Rejuvenation Research
© Mary Ann Liebert, Inc.
DOI: 10.1089/rej.2018.2075

Dialogue Systems and Conversational Agents for Patients with Dementia: the human-robot interaction

Alessandro Russo[1], Grazia D'Onofrio[2, 3], Aldo Gangemi[1], Francesco Giuliani[4], Misael Mongiovi[1], Francesco Ricciardi[4], Francesca Greco[2], Filippo Cavallo[3], Paolo Dario[3], Daniele Sancarlo[2], Valentina Presutti[1], Antonio Greco[2]

[1] Semantic Technology Laboratory (STLab), Institute for Cognitive Sciences and Technology (ISTC) - National Research Council (CNR), Rome, Italy.

[2] Complex Unit of Geriatrics, Department of Medical Sciences, IRCCS "Casa Sollievo della Sofferenza", San Giovanni Rotondo, Foggia, Italy.

[3] The BioRobotics Institute, Scuola Superiore Sant'Anna, Pontedera, Italy.

[4] ICT, Innovation and Research Unit, IRCCS "Casa Sollievo della Sofferenza", San Giovanni Rotondo, Foggia, Italy.

***Correspondence**:
Grazia D'Onofrio
Email: g.donofrio@operapadrepio.it
Phone Number: +39 0882 410271
Fax Number: +39 0882 410271
Address: Viale Cappuccini, 1 – 71030, San Giovanni Rotondo (FG), Italy

Keywords: human-robot interaction; natural language processing; spoken dialogue systems; conversational agents; socially assistive robotics.

Abstract

This study aimed to identify and describe the fundamental characteristics of spoken dialogue systems and their role for supporting human-robot interaction and enabling the communication between socially assistive robots and patients with dementia. Firstly, this work provides an overview of spoken dialogue systems, by considering the underlying technologies, approaches, methods and general issues. Then, the analysis focuses on studies, systems and approaches that have investigated the role of dialogue systems and conversational agents in the interaction with elderly people with dementia, by presenting the results of a literature review.

While the overview of spoken dialogue systems relies on existing surveys and reviews, a research was conducted to identify existing works in the literature that have investigated the role of conversational agents and dialogue systems for the elderly and people with cognitive impairments. Inclusion criteria were: 1) use of conversational agents, dialogue systems or language processing tools for people with cognitive impairments; 2) age ≥ 60 years; 3) diagnosis of dementia according to NIAAA criteria; 4) presence of tests or experiments with qualitative or quantitative results.

Initially 125 studies published between 2000 and 2017 were identified, of which 12 were met the inclusion criteria. The review identifies the issues and challenges that reported when conversational agents and speech-based interfaces have been used for interacting with people with cognitive impairments. In addition, the review led to the identification of studies that have investigated speech processing and natural language processing capabilities to assess the cognitive status of people with dementia.

Introduction

Worldwide, about 50 million people have dementia, with nearly 60% living in low- and middle-income countries [1]. The estimated proportion of the general population aged 60 and over with dementia at a given time is between 5 to 8 per 100 people [1]. The total number of people with dementia is projected to reach 82 million in 2030 and 152 million in 2050 [1].

The difficulty with communication is an initial sign in the dementia onset, causing isolation and loneliness. In light of this assumption, the social robotic solutions could be a real and effective help for elderly people with dementia in terms of support and prevention.

The ability to communicate using natural language is a fundamental requirement for a social robot that aims at providing support for patients with dementia. Spoken dialogue is generally considered as the most natural way for social human-robot interaction [2]. Research efforts in the field of Human-Robot Interaction (HRI) increasingly focus on the development of robots equipped with intelligent communicative abilities, in particular speech-based natural-language conversational abilities [3]. These efforts directly relate to the research area of computational linguistics, generally defined as “the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content” [4]. The advances and results in computational linguistics provide a foundational background for the development of so called Spoken Dialogue Systems, i.e., computer systems designed to interact with humans using spoken natural language [5]. The strong interest in dialogue systems and conversational agents derives from their great potential in several domains. Significant R&D investments from major companies (such as Google, Apple, Microsoft, Nuance) resulted in the integration of speech-based technologies in mobile platforms, with “digital personal assistants” entering the market in the form of well-known applications like Apple Siri, Google Now or Microsoft Cortana. Therefore, for some real-world scenarios dialogue systems are mature enough to be commercially deployed and provide support for specific tasks that require access to information and services, as in the case of applications for mobile platforms or in-car navigation and control systems [6]. Hirschberg and Manning [4] identify four main factors that enabled the significant progresses in computational linguistics and dialogue systems witnessed in the last years, namely: (i) the availability of increased computational power;

(ii) the availability of very large amounts of linguistic data in digital form; (iii) the development of effective machine learning techniques; and (iv) the increased understanding of the structure and role of natural language, in particular in social contexts. Despite these factors, the availability of dialogue systems for supporting complex scenarios is limited to research prototypes, including the integration in service robots and assistive technologies requiring human-robot interaction to provide cognitive and physical assistance for disabled or elderly persons. Simulating human-to-human communication to enhance and ease human-to-machine communication is still a very challenging research task, in particular when the goal is to enable a natural, context-aware, adaptive and intelligent interaction. Achieving such a complex goal requires a multidisciplinary approach with a contribution from different scientific disciplines, including linguistics and cognitive sciences, artificial intelligence and software engineering. The aim of this work is to identify and describe the fundamental characteristics of spoken dialogue systems, by considering their main components and functionalities and outlining the methods and approaches that have been proposed and adopted. While performing a detailed and comprehensive review of the increasing number of existing dialogue systems is out of the scope of our analysis, we mainly focus on providing an overview of the underlying technologies, approaches, methods and general issues. We then focus on studies, systems and approaches that investigated the role of dialogue systems and conversational agents in the interaction with elderly people with dementia.

1. Conversational Agents: Chatterbots and Dialogue Systems

Early spoken language systems have been mainly developed for telephone-based conversational interfaces (Figure 1) [7, 8, 9, 10, 11]. Modern conversational interfaces are more sophisticated in that they perform a deeper analysis of utterances, make use of internal and external knowledge, and maintain a conversational status. Conversational agents can be classified in Chatterbots and Dialogue Systems [12, 13]. The main difference between them is that Chatterbots aim at keeping a conversation going (e.g., for entertainment purposes), without an explicit attempt at understanding the meaning of the utterances, while dialogue systems try to establish an effective communication channel between the system and the user, for sharing information, responding to commands and building a common knowledge. More in detail, chatterbots and dialogue systems usually differ in the following features: Text processing: dialogue systems perform a more sophisticated processing (semantic parsing of the sentences). Conversational state: dialogue systems can manage conversational plans and maintain a conversation state (e.g., by keeping a dialogue history). Internal knowledge: dialogue systems generally make use of internal knowledge to reply appropriately. Clearly, the classification in chatterbots and dialogue systems is blurred since good chatterbots need to perform some kind of understanding of the language to be able to respond appropriately, while existing dialogue systems are not able to fully understand (as a human would) the meaning of sentences. Chatterbots usually have limited language understanding capabilities and they mainly focus on replying in a natural way, rather than really understanding the language. For this reason they are usually more robust to errors but they are quite limited in all tasks that require a real understanding of the utterances (e.g., learning from speech and provide useful answers). Modern chatterbots, however, are quite complex and open up to more advanced tasks that are usually carried out by complex dialogue systems. One of the first chatterbots was ELIZA [14], developed in 1966, which simulated a psychotherapist. After the success of ELIZA, hundreds of chatterbots have been developed. For instance, PARRY [15] simulates a paranoid person. Chatterbot [16] was one of the first chatterbots to pass a restricted version of the Turing test. More recently, ALICE [17], a novel chatterbot based on the AIML language [18], was developed and won the Loebner Prize (a competition based on the Turing test). Other modern chatterbots, such as Cleverbot [19], learn conversational patterns from interaction corpora. Chatterbots have also been proposed for therapeutic purposes, e.g., monitoring and reducing adolescents stress [20] and treating people with Parkinson's disease [21]. Other chatterbots include Facade [4] and The Personality Forge [5]. ChatScript [22] is a recently developed chatterbot engine with advanced capabilities. The first chatterbot developed with ChatScript was Suzette [6], which won the 2010 Loebner Competition. Several languages, frameworks and toolkits for the development of dialogue systems have been recently proposed. The project TRINDI [23, 24], for instance, developed the framework TrindiKit for simplifying the implementation of dialogue processing theories, making the development of dialogue systems easier. Pamini [25] is another framework for defining human robot interaction strategies based on generic interaction patterns. A recent work on spoken dialogue systems was done by Berg [26] who designed and developed NADIA, a framework for the creation of natural dialogue systems. Both chatterbots and dialogue systems have their advantages and disadvantages. Chatterbots are usually more tolerant to errors and noise but they are often too simplistic for complex applications. On the other hand, dialogue systems are harder to build, less robust to noise and their conversation domains are often more restricted, but they can handle more complex conversations. Some approaches have been recently proposed to combine chatterbots an dialogue systems into hybrid systems, in order to take advantage of both paradigms, as discussed for example in [12] and [13].

