from the language, and a confidence score. TSR stores static context in a knowledge base, using a structured ontology, retrievable by queries. They use an instance of the OWL representation of the Protégé ontology (URL 40), developed for TeamTalk. The authors claim that the results showed that a robot could learn how to select human-inspired recovery strategies, using a numerical representation of situations. They believe that these findings are relevant to the design of human–robot interfaces that use experience to improve performance.

The project No. 26 of Recupero & Spiga (2019) focuses on the ability of an interactive and programmable humanoid robot to understand natural language expressions in English, about the action of commands given by humans. Action robot commands refer to the movements of the legs, hands, head, eyes, and movements of the entire robot (e.g., walking forward, backward, sideways, sitting, crouching, standing, etc.) and its eyes (e.g., change color, scroll, blink). The robot used is the NAO (URL 46). To achieve this goal, the robot should be able to:

(i) recognize natural language actions that express natural movements, with the support of ontology that has been set;
(ii) ask the user to define a specific term, when it has not already been mentioned;
(iii) in the input text, performing sequential and complex actions, caused by natural language commands;
(iv) And understanding if a particular action is compatible with the current state, when the robot is in STATEFUL mode.

STATEFUL is one of the two modes, where the system can operate; the other one being STATELESS. In STATELESS, any human expression, which is correctly interpreted by the robot as an action command, after it is executed, will return to its original position. When in STATEFUL mode, the robot will know its current position and execute the command, only if it is compatible with the current status. In this mode, the robot does not return to its default position. For example, if the user tells the robot to stand on his right foot with a first command, the robot cannot execute a following command standing on the left foot, because they are two incompatible moves. Based on previous projects, the basic method used in this presented work was the creation of a specific ontology, with robot actions and robot body parts as the two main classes, as well as all the

specific actions that the robot can perform and all its body parts as individuals. In addition, a NLP engine module was created. This module analyzes the input expression in natural language given by the human and tries to extract and map, in the ontology, the actions that the robot will have to perform. Finally, the NAO robot has been programmed to check, to ask for clarifications (if necessary) and perform the specified actions. The Choregraphe suite (URL 75) was used, both to design the robot's movements and to manage the speech coming. The final speech file is forwarded in an IBM Watson Speech to Text module (URL 76) and then for analysis in the Stanford CoreNLP unit (URL 68). All natural language sentences, provided by the users, and the identified commands will be saved by an architecture in the cloud, which leverages HDFS (URL 77). The reason for that is twofold: firstly, there is the opportunity to look at a set of sentences given by the users, which have not been recognized by the NLP engine for some reason; secondly, all these data will be employed to acquire knowledge and understanding on how different groups of people interact with the robot. For the NLP module, the CoreNLP pipeline (URL 78) has been used. There is a reference about improvement for some tools in this part of the system. For example, the CoreNLP POS tagger may be replaced by Spacy (URL 79), which seems to work better. Finally, the system performance about the obtained precision, for sentences provided by the users to the robot, was 91% for the single action and 90% for multiple actions.

In the work (No. 27) of Saravanan & Sivaramakrishnan (2019), a proposal is presented to create an industrial robot control system of 5 degrees of freedom, with commands in Indic Languages, namely English (Indian accent), Hindi and Tamil. It is an initial attempt to implement speech recognition in the field of robotics, to control the robot from point to point, so that it can eventually perform a specific task. The whole process was developed both for voice commands on a microphone, directly connected to the system, and for voice commands via mobile phone. The system must recognize and edit voice, in order for a text to appear without an online connection, i.e., it is not desirable to use a well-known web speech-to-text web application. For this reason, the CMF Sphinx (URL 16) open source platform is selected, through which an SRS (speech recognition system) is built, which includes the following functions:

- Creating the text corpus.
- Creating the speech corpus.
- Extracting features.
- Train the system for speech recognition.
- Developing a decoder, which can score and match the spoken word.

For each of the above cases, the following techniques and tools were used:

- For text corpus, a list of commands for each language and a dictionary with phonemes for each language as well.
- For speech corpus, binary files and models using the CMU Sphinx API that can be developed in any operating system and in any programming language.
- For Extracting features, the MFCC (Sahidullah & Saha, 2012) technique on recorded speech.
- For training, it uses the Baum-Welch re-estimation procedures.
- In the decoder for offline recognition, we find the forward-backward algorithm and the Viterbi algorithm and it consists of 5 levels that use language, dictionary, acoustic models (to collect HMMs) and transition probabilities in HMMs.

In the case of mobile phone commands, a mobile app is developed to control a robot through speech-based interfaces using HC-05 Bluetooth technology (URL 33). Our conclusion is that this paper does not deal with the development of the robot's understanding, because it simply identifies specific commands with specific movements. It focuses mainly on the offline SRS process and for specific languages.

