PART 1

Historical and Methodological Landmarks

COPYRIGHTED MATERIAL

Chapter 1

An Assessment of the Evolution of Research and Systems

Man–machine dialogue (MMD) systems appear more present in the works of science-fiction than in reality. How many movies do we know which show computers, robots, or even fridges and toys for children who can talk and understand what they are told? The reality is more complex: some products that have come from new technologies, such as cell phones or robot companions, talk and understand a few words, but they are far from the natural dialogue which science-fiction has been promising for years.

The ideas for application are not lacking. Implementing a dialogue with a machine could be useful for getting targeted information, and this could be for any type of information: transportation [LAM 00], various stores, tourist or recreational activities [SIN 02], library collections, financial administrative procedures [COH 04], etc., see [GAR 02] and [GRA 05]. The dialogue is indeed adapted to the step-by-step elaboration of a request, a request that would be difficult to hold in a single utterance or in a command expressed in a computer language. The first field of application of MMD that includes question–answering systems (QAS) is sometimes defined as *information-seeking dialogue*. When the dialogue only concerns a single topic, for example railway information, we talk of closed-domain dialogue. When the dialogue can be about pretty much anything, for example the questioning of the encyclopedic database as IBM Watson recently did with a TV show task, we talk of open-domain dialogue [ROS 08]. If we reuse the example of the

introduction, a unique utterance with no dialogue could be as follows: "I would like to book a single journey to Paris taking the shortest itinerary as long as it takes less than half an hour (otherwise I do not wish to make a reservation)". The elaboration of a natural dialogue is much more flexible: it allows the user to express a first simple request and then improve it according to the machine's answer; it allows the machine to transfer information for a future action, and confirm or negate along the way [PIE 87]. The total number of words to arrive at the same result might be greater, but the spontaneity of the utterances and their speed as well as the ease of production is more than fair compensation. The example of questioning a yellow-page-style directory, [LUZ 95] shows another advantage of dialogue: the user can obtain the address of a taxidermist even when he/she does not know the name of this profession. Through the conversation, the dialogue, the user gets the machine to understand exactly what he/she is looking for. There is a joint construction of a common concept to both interlocutors, and this joint construction is the point of the dialogue compared to the unique utterance or the computer language request.

Beyond the information request or the consultation of a database, installing a dialogue with a machine can also be useful to manage a computer system, for example digital design software (drawing, image processing, three dimensional (3D)) or simply a computer's operating system. We can also imagine that instead of looking for the accurate function in the numerous menus and submenus of the software in question, the user carries out vocal commands that are much swifter and more direct, at least if he/she is not familiar with the software. This second field of application of MMD is close to that of man–machine interfaces (MMI), and is sometimes defined as *control command dialogue*. Including the computer science software development, we can almost imagine a user who would use the language to *program in natural language* [LUZ 95]. Including robotics, this is the field of robot command, the key application of modern artificial intelligence (AI) [GOR 11]. Moreover, it is also the field of professional, civil and military systems whose design I took part in at Thales: air traffic control and management, maritime surveillance, supervision of the situation on the field and system command in dangerous areas. These systems are currently complex MMI and the research team's work in which I participated was to test the potential of *giving them speech.* We remained there in the closed-domain control command dialogue, but with many robustness limits.

Information dialogue, control command dialogue: all the existing systems will not fall into one or the other of these strict categories. Some systems

allow for both types of interaction, for example some companion robots can both give their users information and carry out simple tasks on demand, such as walking or dancing. Other systems do not aim to provide information or carry out specific tasks. These are, for example, purely recreational systems, with the example of conversation robots on the Internet. The examples given here are taken from general public MMD systems, or at least systems that are destined to be such. But do they really work? In fact, what we can note when using this type of system is that it is quite difficult to establish a proper dialogue. When a word is recognized and understood, which is not systematic, the machine tries to give an answer based on this word or attempts to restart the dialogue in its own way, which is rarely a relevant one. As Vilnat [VIL 05, p. 5] states, the MMD systems only work in a very imperfect manner and thus are greatly criticized, up to the point where "it will never work" is often heard. The criticism comes, first and foremost, from the users who notice that there is a wide gap between what they test and what they hope for, and they often believe that a classic MMI is quicker, more efficient and even easier, or less confusing, to use. The criticism also comes from researchers and developers in the MMD field. Indeed, the amount of work required to achieve a system is such that there is a lot of discouragement. The amount of work corresponding to a doctoral dissertation is not sufficient, at least when trying to achieve an innovative system. As an example, Guibert's [GUI 10, p. 60] discouragement when designing a system called *A* is striking: "following the termination of the development of this system A, taken as an example among others, this body of work is actually the chronicles of the foretold failure of current dialogue systems".