2. Dialogue Systems: Architectural Model, Components and Approaches

Dialogue systems integrate multiple components, each providing a specific functionality, according to the reference conceptual architecture shown in Figure 2, where the modular structure of such systems and the information flow between the components are illustrated. A dialogue system consists of five main components providing the following functionalities: Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialogue Management (DM), Natural Language Generation (NLG), Text-to-Speech synthesis (TTS). Basically, user's utterances are converted in a textual representation by a speech recognizer and then processed by a language understanding unit. A dialogue manager interprets the semantic information, manages the conversation flow, and relies on a language generator for producing textual messages that are then rendered by a speech synthesizer. Simple dialogue systems and chatterbots may not have an internal knowledge, and hence they are stateless: they reply to every single request, without contextualization. More complex systems maintain a simple status of the conversation (e.g., memorize which answers have been already given) and hence their conversation appears more natural. Other systems (known bots) use a huge amount of knowledge (typically gathered from the Web) to improve the interaction and answer to questions. The internal knowledge can also be expanded with the knowledge gathered from speech analysis. In the following we detail the role of each component, primarily focusing on Natural Language Understanding and Dialogue Management, mainly in line with the analyses available in [27, 28, 29].

2.1. *Automatic Speech Recognition*

Automatic speech recognition (ASR) capabilities are provided by a Speech Recognizer or Speech-to-Text component. The Speech Recognizer is in charge of converting an input speech utterance into a textual representation. It receives as input the user's speech (typically captured by a microphone or provided as an audio file) and produces as its output a recognition hypothesis, as a sequence of words that most likely corresponds to what the user said. The recognized strings are often associated with some kind of confidence scores, where low values (typically in the range [0, 1]) represent low confidence in the correct word recognition. Confidence scores can be also associated with entire word sequences, according to a N-best recognition approach where a list of N recognition hypotheses ranked in terms of likelihood is generated as output. Depending on several factors, the textual result of the speech-to-text conversion may contain errors, e.g., in terms of inserted, substituted or missing words. Lopez-Cozar et al. generally relate possible errors to environmental conditions (such as noise), acoustic similarity between words, and specific phenomena concerned with spontaneous speech, such as false starts, filled pauses and hesitations [27]. These factors are systematically considered in [28], where the authors relate possible recognition errors to the variability in the speech signal, which in turn depends on: (i) linguistic variables, due to linguistic phenomena such as co-articulation (i.e, the acoustic realization and sound of a given phoneme may change depending on the context given by the preceding or following phonemes), (ii) speaker variability, covering differences between and within speakers, and (iii) channel variability, including the impact of noise and the properties of the transmission channel or input device (e.g., a microphone). The speech recognition process relies on the availability of a set of models to be matched with user's utterances encoded in the incoming speech signal. The vast majority of speech recognizers adopt a stochastic/probabilistic approach, by combining an acoustic model and a language model for a given language (e.g., English, Italian, French, etc.). In a nutshell, the acoustic model represents the basic speech units (phonemes or tri-phones) and their relationships. Acoustic models are typically represented as Hidden Markov Models (HMMs) and derived through a previous training process [30]. The language model provides knowledge about words' likelihood in a given sequence, in order to determine the words or sentences that are expected from the user.

Language models may be manually encoded in the form of finite state networks or grammars for a specific domain. These approaches provide good recognition accuracy in the case of well-defined sequences, but lack in flexibility as they assume that possible speech inputs are specified in advance and may force the user to comply with a static controlled vocabulary or restricted set of commands [31]. Advanced approaches aim at deriving probabilistic language models by applying a learning process over language corpora. Statistical language models estimated from data, such as N-gram models, provide significant flexibility but may have higher error rates. Recent approaches, as implemented in the Google speech API, are relying on knowledge graphs able to provide vocabularies with millions of entities, while advanced machine learning techniques, such as deep learning approaches, are being investigated [32, 33].

2.2. *Natural Language Understanding*

NLU component operates over the output of the Speech Recognizer, with the goal of deriving a semantic, meaningful representation of user's utterances. Language understanding is a complex task, that may involve both syntactic and semantic analysis, taking into account, on the one side, the inherent complexity of natural language processing (e.g., resulting from the presence of ambiguities, anaphora, ellipsis, etc.) and, on the other side, the quality of the output of the Speech Recognizer. The NLU component thus combines different modules, including error correction tools, semantic parsers, and reference resolution and contextual interpretation engines. Different solutions have been proposed and adopted for addressing the task of producing semantic representations from textual representations of user's utterances, as provided by the speech recognition component. Natural language parsers aim at building a representation of the input text taking into account its grammatical structure. According to the "principle of compositionality" identified in [28], sentences are analyzed on the basis of their constituent structure under the assumption that "the meaning of a sentence is a function of the meanings of its parts". Grammars thus represent the fundamental NLP approaches for text parsing and analysis to derive a semantic interpretation of user's utterances [31]. Basic shallow parsing approaches, widely used in dialogue systems, are based on the (manual) definition of task- or domain-specific patterns that are matched against the input text for identifying and extract specific parts. The processing aims at recognizing or spotting specific information-bearing phrases or constituents (e.g., locative phrases and temporal expressions), ignoring the rest. Shallow parsing is often coupled or even directly integrated with a speech recognition approach based on a controlled vocabulary or restricted grammar. These techniques are generally fast and efficient, and have the advantage of being easy to define and develop. Semantic shallow parsing approaches aim at identifying syntactical phrases, or chunks corresponding to basic semantic concepts, in natural language sentences to be labeled with semantic roles. In addition, a direct mapping of the understanding patterns to domain-specific concepts and semantics is possible. However, this domain-specificity, together with development efforts, pose limitations in terms of language coverage and robustness. Nevertheless, simple methods using words recognition, part-of-speech (POS) sequences (with the identification of nouns, verbs, preposition, etc.), or simple templates can often lead to notable results in domain-specific application scenarios [34]. Advanced approaches for NLU rely on domain-independent grammars and models, and aim at performing a grammatical and semantic analysis of the utterances with the goal of getting a rich understanding of their linguistic structure. These approaches rely on machine learning techniques and are grounded in formal languages (which enable automated reasoning) and computational linguistics theories. NLU tools, such as Stanford CoreNLP [35] and FRED, that are able to identify syntactic and semantic information, taking also into account the discourse context, have been developed. They rely on complex processing workflows that include POS tagging, named entities resolution, semantic role labeling, co-reference resolution and word-sense disambiguation. Multiple background theories are leveraged, such as Discourse Representation Theory, Combinatory Categorial Grammar and Frame Semantics. Linguistic frames are increasingly used for representing the output of the NLU process. Frame Semantics is a formal theory of meaning, whose basic idea is that humans can better understand the meaning of a single word by knowing the contextual knowledge related to that word [36]. Frame semantics allows real-world knowledge to be captured by semantic frames, which describe particular types of situations, actions, objects or events, and their semantic arguments or participants, characterized by specific semantic roles with respect to the situation described by the frame. Natural language understanding can thus be mapped to the frame detection or frame recognition task, which has the goal of recognizing complex relations in natural language text and build a machine-readable semantic representation. These approaches and techniques are strongly linked to machine reading, where interpretable structured knowledge is built from written text. However, written and spoken language differ in many ways. Although speech recognition provides a textual representation of user's speech, the textual encoding of user's utterances may have distinctive characteristics with respect to written language. In addition to possible errors introduced by the speech recognizer, user's utterances may not represent grammatically well-formed strings and can include linguistic constructions and irregularities that are typical of spontaneous speech, such as self-corrections, hesitations, repetitions and sentence fragments. These factors, which are particularly relevant in the utterances of elderly people and patients with dementia [37, 38], pose challenges to the parsing capabilities of language understanding systems and require specific techniques for robust processing, that range from syntactic/semantic input pre-processing (e.g., with the introduction of additional grammar rules able to recognize and handle some of these phenomena occurring sufficiently regularly) to partial parsing (similar to shallow parsing approaches) when a full linguistic analysis on each input term fails.