In the case (No. 28) of Thomason *et al.* (2019), an end-to-end pipeline is presented, initially low-resource, for understanding and translating natural language commands to discrete robot actions, and use clarification dialogs to jointly improve language parsing and concept grounding. On this concept, a robot agent is proposed and evaluated that leverages conversations with humans for this expansion. The agent uses the Combinatory Categorial Grammar (CCG) formalism (URL 52) to facilitate parsing. Word embeddings (URL 53) augment the lexicon at test time to recover from out-of-vocabulary words. The BWIBot (Khandelwal *et al.*, 2017) is used, which includes the Google Speech API (URL 54) for Speech-to-text conversion, and the Festival Speech Synthesis System (URL 55) for verbalizing audio responses. Tabletop perception is implemented with RANSAC algorithm (URL 56) plane fitting and Euclidean clustering, as provided by the Point Cloud Library (URL 57). The agent is trained on Mechanical Turk conversations (URL 23), transferring learned linguistic and perceptual knowledge across platforms, from simple simulations to real world applications.

Finally (No. 29) in Tada *et al.* (2020), a new method is described that enables a service robot to understand spoken commands in a robust manner, using off-the-shelf ASR systems and an encoder-decoder neural network with noise injection. This method of language understanding for robot-directed spoken commands is

called *sequence-to-sequence with noise injection* (Seq2Seq-NI). To construct the method for converting an input speech recognition result with errors into a sequence of elemental commands, by considering a set of actions that can be carried out by the robot, a semantic parser is prepared that is highly resistant to recognition errors by injecting artificial noise. Noise injection has often been used to increase the robustness of neural networks. So, noise is injected into a phoneme sequence, which is the input data of the semantic parser, in the training datasets. The tools and individual methods that were used during the development and evaluations of the proposed method of Seq2Seq-NI are:

- a Seq2Seq (Sutskever *et al.*, 2014) model, which is the heart of the semantic parser;
- a suitable unit that injects noise into the phoneme sequences, which is the stochastic deformation model (SDM);
- the CMU Pronouncing Dictionary (URL 14) to represent the English phonemes;
- two different ASR systems, namely, the Google Speech API (URL 15) and CMU sphinx (URL 16);
- a GPSR (URL 17) sentence generator to generate the English sentences representing robot-directed commands.

One more ASR system can be also used, that is Julius (URL 18). The results of the experiment indicate that Seq2Seq-NI outperforms the baseline methods. Noise injection clearly improves the understanding of spoken commands. It was also shown that an attention mechanism contributes to an improved performance of the semantic parsing. In addition, an experiment has been conducted to evaluate the influence of the injected noise, and it was found that a noise level, simulating the actual recognition error rate of the ASR, improves the performance of Seq2Seq-NI.

3.2 Discussion

Tellex *et al*. (2020) present a system flow diagram for a language-using robot. This diagram constitutes a complete HRI system, based on speech with real actions (Fig. 1). The authors describe this system, starting with natural language input as sound collection, with a microphone or alternatively as text. In any case, every input is transformed in text for the next step of processing. Following NLP methods, the text is converted to a semantic representation. The next step is the "Grounding" unit. This unit, in addition, collects information from a unit that is called "World representation", which concerns the environment that is covered by the robot through the "Perception" unit. The World representation unit is also part of the system's decision-making. Completing the input of the "Grounding" unit and after processing, the results are transferred to the "Grounding representation" unit and then to the "Decision-making" unit. Here it is determined exactly what the robot should do. This will be either a real-world action, through the "Physical actions" unit, or a natural language audio output, through the "Language actions" and "Language output" units. We end up with the authors' statement that this study does not cover the flow diagram (Fig. 1) in the levels of the robot's perception, the physical world's representation, as well as action in the real world.