We will see that when the dialogue is directed by a clearly defined task, it is possible to design a performing MMD and this design has actually greatly progressed in the past few years. After discussing a few historical landmarks (section 1.1), we will quickly cover the functionalities that are more and more present on current systems (section 1.2), and from this we will deduce a primary list of potential challenges for the years to come (section 1.3).

1.1. A few essential historical landmarks

The dialogue between human being and machine is a key field in computer science: a kind of quest for the Holy Grail, which was the source of computer science developments and researcher vocations. As it so happens, the first system to become a landmark, ELIZA [WEI 66], is also a huge subterfuge

(which was assumed, as we will see in section 1.1.1). Various paths were then taken in serious MMD system design: a path close to AI, with a focus on interpretation and reasoning issues, and a path that consisted of enriching the automatic speech recognition systems. Both paths with their two separate communities [VIL 05, p. 47], have recently come together again and allowed various consistent MMD systems to reach fruition. These are the systems we will now present.

1.1.1. *First motivations, first written systems*

Can a machine think? In 1950, Alan Turing relaunched this question that had recurred throughout technology's history: he substituted the question "can a machine imitate a human?" and suggested a game, or test, based on imitation, which became famous by the name of the Turing test. At first, the imitation concerns a man and a woman: the test subject talks with a man and a woman in turn, through machine-typed pieces of paper, without seeing or knowing anything about his/her successive interlocutors. The man has to try and pass for a woman, and the subject thus has to guess which one is the man and which one is the woman. Then, without the subject's knowledge, the man is replaced by a machine. If the subject cannot identify either of the interlocutors, then the machine passed the Turing test. This game, created at a time when it was impossible to program an MMD system, was the source of innumerable discussions, of various and varied assertions on the nature of the machine or of the human being. The interesting thing here is the challenge for computer science: to program an MMD system that can be thought to be a human being. Turing does not give us many hints as to how to achieve that result. The description of the test is focused on experimental conditions and does not address the importance of language and dialogue in this approach to thought [TEL 09]. Nonetheless, there are competitions organized today (such as the Loebner prize) inspired by the Turing test. The 1950s correspond to the first research motivations for MMD, information seeking and NLP. We should point out that the Atala association, created in 1959, was originally called *Association for the study and development of automatic translation and applied linguistics* ("Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée" in French), and then became the *Association for Natural Language Processing* ("Association pour le Traitement Automatique des Langues").

The 1960s mark the appearance of the first MMD systems. ELIZA[1] [WEI 66], which we mentioned earlier, is fascinating in more than one way. First of all, this is a written dialogue system that really works without looping or randomly stopping. It is always possible to carry out conversations on hundreds of speaking turns. Moreover, the chosen task is itself fascinating: the system is supposed to play the role of a non-directing psychotherapist, which means it simply listens to the speaker to tell it about his/her problems ("I have a problem with my parents") and sometimes reacts to certain sentence ("tell me about your family"). The realism is so strong that some users have spent hours talking with ELIZA, and J. Weizenbaum had to decide against openly adding a dialogue-saving module, faced with accusations of spying and violating privacy. This task has two advantages: it does not have to carry out a complex dialogue, for example with negotiation or argumentation, while keeping a spontaneous and natural aspect, since the user can say what he/she wants when he/she wants to; and on the other hand, it is easy to program, since the system does not need to understand absolutely everything: utterances such as "what makes you say that?" or "I see, please go on" are vastly sufficient. Indeed, and this is fascinating for NLP, AI or MMD researchers, J. Weizenbaum managed to develop a system that appears to master language and pass the Turing test, whereas it does not even approach the most basic issues of automatic understanding.

Indeed, all of ELIZA's operating relies on a few well-chosen heuristic rules. The system knows a few words, especially those linked to family: "parents", "mother" and "father". It is thus able to bounce off the utterance "I have a problem with my parents" without any understanding involved in this process: the system just detected "parents" and answered with a new question on "family", a new question that actually allows it not to have to take into account the meaning of the user's utterance. The system also knows the personal pronouns referring to the two interlocutors, "I", "me", "my", "you" and "your" that allows it to carry out replacements and build an utterance taking up parts of the user's utterance, such as "what makes you believe that