2.3. *Dialogue Management*

Dialogue Management (DM) capabilities are provided by a Dialogue Manager. The Dialogue Manager represents the core of a dialogue system: it keeps the conversational state of a dialogue and manages the overall human-machine interaction flow, taking into account user's utterances, the dialogue history, contextual information and background knowledge. Typical activities performed by the dialogue manager relate to decision-making and include: trigger a specific task or action (e.g., in the presence of a command issued by the user); provide information to the user (e.g., in response to a question); ask the user to provide information (e.g., through the generation of a question); undertake verification and/or confirmation steps in the case of uncertainty, recognition errors and incomplete information. An important task that crosses the boundaries between NLU and dialogue management is the recognition and identification of speech or dialogue acts expressing the function of an utterance. Speech act analysis goes beyond syntactic and semantic analysis, as it may rely on external information, such as the discourse context and the user model. As the function of an utterance relates to user's goals, it plays a significant role in the decision-making process of a dialogue manager. Bunt et al. provide a taxonomic classification of domain-independent, general-purpose functions for classifying utterances [39]. On the one side, information-transfer functions distinguish between information-seeking acts (i.e., the speaker asks for information) and information providing acts (i.e., the speaker provides information). On the other side, action discussion functions distinguish between commissive acts (i.e., the speaker commits to something, such as accepting/declining suggestions) and directive acts (i.e., the speaker suggests or instructs the interlocutor to do something). The different approaches that have been proposed and can be found in the literature for dialogue management can be generally classified according to two related dimensions: the dialogue control strategy and the dialogue control model.

2.3.1. Dialogue Control Strategies

Dialogue control strategies consider the extent to which the user and/or the system maintains the initiative in a dialogue. Three possible strategies can be implemented. In a dialogue based on system-initiative the dialogue strategy is system-led and the system asks a sequence of questions to gather information from the user. The user can only perform question answering in response to system-initiated speech acts. In a dialogue based on user-initiative the dialogue strategy is user-led and the user can ask questions to the system, which provides question answering capabilities, or issue commands. In a dialogue based on mixed-initiative the previous strategies are combined and both the system and the user can take the initiative: the user can ask questions at any time, but the system can also take control of the dialogue to gather information from the user.

2.3.2. Dialogue Control Models

System-initiative and mixed-initiative strategies require the system to rely on a dialogue control model in order to manage the dialogue flow. Main dialogue control and management approaches include finite state based models, frame-based approaches, information state models and agent-based approaches [27, 28, 29].

2.3.2.1. Finite State-Based Models

In a finite state-based dialogue model a dialogue is represented as a finite state transition system, where the nodes represent the dialogue state and correspond to system's utterances or actions, while the transitions correspond to user's utterances. Possible interactions are thus defined by possible paths through the graph. Small-sized finite-state automata for dialogue control are easy to define and the corresponding dialogues result in a predictable system behavior. These approaches are thus well-suited for representing dialogue flows that correspond to well-structured, pre-definable task-oriented interactions. However, state based models lack in flexibility and naturalness, and the overall dialogue strategy is mainly driven by system-initiative. The dialogue is basically defined as a scripted interaction, with limited support for deviations.

2.3.2.2. Frame-based Models

In frame-based dialogue control strategies, the dialogue state is represented as a frame with slots that have to be filled by the user's answers. The basic idea of frames with slots to be filled is analogous to the concept of a form with fields to be filled. Basically, for each slot in a frame the system ask a question to gather information from the user and fill the corresponding slot, although the information provided by the user with a utterance may contribute to filling more than one slot, as in the case of over-informative answers. The goal of the system is to gather all the information required to fill the slots of a frame. A frame thus acts as a template that drives the interaction flow. The sequence of questions issued by the system is not predetermined and does not follow a rigid question-answer scheme. Possible dependencies between frame slots can be declaratively expressed with precondition rules or priorities. Depending on the information gathered for each slot, the system may undertake verification or confirmation actions. Frame-based models increase the degree of flexibility with respect to finite state-based models and a mixed-initiative strategy can be supported. Frame-based approaches are well suited for information retrieval tasks, where the system aims at gathering a fixed set of information from the user. Specific dialogue behaviors, such as information verification or confirmation steps, still need to be hard-coded into the dialogue's control structure.

2.3.2.3. Agent-based Approaches

Agent-based approaches include different techniques that rely on artificial intelligence and machine learning techniques for modeling and managing a dialogue as an interactive collaborative process between intelligent agents (basically, the system and the user). Plan-based approaches, for example, adopt planning techniques for dialogue management. The basic assumption is that a dialogue is undertaken to achieve a specific goal and utterances modeled as speech acts can be considered as action operators in a planning domain, in terms of constraints, preconditions and effects. A dialogue strategy can thus be devised by solving a planning problem. Some agent-based approaches for plan-based dialogue management are grounded in the belief-desire-intention (BDI) model. Basically, on the basis of its current beliefs about the domain (which include nested beliefs about shared knowledge) and the discourse obligations it has, the agent selects communicative goals, decides the speech acts to be performed, generates an utterance, analyses the response, and updates its beliefs about the discourse state and its own discourse obligations. Recent research efforts relate dialogue management with partially observable Markov decision processes (POMDPs) [40]. These approaches aim at identifying an optimal or approximated dialogue management policy by relying on a probability distribution over possible dialogue states, and updating the distribution as a result of the observed dialogue behavior (i.e., possible interpretations of linguistic utterances), in an action space given by the set of possible dialogue moves. POMDP-based approaches allow for directly incorporating state uncertainty in the dialogue management strategy and exploit reinforcement learning techniques for devising a suitable policy. Agent-based approaches allow implementing robust dialogue control strategies for flexible mixed-initiative dialogues. However, the planning and reasoning stages required for dialogue management assume a complex representation of the dialogue state, introduce an additional source of complexity and represent computationally intensive tasks that may contrast with the real-time performance requirements of dialogue systems.