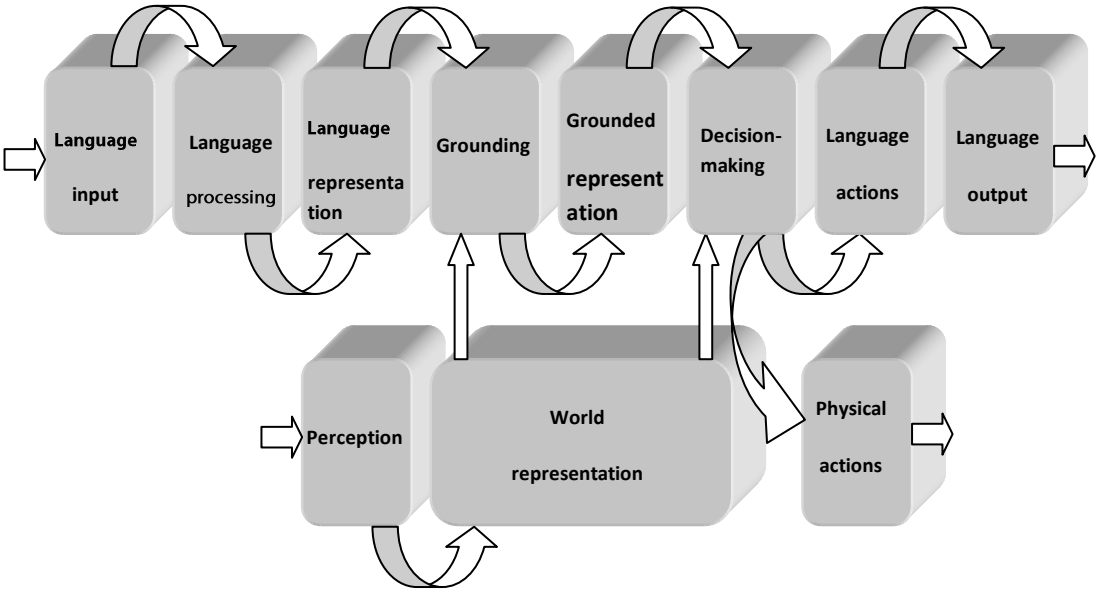

Figure 1: Typical flow diagram for a language-using robot (from: Tellex *et al*., 2020, after adaptation).

The works studied can be classified according to the above diagram and where is the interest of each one centered (Fig. 2). This is done with the intention of gaining a good knowledge of the research trends during these last years, to holistically cover such a formal system. Referring to the work table (Table 1), the discussion starts with work No. 1:

- This work (No. 1), a social robot, covers the entire diagram, although emphasis is placed on augmented dialogue.
- No. 2 is not focused on the subject of speech.
- No. 3 focuses on lexical entrainment but covers the whole diagram, using the movement of the hands of a social robot.
- No. 4 is developing a Natural Language Interface for the Hungarian language, into a social robot, and does not cover the area of speech creation.
- No. 5 does not cover the decision area and the speech output area, so it lags behind in the completion of the understanding.
- No. 6 focuses on dialogue and not on grounding and decision-making.
- No. 7 delves into the acquisition of knowledge, through Multi-mode Natural Language Processing for social robots and also involves the decision-making unit.
- No. 8 focuses on speech but covers the whole diagram.
- No. 9 researches the Spoken language understanding and does not cover the speech generation.
- No. 10 researches the Real-time natural language corrections and does not cover speech generation.
- No. 11 concerns an Ambient Assisted Living environment, as No. 15 does.
- No. 12 does not cover speech generation.
- No. 13 researches the perception in dialogue errors recognition, while the whole application is realized in simulation, so it does not cover motion.
- No. 14 does not cover the perception and the world representation in units, as well as speech output, but there is text in natural language.
- No. 15 concerns an Ambient Assisted Living environment and, although there are actions from the robot, there are not movement actions but only devices' controls are performed.
- No. 16 can be identified with the diagram, without having speech as input/output, without motion, and the production of a text description in natural language being done in the grounding representation unit.
- No. 17 does not cover speech generation.
- No. 18 is based on an autonomous mission vehicle and creates a dialog box, to monitor the machine's behavior by answering the questions "why" and "why not".
- No. 19 is a proposal for the understanding and communication actions of a robot. Because there, it does not seem to have been any practical testing of the proposed model, this topic is placed at the center of the diagram, related to understanding, regardless of input and action type.
- No. 20 is related to Voice Control for Web Applications and does not cover grounding and movement.
- No. 21 does not cover speech generation.
- No. 22 covers the entire diagram with a campus navigation robot, with an emphasis on dialogue.
- No. 23 does not cover the speech output of the diagram.
- No. 24 researches into speech recognition in noisy environments and identifies only with the first part of the diagram.
- No. 25 covers the whole diagram but the robot is in a virtual environment.
- No. 26 does not cover the perception and world representation parts of the diagram.
- No. 27 is a voice controlled robotic arm for Indian dialects that does not perform speech and decision-making, while it has no environment perception.
- No. 28 covers the entire diagram with a system that acquires information from dialog with humans.
- No. 29 creates a powerful system for understanding the incoming commands, with noise injection, and it covers the first blocks of the diagram.

The results of the above procedure are shown in Figure 2. We observe that the area with the largest gap is that of speech generation. We also note that only six studies cover the entire diagram, two of which refer to social robots, two to laboratory experimental models, one to campus robots and one to robots in a virtual environment. In general, it has been noticed that most research is done on ready-made platforms that are

basically social humanoid robots. It has been also noticed that some studies were conducted in languages other than English and more specifically in Hungarian, Indian dialects and Spanish.