1 The name came from the ELIZA Doolittle character in the movie *My Fair Lady* (1964, G. Cukor), itself an adaptation of the play *Pygmalion* (1914, G.B. Shaw), which has also been adapted for the movies. ELIZA Doolittle is a florist from a very poor neighborhood, and becomes the subject of a bet when an aristocrat claims that by changing her manner of speech, he will be able to make her pass for an aristocrat herself. The idea of duping someone through language and dialogue is thus the origin of the system's name.

you are listening to *my* advice", generated after "*I* am listening to *your* advice". With this example, we can note that the system does not understand much, but it is able to switch the persons around and frame the user's utterance in a question "what makes you believe that", a deliberately open question. The techniques implemented by the input utterances are word sequence detection and keyword detection. Those implemented for the system's output utterance generation are the direct production of typical sentences, the concatenation of text span, whether they are typical spans or spans obtained through a user's utterance. The system also has the beginning of memory, inasmuch as it is able to return to a familiar term used a few speaking turns prior.

A few years after ELIZA, the PARRY [COL 71] system had an impact due to its supplementary techniques. This time the machine simulates a paranoid subject during his first (written) interview with the user who is supposed to play the role of a psychiatrist, a profession to which the main author incidentally belongs. The claimed scientific approach is the studying and modeling of paranoia, and this goes so far as the funding that comes in part from the National Institute of Mental Health, and the methodology that includes not only the modeling and computer science development of the model, but also its assessment by mental health professionals: a total of 25 psychiatrists were involved, and the overwhelming majority of them (23) diagnosed the system as paranoid, making it pass the Turing test with flying colors. The dialogues are carried out as interviews and start with the factual questions that the user asked the system: name, age, occupation. Thus, PARRY has in his memory a set of answers to these typical questions: his name is Frank Smith, he is 28 years old, and interned in a hospital. He also has in his memory various questions that the system can ask, thus inverting the dialogue orientation: "who are you?", "what do you want with me?", as well as anecdotes, and especially words around a relatively well-elaborated concept, such as that of *mafia.* The techniques implemented are also techniques of text span research, keyword detection, first and second person pronoun management, but all with more finesse than ELIZA had. For example, the word "fear" has a set of predefined spans, and verbs such as "to believe" have specific processes. Moreover, the system is characterized by an attempt at personality or *mental states* through variables: fear, anger and distrust. The values of these variables increase or decrease as the dialogue unfolds, according to what the user says. The system's behavior evolves in a

consequent manner: it becomes aggressive if the anger value passes a certain threshold. The rules or heuristics, on the contrary from ELIZA's rules, are based both on the user's utterances and on the variables of state. PARRY marks an evolution of MMD systems, with the technical means of the time: the program, written in a variant of the Lisp language, takes 35 Kb of which 14 Kb belong to the database.

The 1970s were the time of the first (written) understanding systems, with significant improvements in NLP, especially in syntactic and semantic analyses and, thus, the first true systems of MMD written that model a field of knowledge, know how to interpret an utterance in this field, and start to manage a structure dialogue. This progress follows a few landmark works in linguistics and computer linguistics, especially B.J. Grosz and then C.L. Sidner; as Jurafsky and Martin [JUR 09, p. 892] underline it. That corresponds to the first path mentioned on page 6, with two key systems, SHRDLU and genial understanding system (GUS). In parallel, the speech recognition system path is also progressing strongly, especially with systems developed within the American Advanced Research Projects Agency (ARPA) projects: Harpy, Hearsay, Hwim. We thus go from the recognition of isolated words, which is not at all adapted to MMD, to the recognition of continuous, and eventually multi-locutor words, with concerns which start to reach those of MMD, for example the question of software architecture to get various sources of knowledge communicating inside systems, see Pierrel's [PIE 87] historical outline. We will return to this in section 1.1.2 with the first oral MMD systems.

The SHRDLU[2] system [WIN 72] gives a new boost to written MMD by showing the deeper understanding and dialogue possibilities as soon as you limit yourself to a clearly limited and modeled task. This time, let us forget about the Turing test and turn to targeted applications: the task consists of displacing geometrical objects (cubes, cones and pyramids) with a machine. It involves the display of a scenario on a screen, with a representation of the system itself with a kind of robot arm manipulating objects. The user creates utterances such as "pick up a green block" or "find a block that is taller than the one you are holding and put it into the box", and the system carries out

2 The name comes from the sequence of letters E T A O I N S H R D L U that is, in decreasing order, the sequence of letters most often used in English, in the way they are vertically shown in the middle of some printing machine keyboards.