2.3.2.4. Information State Model

The Information State approach [23] defines a general model for dialogue management, on the basis of information state-based theory. The general theory defines the following main components: a description and formal representation of the informational components that collectively define the information state, including both static and dynamic components such as dialogue participants and user models, background knowledge, beliefs, intentions, etc.; a set of dialogue moves that produce updates over the information state; a set of update rules that formalize and control the way the information state is updated and dialogue moves are chosen; update rules basically consist of applicability conditions over the information state and effects that represent changes on the information state when the rule is applied; an update strategy that defines how to select and apply update rules among the applicable rules at each stage of the dialogue. The general model allows for a rich representation of the dialogue state and does not impose specific formalisms or algorithms to be used for the formalization of dialogue moves, update rules and update strategies. Rather than a specific dialogue management approach, it represents a framework that allows dialogue theories to be formalized, implemented and evaluated.

2.4. Natural Language Generation and Text-to-Speech Synthesis

Natural Language Generation (NLG) capabilities are provided by a Natural Language Generator component. It has the goal of transforming the information or decision produced by the dialogue manager into a grammatically and semantically correct textual format to be sent to the speech synthesizer to be spoken to the user. The textual message has to be coherent with the current dialogue status and its construction requires to identify: (i) the information that should be included; (ii) how this information is structured; and (iii) the linguistic realization of the message, i.e., the formulation of the message, including the choices of lexical items and syntactic structures, as well as politeness and formality aspects. Basic approaches for language generation are template-based, i.e., they rely on templates that allow generating different sentence types. A template typically includes fixed parts/words with predefined slots that must be filled/instantiated with data provided by the dialogue manager. More advanced NLG approaches dynamically produce textual messages using explicit domain and language models, such as generation grammars or statistical models [41, 42]. The textual sentences generated by the NLG process represent the input for the Text-to-Speech Synthesis (TTS) component, in charge of transforming the sentences into speech. In the case of speech capabilities based on a restricted dictionary, pre-recorded words or sentences can be concatenated to produce an output utterance. However, modern speech synthesizers allows transforming arbitrary text into speech. Although out of the scope of this discussion, speech synthesis is still a complex task, where speech generation is preceded by a text analysis stage that includes text segmentation and normalization, morphological analysis, and the modeling of continuous speech effects, with the final goal of producing natural-sounding speech.

In the following paragraphs, a literature review was made about the studies, systems and approaches that investigated the role of dialogue systems and conversational agents in the interaction with elderly people with dementia.

3. Methods

The literature review has been performed according to the general guidelines defined in [43]. A keyword-based search strategy was defined to identify relevant studies through Google Scholar and accessing the SpringerLink digital library, the IEEE408 Xplore Digital Library, the ACM Digital Library and PubMed. The search focused on articles published from 2000 to 2017. The articles were evaluated according to the following inclusion criteria: 1) use of conversational agents, dialogue systems or language processing tools for people with cognitive impairments; 2) age ≥ 60 years; 3) diagnosis of dementia according to the criteria of the National Institute on Aging-Alzheimer's Association (NIAAA) [44] and the Diagnostic and Statistical Manual of Mental Disorders - Fifth Edition (DMS-5) [45]; 4) presence of tests or experiments with qualitative or quantitative results. Articles presenting work on socially assistive robots where speech-based interaction is not explicit considered were excluded. Similarly, we excluded the articles presenting conversational agents or dialogue systems not specifically targeted to people with cognitive impairments. As shown in Figure 3, initially 125 articles were identified by the search strategy and 67 articles were then selected after removing duplicates and a preliminary analysis of their titles. We then considered the abstract, introduction and conclusions of the selected articles and 44 of them were excluded accordingly as they did not met the inclusion criteria. 23 articles were further assessed and 11 of them were excluded after an in-depth examination. 12 articles were finally selected for the review (Table 1). During the analysis, the bibliographies of the selected articles were considered to identify additional works potentially excluded by the search strategy.

The results section was divided as shown below:

1) Conversational Agents

2) Socially Assistive Robot

3) Speech Analysis and Classification

4. Results

4.1. Conversational Agents

Heerink et al. present the results of a field experiment conducted with a Wizard-of-Oz approach to investigate the role of perceived social abilities on the acceptance of a communicative robotic interface by elderly users [46]. The study was performed in 2005 with the Philips iCat robot with simulated conversational capabilities. Conversational scripts were developed for the robot to simulate two conditions: a more socially communicative interface (with the robot able to remember and use the interlocutor's name, and configured to look at the user, nod and smile) and a less socially communicative interface. 40 elderly inhabitants (13 male, 27 female) of an eldercare institution in the Netherlands were observed during interaction sessions lasting 5 minutes, and were then interviewed with questions on the perceived social abilities of the robot and technology acceptance. Specifically, the participants were invited to have a conversation with the robot to ask it to perform the three simple tasks (setting an alarm, give directions to the nearest supermarket and giving the weather forecast for the next day). From a general perspective, the experiment was not able to highlight any significant correlation between perceived social abilities and technology acceptance. However, significant differences were observed in the results regarding the acceptance of the robot as a conversational partner, asking the participants whether they felt uncomfortable when talking to a robot. All participants interacting with the robot configured with more socially communicative abilities reported to feel comfortable, while nearly half of the participants interacting with under less socially communicative conditions reported to feel a little or very uncomfortable. In addition, behavior analysis over recorded interactions showed that elders tend to be more expressive when interacting with a more sociable robot. The authors also report that, for several participants, the conversation with the robot went beyond the possible tasks and the presented possible functionalities of the robot. The authors describe the development of an agent system able to serve as a conversation partner for individuals with dementia [47]. The agent is an animated face of a child shown on a computer screen, and it was used to conduct an evaluation experiment with eight subjects with mild Alzheimer disease (two male and six female with and average age of 78.5 years and a mean Mini-Mental State Examination score of 22.2). Participants were asked to reply to the same 15 reminiscent questions under two conditions, i.e., with questions asked by the virtual agent and by a human speech therapist. Each question was preceded by introductory comments and also shown in textual form on the screen. The analysis focused on evaluating the effectiveness of the conversational agent in eliciting information from the subjects, by comparing the length of the utterances (using the syllable number) under the two conditions. Reported results show that the overall number of syllables used by all the participants to reply to the agent was lower (74%) than when replying to the human partner. Although the methodology and results of the study are questionable, the authors claim that conversational agents may represent a practical and valuable alternative when no human conversation partner exists. Relevant observed phenomena include: (i) the fact that some subjects replied to the introductory comment by the agents before the actual question was asked; (ii) the difficulties faced by the speech recognition system to deal with non-speech utterances (such as coughing or sighing) and background noise; (iii) the comment made by one of the participants about the possibility to “talk freely without any hesitation or anxiety” when interacting with the agent. Yaghoubzadeh et al. investigated the role of virtual agents as daily assistants for elderly or cognitively impaired people, also focusing on how to structure and present the dialogue-based conversational behavior of the agent to support a robust and effective interaction with the user [48]. To this end, a prototype of a virtual agent (acting as an assistant in managing appointments on the calendar) was built and evaluated. Speech recognition capabilities are provided by Windows Speech Recognition or Nuance Dragon NaturallySpeaking systems, and natural language understanding capabilities are based on the ability to identify keywords (i.e., keywords spotting) and simple grammatical structures. Dialogue management was implemented on the basis of the Information State

Model described in the previous Section. However, the evaluation was conducted according to a Wizard-of-Oz approach, to evaluate users' interaction with a conversational agent and their reaction to system misunderstandings introduced on purpose. The study was conducted with two user groups, 6 elderly users and 11 cognitively impaired users. Participants were instructed to verbally interact with the system to insert appointments in the calendar and were then interviewed about system usability and acceptability and about desirable features of the virtual agent. During the dialogues, two data confirmation and verification strategies were experimented, with the agent either summarizing an appointment in a single utterance to be confirmed or presenting the information slots one-by-one (date, time, topic, etc.) with sequential confirmations. On the basis of their results, the authors conclude that both user groups were able to handle interaction problems that may occur when interacting with speech-based systems (such as errors in the recognition and understanding process), but the corresponding dialogue strategies were simulated and not really implemented in the system. In addition, the analysis suggests that explicit confirmation dialogues are an important aspect of the dialogue strategy, and better results are achieved when limited amount of information is presented for confirmation.