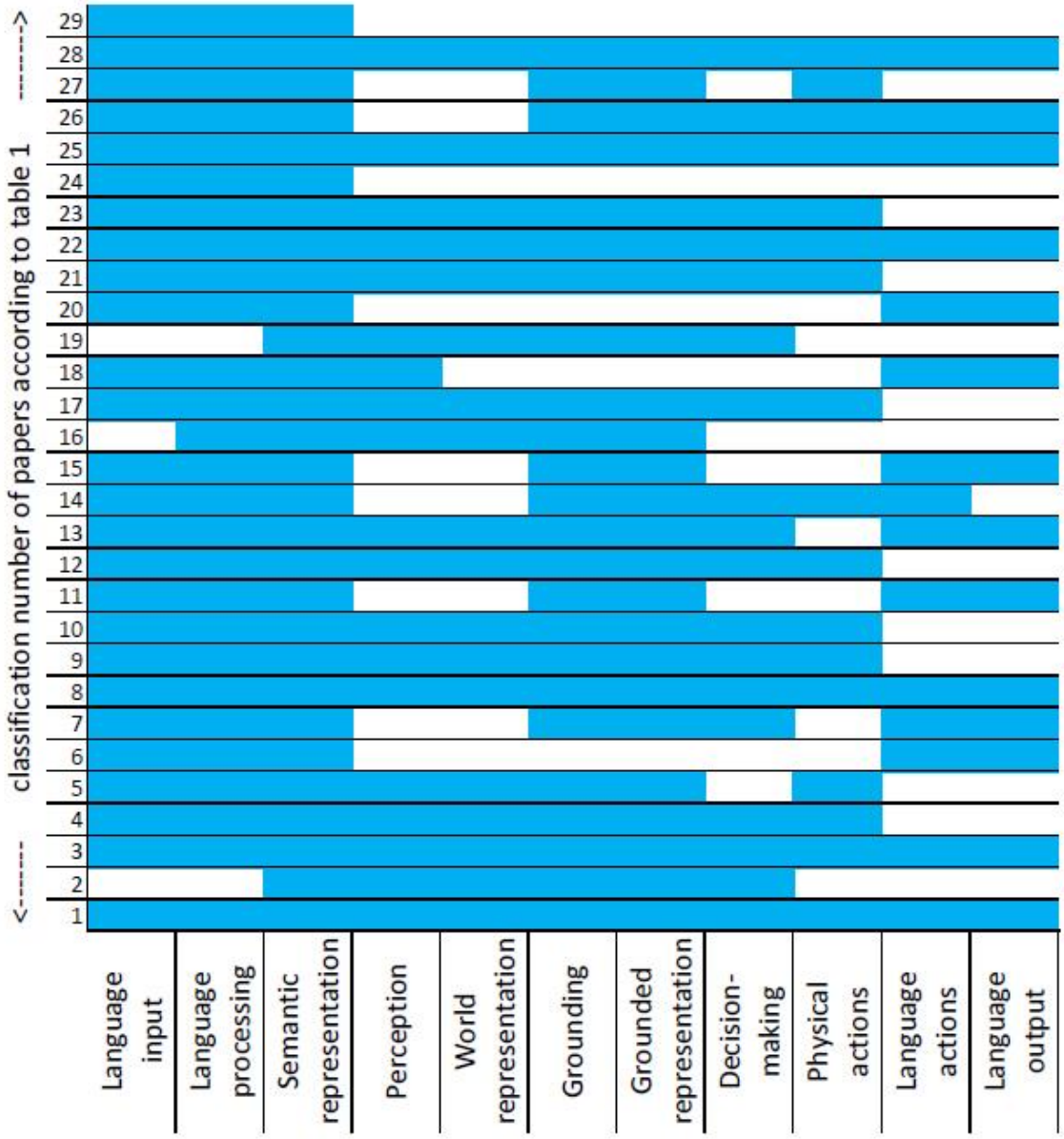

Figure 2: Coverage area of each numbered study, according to the standard flow diagram for a language-using robot (Fig. 1).

4. Conclusion

This study was conducted to gather information about the state of the art in HRI/HMI that uses natural language with the ability of understanding and taking action. There were 29 papers, selected for a final study, from a large number that appeared in our search and were close to our research interests/scope. In addition, through these works, the research interest of the scientific community is outlined. Initially, in order to implement our proposal, a gap is noticed in applications for the Greek language. It has been also noticed that, to a large extent, the issue of space is covered, unlike the issue of time. In a series of commands that enter a system, the time order of the commands must not be a mandatory condition. There could be an amount of commands without time continuity, where the machine should be able to place these commands in the right time-order, using every way to acquire further information, as even to ask questions or notice the space itself. In previous works (Giachos, 2015; Giachos *et al.*, 2017a; 2017b), this last topic has been covered in a

simulation environment, by using an artificial language, with very promising results. The most important element of this study was a computational semantic model, called OMAS-III, which was the heart of the system. Nothing similar has been encountered in the presented studies. In conclusion and given the gap mentioned above, it will be attempted, in the near future, to expand this previous work, from the level of simulation to the level of real-world implementation.