4.2. Socially Assistive Robot

An analysis of the functional requirements and design principles for socially assistive robots for elderly people with mild cognitive impairments is carried out in the study of Bruno et al. [49]. Starting from the observation that dialogue is one of the most important social interaction abilities for a socially assistive robots, the authors identify the ability to convey information in spoken natural language (i.e., the ability to produce speech) as a key requirement for the robot's interaction system. The analysis considers the role of emotions in social robots, and identifies the need to provide the robot with the ability to perceive, interpret and convey emotions. The authors argue that it is possible to design a robot able to perceive and convey emotions within speech. Their claims draw on existing results that, on the one side, have investigated the role of speech as a valid method to convey emotions and, on the other side, have shown that emotions produce vocal effects which are consistent between speakers.

The combination of a voice-based and graphical interface for supporting the interaction between an assistive robot and elderly people with mild cognitive impairments is discussed in [50]. Starting from the observation that memory deficits affect language comprehension and production, the authors identify the need to design an interface able to deal with potential problems of human speech, including sentence fragments, interruptions and false starts. In order to identify the required features for the vocal and graphical interface, two tests were conducted with a Kompai mobile robot. The implemented dialogue system relies on speech recognition, dialogue management and text-to-speech capabilities augmented with a multimodal approach that allows combining voice and graphical input/output. Specific tests were conducted with 6 patients from a geriatric hospital in Paris to identify the vocabulary and syntactic structure that characterize the utterances of people with mild cognitive impairments. The dialogue system was configured with a vocabulary of about 170 French words and different syntactic models related to the specific tasks supported by the robot. Speech capabilities of the robot were based on the production of short and simple sentences, as a simple syntactic structure avoids a cognitive overload for the user. The patients were asked to perform selected tasks with the robot (such as make an appointment, add/remove items from a shopping list, or asking for information about, for example, date and time or appointments), using a Wizard-of-Oz approach. The sentences uttered by the patients were then manually annotated and compared with the pre-configured vocabulary and syntactic models. The authors report a significant difference between the vocabulary used by elderly and those used by non-elderly adults (and used to configure the dialogue system). Some differences were also observed regarding the syntactic structure of the utterances. The authors highlight the importance of taking into account, in the design process, the peculiarities of the vocabulary and syntactic structure of the target users, as age and cognitive impairments may introduce differences with respect to reference models used by non-elderly adults. In early 2000, the Nursebot project started investigating the development of a personal service robot, Flo, with social interaction abilities and specifically targeted at people with mild forms of dementia and cognitive impairment [51]. Spoken interaction capabilities were considered as an essential requirement to enable a natural interaction between the users and the robot. Flo was equipped with a speech recognition interface configured with a vocabulary of approximately one hundred words and controlled by a dialogue system. To deal with simple domains, language understanding was based on keyword spotting techniques. Although advanced NLP techniques were considered as not necessary, the dialogue manager was able to exploit external information sources and connect to the World Wide Web. Flo was superseded by the Pearl robot and some experiments were conducted in an assisted living facility, as reported in [52]. In Pearl software components were implemented relying on probabilistic techniques to take into account uncertainty factors. Regarding speech capabilities, this was required due to the difficulties observed in people with cognitive deficiencies in articulating computer-understandable responses.

4.3. Speech Analysis and Classification

Techniques based on partially observable Markov decision processes (POMDPs) were used for robot control and dialogue management. Among the experiments, the speech interface was compared with a system unable to deal with uncertainty. The average time to task completion, the average number of errors, and the average user-assigned reward were measured in the presence of good, average and poor speech recognition conditions. Reported results show that, especially in the case of poor speech recognition, the POMDP-based approach exhibits a greater time to task completion due to the generation of clarification questions, but the corresponding error rate is much lower. The authors generally identify probabilistic models and techniques that cope with uncertainty as a viable solution for building robust assistive robotic solutions for people with mild forms of dementia and cognitive impairment. The authors investigate the possibility to assess the cognitive status of people with dementia through conversations, focusing on patients' communication responsiveness [53, 54, 55]. To this end, a prototype of a virtual listener agent (shown on a computer screen) was developed, and conversations between the agent and 10 people with dementia (2 male and 8 female) were recorded and evaluated. The conversational agent is able to sequentially ask the user a set of questions regarding, for example, the patient's physical conditions or related to reminiscence and patient's memories. Understanding capabilities are based on a combination of keyword spotting and a similarity-based method to compare keywords with a list of possible utterances to be recognized. While interacting with the user, the agent records the number of utterances, their duration and the pitch. User's behavior in providing each response was then manually classified as high responsive (HR) and low responsive (LR). For these two classification groups, the following metrics were evaluated: (i) the pause between the end of the agent's question and the start of the user's answer; (ii) the pitch of the user's response; (iii) the average utterance duration; and (iv) the average frequency of head nods per response. The authors report that higher responsiveness correlates with shorter pause, higher pitch, longer utterance, and more frequent head nods. They thus claim that patient's responsiveness in the conversation with an agent, as an indicator of patient's cognitive status, can be assessed using the aforementioned metrics. Roark et al. have recently investigated the application of natural language processing techniques to analyze spoken responses produced by subjects during neuropsychological exams [56]. The study aims at identifying diagnostic markers in spoken language that allow discriminating between healthy elderly subjects and subjects with mild cognitive impairment. To this end, the authors evaluate the role of different linguistic complexity measures and pause statistics to derive early markers of Alzheimer disease. Audio recordings were collected from 74 neuropsychological examinations, partitioned into 37 healthy subjects and 37 subjects with mild cognitive impairment. The recordings were manually transcribed and annotated with syntactic annotations. In a first stage, multiple linguistic complexity measures and other spoken language derived measures were evaluated, covering both measures derived from the syntactic structure of the utterances and measures derived from temporal aspects of the speech samples (e.g., pauses and their duration). In a second stage, automated NLP techniques were used for parsing and processing transcripts and audio recording. The study highlights a high correlation between the measures derived from automatic processing and the manually derived measures. According to the authors, this correlation indicates that the automatically derived measure are able to preserve the discriminative utility of manually derived measures in differentiating between healthy subjects and subjects with mild cognitive impairment. Presented results show the potential of applying NLP and speech processing techniques to spoken language samples, in order to automatically derive measures for discriminating between healthy subjects and those with mild cognitive impairment. The authors propose an approach for classifying dialogue acts in spontaneous speech of the elderly with mild dementia [57]. The goal is to provide a dialogue system with the ability to classify dialogue acts and assign functional tags to user's utterances, to enable the understanding of the communicative intentions of the utterances and produce an appropriate response in the context of a conversation. Rather than undertaking a language modeling approach, considered by the authors as too expensive for spontaneous speech, the proposed solution introduces a sub-lexical dialogue act classifier that processes sequences of phonemes provided by a phoneme recognizer. The study builds on the hypothesis that sub-sequences of phonemes should be effective indicators for identifying dialogue acts, in the same way as word n-grams can be used for this task. The analysis and classification is performed by a sub-lexical classifier based on a support vector machine (SVM). An empirical study was conducted to compare the sub-lexical dialogue act classifier with a classifier based on a hidden Markov model (HMM). The approaches were compared on a dialogue corpus containing 4080 user utterances derived from the interactions between 20 participants (3 male and 17 female, with an age ranging from 67 to 97 and a mini mental state examination score ranging from 9 to 30) and the system. 12 dialogue acts were identified for representing the communicative intention of the utterances (e.g., questions, confirmations, affirmative answers, etc.). Both the classification approaches were evaluated and compared in term of accuracy. The sub-lexical SVM-based classifier was also compared with a SVM-based classifier operating on lexical features. Presented results show that the sub-lexical classifier is robust against the poor language modeling of language and it performed better than the HMM-based lexical classifier, whose performances decreases when the mismatch between the language model and the actual corpus increases.

5. Conclusion

Speech is largely considered as the most powerful and effective communication mode for an assistive social robot to interact with its users. Recent technological developments and research results are contributing to solving the challenges that characterize the design and implementation of spoken dialogue systems for human-robot interaction. In particular, the increasing accuracy of speech recognizers and the advances in the field of natural language processing provide viable solutions for the adoption of dialogue systems in different domains. In addition, probabilistic models and machine learning techniques are increasingly used to deal with the uncertainty factors that characterize language processing and dialogue management. We have provided an overview of the technologies, approaches, methods and general issues that come into play in the complex process of building a dialogue system. Each functional component in the processing chain plays a fundamental role in the overall system performance and has to be carefully designed. However, the approach and techniques to be adopted for each stage and their effectiveness largely depend on the target domain, the intended users and the specific tasks that the system has to support. For example, the usage of a restricted vocabulary, keyword-spotting techniques and simple scripted dialogue management strategies can be effective in supporting predefined tasks and dialogues where the interaction is mainly driven by the system (e.g., for eliciting specific information from the user through a set of questions). While advanced stochastic techniques and deep NLP would be an overkill in these settings, they improve the ability of the system to deal with uncertainty and open-ended dialogues with a mixed-initiative interaction pattern. The presented analysis supports the communication between the elderly with cognitive impairments and a social robot, and introduces additional issues. The ability of a dialogue system to respond to user needs affects the perceived social abilities of the robot and has thus an impact on the acceptance by the users. Speech disfluencies, such as repetitions and incomplete words or sentences observed from the analysis of conversational dialogs, pose challenges to speech recognition and language processing capabilities. In addition, some authors have observed that age and cognitive impairments produce modifications in the vocabulary and syntactic structures used by the elderly or people with dementia, resulting in the need for adaptive techniques able to take these changes into account. Similarly, it has been observed that the communication abilities of the system have to be adapted to the target users to improve the understanding and reduce the cognitive load. This includes, for example, a relatively slow speech rate, closed-ended questions, repetitions and small verification questions, a reduced syntactic complexity and sentences with few clauses. Most of the experiments considered in the analysis were conducted with using a Wizard-of-Oz approach, and thus aimed at deriving requirements and evaluating the potential feasibility and acceptability of vocal interfaces and conversational agents, rather than at evaluating the actual capabilities of dialogue systems in real-world scenarios. Other authors have investigated speech processing and NLP capabilities form a different perspective. They focused on the possibility to assess the cognitive status of people with dementia by identifying and analyzing metrics and diagnostic markers able to serve as indicators for evaluating patient's cognitive status and discriminating between healthy subjects and people with cognitive impairment. Reported results show that these approaches have a potential and their integration in a dialogue system deserves further investigation, along with emerging NLP techniques that focus on sentiment analysis and the identification of the emotional state of the speaker. Similarly, the ability of a social robot to convey emotions through the produced speech is still an open research challenge, where further efforts are necessary with multidisciplinary contributions towards socially assistive robots.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The research leading to these results has received funding from the European Union Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) under grant agreement 643808 Project MARIO "Managing active and healthy aging with use of caring service robots".

Table 1. Conversational agents, dialogue systems and NLP techniques for patients with dementia.

Studies	**Ref.**	**Methods**	**Outcomes**
Heerink *et al.*, 2006	[46]	CRI	Acceptance of the robot as a conversational partner increases if the robot exhibits socially communicative abilities.
Yasuda *et al.*, 2013	[47]	VCA	Conversational agent practical and valuable alternative to humans for reminiscence; issues in speech recognition for non-speech utterances.
Yaghoubzadeh *et al.*, 2013	[48]	VCA	Importance of explicit confirmation dialogues as part of the dialogue strategy; better results achieved when limited amount of information is presented for confirmation.
Bruno *et al.*, 2013	[49]	SAR	Need to provide a socially assistive robot with the ability to perceive, interpret and convey emotions through speech.
Granata *et al.*, 2010	[50]	SAR - DS	Take into account the difference between the vocabulary and syntactic structures used by elderly and non-elderly adults.
Roy *et al.*, 2000	[51]	SAR - DS	Keyword spotting techniques considered sufficient to provide support for simple tasks for people with mild forms of dementia.

Montemerlo *et al.*, 2002	[52]	SAR - DS	Partially observable Markov decision processes (POMDPs) as a robust solution to cope with uncertainty and poor speech recognition.
Nonaka et al., 2012 Huang et al., 2012 Sakai et al., 2012	[53-55]	VCA - SA	Pause duration, utterance duration, pitch and frequency of head nods related to patient's responsiveness in the conversation, as indicators of patient’s cognitive status.
Roark et al., 2011	[56]	NLP - SA	NLP and speech processing techniques able to automatically derive measures and diagnostic markers for discriminating between healthy subjects and those with mild cognitive impairment.
Sadohara *et al.*, 2013	[57]	DS - DAC	Machine learning techniques for a sub-lexical classifier for dialogue acts using sequences of phonemes.

Legend

SAR: Socially Assistive Robot; **CRI:** Communicative Robotic Interface; **VCA:** Virtual Conversational Agent; **DS:** Dialogue System; **SA:** Speech Analysis; **NLP:** Natural Language Processing; **DAC:** Dialogue Act Classification

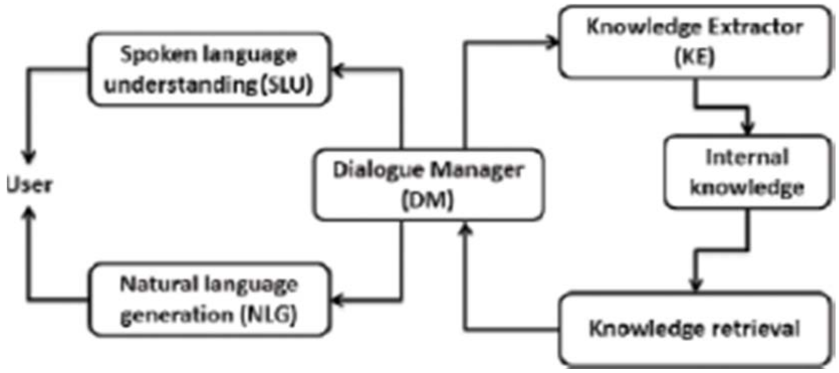

Figure 1. General architecture of a dialogue system

Rejuvenation Research
Dialogue Systems and Conversational Agents for Patients with Dementia: the human-robot interaction (DOI: 10.1089/rej.2018.2075)
This paper has been peer-reviewed and accepted for publication, but has yet to undergo copyediting and proof correction. The final published version may differ from this proof.