References

[1] Almagro-Cádiza M., Fresnoa V. & López F. d. l. P., 2018, Speech gestural interpretation by applying word representations in robotics, *Integrated Computer-Aided Engineering*, vol. 1, pp. 1-13.
[2] Alonso-Martín F., Castro-González A., Luengo F. J. F. D. G. & Salichs M. Á., 2015, Augmented Robotics Dialog System for Enhancing Human–Robot Interaction, *Sensors*, vol. 15, iss. 7, pp. 15799-15829.
[3] Altman N. S., 1992, An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician*, vol. 46, iss. 3, pp. 175–185.
[4] Ash R. B., 1966, *Information Theory: Tracts in Pure & Applied Mathematics*, New York, John Wiley & Sons Inc.
[5] Asimov I., 1942, Runaround, *Street & Smith*, World Heritage Encyclopedia, http://self.gutenberg.org/articles/eng/Runaround_(story)
[6] Asimov I., 1950, *I, Robot*, New York City, Gnome Press.
[7] Barthelmess U. & Furbach U., 2014, *Do we need Asimov's Laws?*, Cornell University, https://arxiv.org/abs/1405.0961
[8] Bastianelli E., Castellucci G., Croce D., Basili R. & Nardi D., 2017, Structured learning for spoken language understanding in human- robot interaction, *The International Journal of Robotics Research*, vol. 36, pp. 660–683.
[9] Bastianelli E., Nardi D., Aiello L. C., Giacomelli F. & Manes N., 2016, Speaky for robots: the development of vocal interfaces, *Applied Intelligence*, vol. 44, pp. 43-46.
[10] Bella M. & Hanington B., 2012, *Universal Methods of Design*, Beverly, MA, Rockport Publishers (p. 204).
[11] Bordag S., 2008, A Comparison of Co-occurrence and Similarity Measures as Simulations of Context, *In 9th CICLing*, pp. 52-63.
[12] Broad A., Arkin J., Ratliff N., Howard T. & Argall B., 2017, Real-time natural language corrections for assistive robotic manipulators, *The International Journal of Robotics Research*, vol. 36, iss. 5-7, pp. 1-15.
[13] Bui H.-D. & Chong N. Y., 2018, An Integrated Approach to Human-Robot-Smart Environment, *In IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, Genoa, Italy.
[14] Bui H.-D., Pham C., Lim Y., Tan Y. & Chong N. Y., 2017, *Integrating a Humanoid Robot into ECHONET-Based Smart Home Environments*, Springer International Publishing, pp. 314–323.
[15] Cascianelli S., Costante G., Ciarfuglia T. A., Valigi P. & Fravolini M. L., 2018, Full-GRU Natural Language Video Description for Service Robotics Applications, *IEEE Robotics and Automation Letters*, vol. 3, iss. 2, pp. 841-848.
[16] Chen C.-H., Wu M.-C. & Wang C.-C., 2019a, Cloud-based Dialog Navigation Agent System for Service Robots, *Sensors and Materials*, vol. 31, iss. 6, pp. 1871-1891.
[17] Chen H., Tan H., Kuntz A., Bansal M. & Alterovitz R., 2019b, *Enabling Robots to Understand Incomplete Natural Language Instructions Using Commonsense Reasoning*, Cornell University, https://arxiv.org/abs/1904.12907v1?source=post_page
[18] Du G., Chen M., Liu C., Zhang B. & Zhang P., 2018, Online robot teaching with natural human-robot interaction, *IEEE Transactions on Industrial Electronics*, vol. 65, iss. 12, pp. 9571 - 9581.
[19] Garcia F. J. C., Robb D. A., Liu X., Laskov A., Patron P. & Hastie H., 2018, *Explain Yourself: A Natural Language Interface for Scrutable Autonomous Robots*, Cornell University, https://arxiv.org/abs/1803.02088
[20] Giachos I., 2015, *Implementation of OMAS-III as a Grammar Formalism for Robotic Applications*, National & Kapodistrian University of Athens and National Technical University of Athens, Joint Postgraduate Dissertation (in Greek),.
[21] Giachos I., Papakitsos E.C. & Chorozoglou G., 2017a, Exploring natural language understanding in robotic interfaces, *International Journal of Advances in Intelligent Informatics*, vol. 3, iss. 1, pp. 10-19.
[22] Giachos I., Papakitsos E.C. & Makrygiannis P., 2017b, An Experiment of Language Communication Learning in a Robotic System., *In Proceedings of the 9th Conference on Informatics in Education (CIE2017)*, University of Piraeus, pp. 46-56 (in Greek).
[23] Griffin A., 2017, Facebook's artificial intelligence robots shut down after they start talking to each other in their own language, *Independent*, Monday 31 July 2017, https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html
[24] Gweon, H. & Saxe R., 2013. Developmental cognitive neuroscience of Theory of Mind. *In Neural Circuit Development and Function in the Brain: Comprehensive Developmental Neuroscience*, J. Rubenstein & P. Rakic (Eds.), Elsevier.
[25] Hellström T. & Bensch S., 2018, Understandable robots, *Paladyn, Journal of Behavioral Robotics*, vol. 9, iss. 1, pp. 110–123.
[26] Hermansky H., 2019, Coding and decoding of messages in human speech communication - Implications for machine recognition of speech Communication, *Speech Communication*, vol. 106, pp. 112–117.

[27] Igarashi T. & Inami M., 2015, Exploration of Alternative Interaction Techniques, *IEEE Computer Graphics and Applications*, vol. 35, pp. 33-41.
[28] Iio T., Shiomi M., Shinozawa K., Shimohara K., Miki M. & Hagita N., 2015, Lexical Entrainment in Human Robot Interaction, *International Journal of Social Robotics*, vol. 7, iss. 2, pp. 253–263.
[29] Kandhari M. S., Zulkernine F. & Isah H., 2018, 2018 IEEE A Voice Controlled E-Commerce Web Application, *In 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver.
[30] Kaven O., 2005, Zensys' Z-Wave Technology, *PC Magazine*, January 8, 2005, https://www.pcmag.com/archive/zensys-z-wave-technology-142295
[31] Khandelwal P., Zhang S., Sinapov J., Leonetti M., Thomason J., Yang F., Gori I., Svetlik M., Khante P., Lifschitz V., Aggarwal J. K., Mooney R. & Stone P., 2017, BWIBots: A platform for bridging the gap between AI and human–robot interaction research , *The International Journal of Robotics Research*, vol. 36, iss. 5-7, pp. 635-659.
[32] Kovács L., Barabás P., 2015, Grammar Representation Forms in Natural Language Interface for Robot Controlling, *In Emergent Trends in Robotics and Intelligent Systems, Advances in Intelligent Systems and Computing*, vol. 316, Sinčák P., Hartono P., Virčíková M., Vaščák J., Jakša R. (eds.), Springer, Cham.
[33] Marge M. & Rudnicky A. I., 2019, Miscommunication Detection and Recovery in Situated, *ACM Transactions on Interactive Intelligent Systems*, vol. 9, iss. 1, pp. 1–40.
[34] Mikolov T., Chen K., Corrado G. & Dean J., 2013, *Efficient Estimation of Word Representations in Vector Space*, Cornell University, https://arxiv.org/abs/1301.3781
[35] Paul R., Barbu A., Felshin S., Katz B. & Roy N., 2017, Temporal Grounding Graphs for Language Understanding with Accrued Visual-Linguistic Context, *In Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI-17.
[36] Poncela A. & Gallardo-Estrella L., 2015, Command-based voice teleoperation of a mobile robot via a human-robot interface, *Robotica*, vol. 33, pp. 1-18.
[37] Rajaraman A. & Ullman J. D., 2011, Mining of Massive Datasets, Cambridge University Press.
[38] Recupero D. R. & Spiga F., 2019, Knowledge acquisition from parsing natural language expressions for humanoid robot action commands, *Information Processing and Management*, in press, https://www.sciencedirect.com/science/article/abs/pii/S0306457319303322
[39] Sahidullah M. & Saha G., 2012, Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition, *Speech Communication*, vol. 54, iss. 4, pp. 543–565.
[40] Saravanan N. & Sivaramakrishnan R., 2019, Command and control of industrial manipulator through speech based interfaces in Indic Languages, *The Journal of Supercomputing*, vol. 75, pp. 5106–5117.
[41] Sarnataro V., 2012, Dragon NaturallySpeaking (DNS) 12 Review, *Technology Guide*, 2012-11-08, http://www.technologyguide.com/review/dragon-naturallyspeaking-dns-12-review/
[42] Schlögl S., Doherty G. & Luz S., 2015, Wizard of Oz Experimentation for Language Technology Applications: Challenges and Tools, *Interacting with Computers*, vol. 27, pp. 592–615.
[43] Schütte N., Namee B. M. & Kelleher J., 2017, Robot perception errors and human resolution strategies in situated human–robot dialogue, *Advanced Robotics*, vol. 31, iss. 5, pp. 243-257.
[44] Sutskever I., Vinyals O. & Le Q. V., 2014, *Sequence to sequence learning with neural networks*, Cornell University, https://arxiv.org/abs/1409.3215
[45] Tada Y., Hagiwara Y., Tanaka H. & Taniguchi T., 2020, Robust Understanding of Robot-Directed Speech Commands Using Sequence to Sequence with Noise Injection, *Frontiers in Robotics and AI*, vol. 6, iss. 144, pp. 1-12.
[46] Tellex S., Gopalan N., Kress-Gazit H. & Matuszek C., 2020, Robots That Use Language, *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, 3 May 2020.
[47] Thomason J., Padmakumary A., Sinapovz J., Walker N., Jiangy Y., Yedidsiony H., Harty J., Stoney P. & Mooney R. J., 2019, Improving Grounded Natural Language Understanding through Human-Robot Dialog, *In 2019 International Conference on Robotics and Automation (ICRA)*, Montreal, Canada.
[48] URL 01: https://www.softbankrobotics.com/us/pepper
[49] URL 02: https://dialogflow.com/
[50] URL 03: http://doc.aldebaran.com/1-14/dev/naoqi/index.html
[51] URL 04: http://doc.aldebaran.com/2-4/naoqi/audio/altexttospeech.html
[52] URL 05: https://web.archive.org/web/20120629211518/http://www.middleware.org/whatis.html
[53] URL 06: http://depot.universaal.org/
[54] URL 07: https://www.knx.org/knx-en/for-professionals/What-is-KNX/A-brief-introduction/index.php
[55] URL 08: https://zigbeealliance.org/
[56] URL 09: https://echonet.jp/about_en/
[57] URL 10: https://blog.dialogflow.com/post/apiai-new-name-dialogflow-new-features/
[58] URL 11: https://echonet.jp/introduce_en/gz-000057_en/
[59] URL 12: https://ontology.tno.nl/saref/
[60] URL 13: https://www.roboticstoday.com/robots/robovie-r-ver2-description
[61] URL 14: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[62] URL 15: https://cloud.google.com/speech-to-text/
[63] URL 16: https://cmusphinx.github.io/