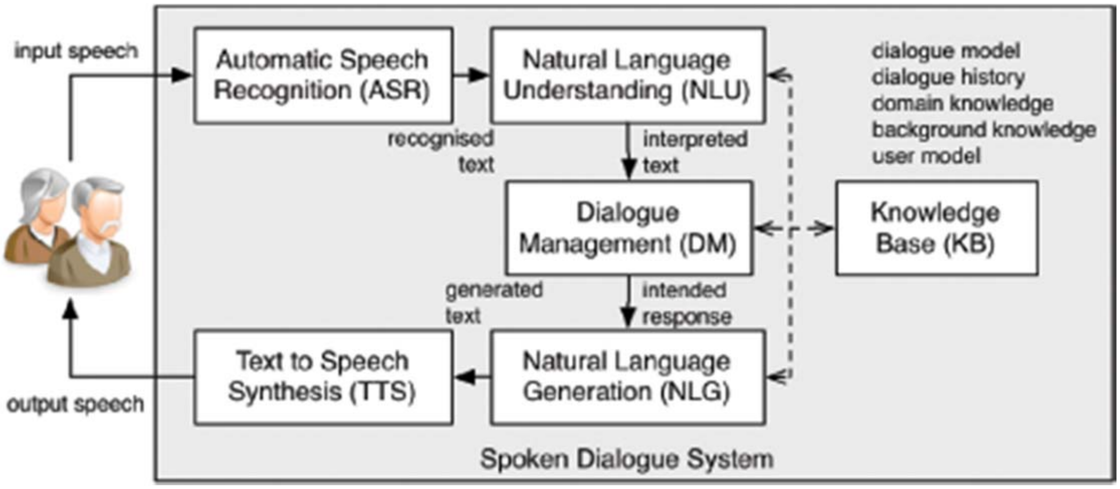

Figure 2. Functional architecture of a Spoken Dialogue System

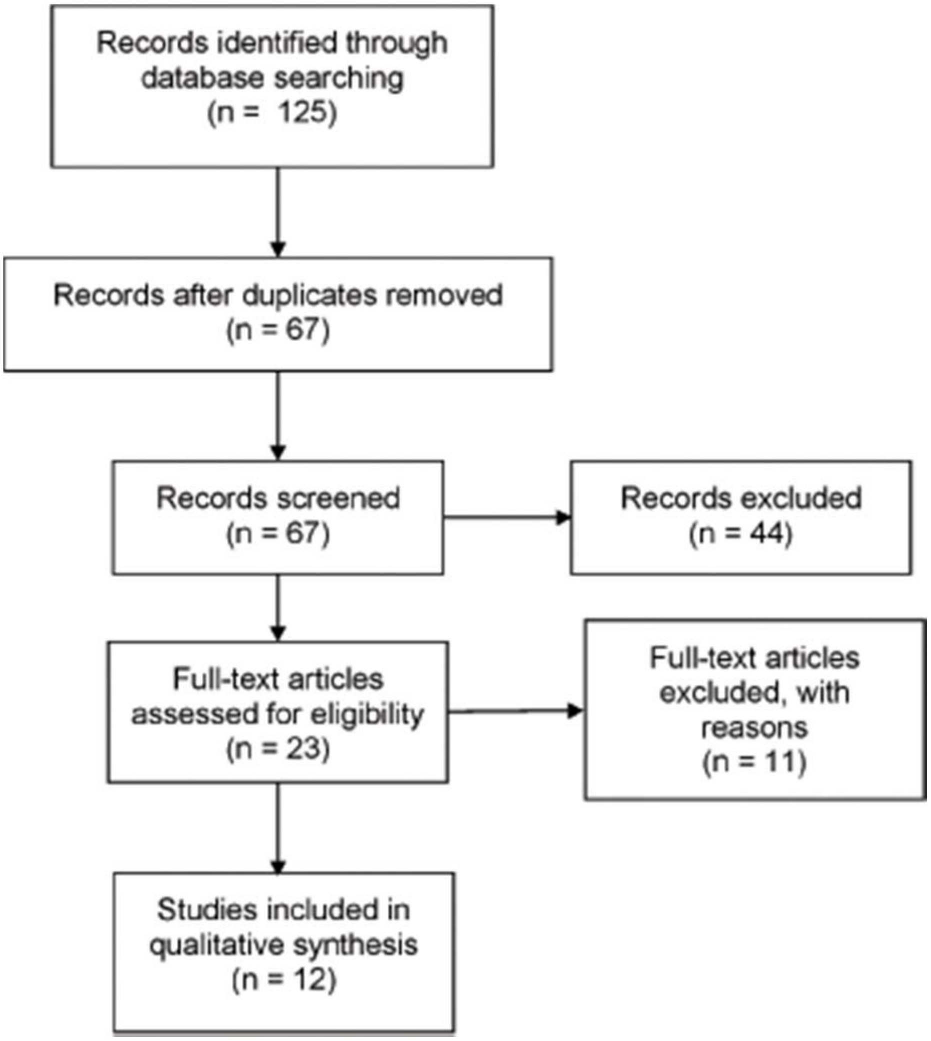
Records identified through database searching (n = 125)
Records after duplicates removed (n = 67)
Records screened (n = 67)
Records excluded (n = 44)
Full-text articles assessed for eligibility (n = 23)
Full-text articles excluded, with reasons (n = 11)
Studies included in qualitative synthesis (n = 12)

Figure 3. Flow diagram outlining the selection procedure to identify articles which were included in the analysis of the state of the art in speech-based robot socialization

References

1. World Health Organization. Dementia: Updated December 2017. http://www.who.int/mediacentre/factsheets/fs362/en/. Accessed February 06, 2018.

2. Fong T, Nourbakhsh I, Dautenhahn K. A survey of socially interactive robots. *Rob Auton Syst* 42(3-4), 143-166 (2003).

3. Mavridis N. A review of verbal and non-verbal human-robot interactive communication. *Rob Auton Syst* 63(P1), 22-35 (2015).

4. Hirschberg J, Manning CD. Advances in natural language processing. *Science* 349(6245), 261-266 (2015).

5. McTear MF. Spoken Dialogue Technology: Towards the Conversational User Interface. Springer, London (2004).

6. Mariani J, Rosset S, Garnier-Rizet M, *et al*. Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems Into Practice. Springer, New York (2013).

7. Möller S. Quality of Telephone-Based Spoken Dialogue Systems, 1st ed. Springer, New York (2010).

8. Zue VW, Glass JR. Conversational interfaces: advances and challenges. *Proceedings of the IEEE* 88(8), 1166-1180 (2000).

9. Gorin AL, Riccardi G, Wright JH. How May I Help You? *Speech Commun* 23(1-2), 113-127 (1997).

10. Zue V, Sene S, Glass JR, *et al*. JUPITER: a telephone-based conversational interface for weather information. *IEEE Trans Speech Audio Process* 8(1), 85-96 (2000).

11. Xu W, Rudnicky AI. Task-based dialog management using an agenda. In: Proceedings of the 2000 ANLP/NAACL Workshop on Conversational Systems - Volume 3, pp. 42-47 (2000).

12. Kluwer T. From chatbots to dialog systems. In: Perez-Marin D, Pascual-Nieto I (eds.) Conversational Agents and Natural Language Interaction: Techniques and Effective Practices, pp. 1-22. IGI Global, Hershey, PA 17033, USA (2011).

13. Dingli A, Scerri D. Building a Hybrid: Chatterbot - Dialog System. In: Text, Speech, and Dialogue, pp. 145-152. Springer, Berlin Heidelberg (2013).