[64] URL 17: https://github.com/komeisugiura/GPSRsentence_generator
[65] URL 18: https://github.com/julius-speech/Julius
[66] URL 19: https://www.kinovarobotics.com/en/knowledge-hub/mico-robotic-arm
[67] URL 20: https://github.com/TalAter/annyang
[68] URL 21: http://wiki.ros.org/roslibjs
[69] URL 22: https://github.com/tmhoward/h2sl
[70] URL 23: https://www.mturk.com/mturk/welcome
[71] URL 24: https://www.nltk.org/
[72] URL 25: http://conceptnet.io/
[73] URL 26: http://www.aclclp.org.tw/use_ckip.php
[74] URL 27: https://nlp.stanford.edu/software/lex-parser.shtml
[75] URL 28: https://eclipse-ee4j.github.io/jersey/
[76] URL 29: https://howtodoinjava.com/jersey-jax-rs-tutorials/
[77] URL 30: https://www.ros.org/
[78] URL 31: http://wiki.speech.cs.cmu.edu/teamtalk/index.php/Main_Page
[79] URL 32: http://wiki.speech.cs.cmu.edu/olympus/index.php/Olympus
[80] URL 33: https://wiki.eprolabs.com/index.php?title=Bluetooth_Module_HC-05
[81] URL 34: http://www.ontolinux.com/community/phoenix/Phoenix_Manual.pdf
[82] URL 35: https://47degrees.github.io/helios/docs/
[83] URL 36: https://dl.acm.org/doi/10.1145/258368.258412
[84] URL 37: https://github.com/kalliope-project/kalliope
[85] URL 38: http://sourceforge.net/projects/usarsim
[86] URL 39: http://people.cs.ksu.edu/~dag/Activmedia/p2opman4.pdf
[87] URL 40: https://protegewiki.stanford.edu/wiki/Protege_Ontology_Library
[88] URL41:https://www.google.com/urlsa=t&rct=j&q=&esrc=s&source=web&cd=16&ved=2%20%20%20ahUKEwi29JWyxLboAhWpSRUIHc2UC_A4ChAWMAV6BAgGEAE&url=http%3A%2F%2Fwww.googoltech.com%2Fdown-383.html&usg=AOvVaw0ybozw4AsvyidR0kQ5UyKN
[89] URL 42: https://www.robots.com/robots/kuka-kr-6-r700-sixx
[90] URL 43: http://www.bbc.co.uk/newsbeat/10996389
[91] URL 44: https://web.archive.org/web/20121003161057/http://www.eetimes.com/electronics-news/4216978/GPS-system-with-IMUs-tracks-first-responders
[92] URL 45: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-sdk?tabs=windows
[93] URL 46: https://www.softbankrobotics.com/emea/en/nao
[94] URL 47: https://freeling-user-manual.readthedocs.io/en/latest/basics/
[95] URL 48: http://multiwordnet.fbk.eu/english/home.php
[96] URL 49: http://corpus.byu.edu/coca/
[97] URL 50: https://www.ultraleap.com/
[98] URL 51: http://www.ai.mit.edu/courses/6.891-nlp/READINGS/adwait.pdf
[99] URL 52: http://groups.inf.ed.ac.uk/ccg/
[100] URL 53: https://en.wikipedia.org/wiki/Word_embedding
[101] URL 54: https://cloud.google.com/speech/
[102] URL 55: http://www.cstr.ed.ac.uk/projects/festival/
[103] URL 56: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/FISHER/RANSAC/
[104] URL 57: http://pointclouds.org/documentation/
[105] URL 58: https://www.nuance.com/omni-channel-customer-engagement/voice-and- ivr/automatic-speech-recognition/nuance-recognizer/recognizer-languages.html
[106] URL 59: http://doc.aldebaran.com/2-1/naoqi/index.html
[107] URL 60: https://web.archive.org/web/20110817012415/http://www.cslu.ogi.edu/toolkit/
[108] URL 61: https://www.academia.edu/17292686/Jaspis_-_a_framework_for_multilingual_adaptive_speech_applications
[109] URL 62: https://www.epfl.ch/campus/associations/list/qc/
[110] URL63:https://www.google.gr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=13&ved=2ahUKEwiUmvm8tOjoAhVNs4sKHZPpDW8QFjAMegQICRAB&url=https%3A%2F%2Fhci.stanford.edu%2Fpublications%2F2000%2Fsuede%2Fsuede
[111] uist2000.pdf&usg=AOvVaw0WpRywa6FWhqazLUJA72LY
[112] URL 64: http://www.softdoc.de/woz/index.html
[113] URL 65: https://www.researchgate.net/figure/Configuration-of-the-NEIMO-platform_fig1_2440673
[114] URL 66: https://www.jst.go.jp/erato/igarashi/en/
[115] URL 67: http://www.kelp-ml.org/?page_id=728
[116] URL 68: https://stanfordnlp.github.io/CoreNLP/
[117] URL 69: https://framenet.icsi.berkeley.edu/fndrupal/WhatIsFrameNet
[118] URL 70: https://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html
[119] URL 71: http://htk.eng.cam.ac.uk/
[120] URL 72: http://wiki.ros.org/roboframenet

[121] URL 73: http://www.seebyte.com/military/neptune/
[122] URL 74: http://www.seebyte.com/oceanography/seetrack/
[123] URL 75: http://doc.aldebaran.com/2-4/software/choregraphe/choregraphe_overview.html
[124] URL 76: https://www.ibm.com/cloud/watson-speech-to-text
[125] URL 77: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
[126] URL 78: https://stanfordnlp.github.io/stanfordnlp/index.html
[127] URL 79: https://spacy.io/
[128] URL 80: http://www.robotics-kejia.com/
[129] URL 81: https://en.wikipedia.org/wiki/Open_Mind_Common_Sense
[130] URL 82: https://web.archive.org/web/20140826071530/http://www.rethinkrobotics.com/products/baxter/
[131] URL 83: https://www.bloomberg.com/news/articles/2014-11-06/amazon-echo-is-a-listening-talking-music-playing-speaker-for-your-home
[132] URL 84: http://start.csail.mit.edu/start-system.php
[133] URL 85: http://upplysingaoflun.ecn.purdue.edu/~qobi/cccp/sentence-tracker-language-learning.html
[134] URL 86: https://medium.com/@alittlepain833/simple-understanding-of-mask-rcnn-134b5b330e95
[135] URL 87: https://moveit.ros.org/
[136] Winograd T., 1971, Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, MIT Libraries, https://dspace.mit.edu/handle/1721.1/7095
[137] Xie J., Chen X. & Ji J., 2015, Multi-mode Natural Language Processing for human-robot interaction, Web Intelligence, vol. 13, pp. 267–278.
[138] Yongda D., Fang L. & Huang X., 2018, Research on multimodal human-robot interaction based on speech and gesture, Computers and Electrical Engineering, vol. 72, pp. 443-454.
[139] Zarchan P. & Musoff H., 2000, Fundamentals of Kalman Filtering: A Practical Approach. American Institute of Aeronautics and Astronautics, Incorporated.