14. Weizenbaum J. ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun ACM* 9(1), 36-45 (1966).

15. Colby KM, Weber S, Hilf FD. Artificial paranoia. *Artif Intell* 2(1), 1-25 (1971).

16. Mauldin ML. ChatterBots, TinyMuds, and the Turing Test: Entering the Loebner Prize Competition. In: Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1), pp. 16-21. AAAI, Menlo Park, CA, USA (1994).

17. Wallace RS. The Anatomy of A.L.I.C.E. In: Epstein, R., Roberts, G., Beber, G. (eds.) Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer, pp. 181-210. Springer, Dordrecht (2009).

18. AIML: Artificial Intelligence Markup Language. http://www.alicebot.org/aiml.html

19. http://www.cleverbot.com/

20. Huang J, Li Q, Xue Y, *et al*. TeenChat: A Chatterbot System for Sensing and Releasing Adolescents' Stress. In: Proceedings of the 4th International Conference on Health Information Science, pp. 133-145. Springer, Cham (2015).

21. Ireland D, Liddle J, McBride S, *et al*. Chat-Bots for People with Parkinson's Disease: Science Fiction or Reality? *Stud Health Technol Inform* 214, 128-133 (2015).

22. https://github.com/bwilcox-1234/ChatScript

23. Traum DR, Larsson S. The Information State Approach to Dialogue Management. In: van Kuppevelt J, Smith RW (eds.) Current and New Directions in Discourse and Dialogue, pp. 325-353. Springer, Dordrecht (2003).

24. Larsson S, Traum DR. Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit. *Nat Lang Eng* 6(3-4), 323-340 (2000).

25. Peltason J, Wrede B. Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns. In: Proceedings of the SIGDIAL 2010 Conference. Association for Computational Linguistics, Stroudsburg PA 18360, USA (2010).

26. Berg MM. NADIA: A Simplied Approach Towards the Development of Natural Dialogue Systems. In: Natural Language Processing and Information Systems, pp. 144-150. Springer, Cham, Switzerland (2015).

27. López-Cózar R, Callejas Z, Griol D, *et al*. Review of spoken dialogue systems. *Loquens* 1(2) (2014).

28. McTear MF. Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Comput Surv* 34(1), 90-169 (2002).

29. Zue V, Sene S. Spoken Dialogue Systems. In: Benesty J, Sondhi MM, Huang YA (eds.) Springer Handbook of Speech Processing, pp. 705-722. Springer, Berlin Heidelberg (2008).

30. Gales M, Young S. The Application of Hidden Markov Models in Speech Recognition. *Found Trends Signal Process* 1(3), 195-304 (2007).

31. Bastianelli E, Castellucci G, Croce D, *et al*. Effective and Robust Natural Language Understanding for Human-robot Interaction. In: Proceedings of the Twenty-first European

Conference on Artificial Intelligence, pp. 57-62. IOS Press, Amsterdam, The Netherlands (2014).

32. Deng L, Li X. Machine Learning Paradigms for Speech Recognition: An Overview. *Trans Audio Speech and Lang Proc* 21(5), 1060-1089 (2013).

33. Yu D, Deng L. Automatic Speech Recognition: A Deep Learning Approach. Springer, New York (2014).

34. Roukos S. Natural language understanding. In: Benesty J, Sondhi MM, Huang YA (eds.) Springer Handbook of Speech Processing, pp. 617-626. Springer, Berlin Heidelberg (2008).

35. https://stanfordnlp.github.io/CoreNLP/

36. Fillmore CJ. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280(1), 20-32 (1976).

37. Vacher M, Aman F, Rossato S, *et al*. Development of Automatic Speech Recognition Techniques for Elderly Home Support: Applications and Challenges. In: Proceedings of the First International Conference on Human Aspects of IT for the Aged Population, pp. 341-353. Springer, Cham, Switzerland (2015).

38. Young V, Mihailidis A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review. *Assist Technol* 22(2), 99-112 (2010).

39. Bunt H, Petukhova V, Traum D, *et al*. Dialogue Act Annotation with the ISO 24617-2 Standard. In: Dahl DA (ed.) Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything, pp. 109-135. Springer, Cham, Switzerland (2017).

40. Young S, Gasic M, Thomson B, *et al*. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE* 101(5), 1160-1179 (2013).

41. Rieser V, Lemon O, Keizer S. Natural Language Generation As Incremental Planning Under Uncertainty: Adaptive Information Presentation for Statistical Dialogue Systems. *IEEE/ACM Trans Audio Speech and Lang Proc* 22(5), 979-994 (2014).

42. Shannon M, Zen H, Byrne W. Autoregressive Models for Statistical Parametric Speech Synthesis. *IEEE/ACM Trans Audio Speech and Lang Proc* 21(3), 587-597 (2013).

43. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. Technical report, EBSE (2007).

44. McKhann GM, Knopman DS, Chertkow H, *et al*. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guide-lines for Alzheimer's disease. *Alzheimers Dement* 7, 263–269 (2011).

45. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Association: Washington (2013).

46. Heerink M, Krose B, Evers V, *et al*. The Influence of a Robot's Social Abilities on Acceptance by Elderly Users. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication, pp. 521-526 (2006).

47. Yasuda K, Aoe J, Fuketa M. Development of an agent system for conversing with individuals with dementia. In: Proceedings of the 27th Annual Conference of the Japanese Society for Artificial Intelligence (2013).

48. Yaghoubzadeh R, Kramer M, Pitsch K, *et al*. Virtual Agents as Daily Assistants for Elderly or Cognitively Impaired People. In: Aylett R, Krenn B, Pelachaud C, Shimodaira H (eds.) Proceedings of the 13th International Conference on Intelligent Virtual Agents, pp. 79-91. Springer, Berlin Heidelberg (2013).

49. Bruno B, Mastrogiovanni F, Sgorbissa A. Functional requirements and design issues for a socially assistive robot for elderly people with mild cognitive impairments. In: 2013 IEEE RO-MAN, pp. 768-773 (2013).

50. Granata C, Chetouani M, Tapus A, *et al*. Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders. In: 19th International Symposium in Robot and Human Interactive Communication, pp. 785-790 (2010).

51. Roy N, Baltus G, Fox D, *et al*. Towards Personal Service Robots for the Elderly. In: Workshop on Interactive Robots and Entertainment (2000).

52. Montemerlo M, Pineau J, Roy N, *et al*. Experiences with a Mobile Robotic Guide for the Elderly. In: Eighteenth National Conference on Artificial Intelligence, pp. 587-592. American Association for Artificial Intelligence, Menlo Park, CA, USA (2002).

53. Nonaka Y, Sakai Y, Yasuda K, *et al*. Towards Assessing the Communication Responsiveness of People with Dementia. In: Proceedings of the 12th International Conference on Intelligent Virtual Agents, pp. 496-498. Springer, Berlin Heidelberg (2012).

54. Huang HH, Matsushita H, Kawagoe K, *et al*. Toward a memory assistant companion for the individuals with mild memory impairment. In: 2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing, pp. 295-299 (2012).

55. Sakai Y, Nonaka Y, Yasuda K, *et al*. Listener agent for elderly people with dementia. In: 2012 7th ACM/IEEE International Conference on Human-Robot Interaction, pp. 199-200 (2012).

56. Roark B, Mitchell M, Hosom JP, *et al*. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE Transactions on Audio, Speech, and Language Processing (TASL)* 19(7), 2081-2090 (2011).

57. Sadohara K, Kojima H, Narita T, et al. Sub-lexical Dialogue Act Classification in a Spoken Dialogue System Support for the Elderly with Cognitive Disabilities. 4th Workshop on Speech and Language Processing for Assistive Technologies,2013.