these actions, whose results are visible on screen. This task puts the accent on *object reference* phenomena: what object is referred to by "the pyramid"? To correctly interpret such a reference, the system must find among the objects on display which one is correct, meaning which one corresponds to the user's intent. If two or three pyramids are visible, the system can thus answer "I do not understand which pyramid you mean". After clarification, it does what it must do, that is carry out the actions and answer questions. Many of the possible questions revolve around the physical world of objects: "what does the box contain?" and "what is the pyramid supported by?". Each time, SHRDLU is able to analyze the image, identify the spatial relations between objects, count and answer. Certainly, a world of geometric objects remains simple. But all these implemented automatic understanding processes are impressive, as well as the matching knowledge modeling: the system is able to solve complex references, such as "a block which is taller than the one you are holding", to solve anaphora such as "put it in the box", to identify speech acts. The resulting dialogue is focused on the essential. There may be a lack of fluidity, but the goal is to satisfy the task, and indeed, all is done for this to happen.

As for GUS [BOB 77], it takes an additional step into the utilitarian MMD, with a flight reservation task. To demonstrate this research prototype, the database only comprises a single flight in California. Beyond this limitation, the linguistic modeling, the computer modeling and the methodological aspects give an idea of what the MMD domain will look like in a few years. Just as SHRDLU, the system is able to solve object and anaphora references, at least when they directly concern the task's objects, that is the flights, days and timetables. For example, it manages to allocate a date-type reference to the referential expression "Friday" used as a return date after specifying "May 28th" as an outward flight. The interpretation of the user's utterances triggers a syntactic and semantic analysis that can be partial, and thus operate on other linguistic materials than just full sentences. It also triggers a recognition of speech acts, notably with the understanding of indirect answers to some questions. The great results of linguistic works on the dialogue structure and the information structure are used, which leads to the system managing a great deal of knowledge on language: lexicon (3,000 roots recorded, which is greater than the precedent systems), morphological rules, syntactic constructions, simplified principles of the informational structure, patterns for the dialogue structure, conceptual models for the travel plans and dates, and finally the agenda model: central structure that will allow the system to manage events and know at any moment what task to carry out.

The computer implementation is being rationalized: the different linguistic analyzers are implemented as independent modules and a communication language between the modules is specified. The fact that the modules are independent allows us to test them, correct them and improve them separately. Actually, all design follows an exemplary methodology: the authors have started by collecting and studying the human dialogues focusing on the same task, that is they carried out a corpus study, the word *corpus* referring to a collection of attested linguistic material, and they even implemented a system simulation experiment (which will later be called a *Wizard of Oz*), to collect the data on the user's behavior when faced with the system they imagined. The fundamental methods of the MMD are set. Obviously, they are applied with the means at the times, and the computer's sluggishness, for example, leads to a wait between 10 and 60 s for each utterance, a wait which is taken into account in the simulation experiment, and is very far from the speed and naturalness of human dialogue.

1.1.2. *First oral and multimodal systems*

While it was possible up until now to present the major advances in MMD through a few emblematic systems, this has no longer been the case after the 1980s. Indeed, this decade saw a multitude of theoretical works that many researchers discovered, the dialogue and its characteristics, a multitude of prototypes and MMD systems, and notably the first oral systems and the first multimodal systems. Moreover, it was also the golden age of video games and the general public discovered adventure games with textual interaction[3], that were the first recreational MMD systems.

---

3 As an example, the SRAM (*Mars* backwards) game, published in 1986 by Ere Informatique, left a strong impression on the 12-year-old player that I was at the time: all the interaction in the game went through written commands, which led me to type, for example, "I want to go west" and see (visualization of the analysis steps carried out by the software) the utterance appear on screen with the words "go" and "west" highlighted in color, and then discover the software's answer: "you arrive near a waterfall", with the display of a visual scene in which the player must look for clues to continue his quest. The techniques implemented here are much simpler than those in SHRDLU or in GUS, with keyword detection instead of word sequence detection, and the keywords are almost always found within the verb–complement pattern, but the limits are not the same: the vocabulary and possibilities are vast, adapted to intensive use, and the game must also be robust, reliable and interesting.

Among the theoretical works that have marked the 1980s, there is the research carried out by conversational analysis [SAC 74] and discourse analysis or discourse pragmatics [ROU 85, MOE 85]. Although the objectives of these two studies differ, their focus – i.e. the recording and transcription of human dialogues – is the same, and the observations will give us a clearer view of the notions defined in the introduction and how they relate to one another (speaking turn, intervention, utterance and speech act). They will also help in the comprehension of the notions of cooperation, planning, conversational organization, dialogue structure, common ground, grounding and relevance, which we will see in Chapter 8. These works would contribute to numerous articles being published [ALL 80, CLA 86, GRO 86, CLA 89, COH 90] that would inspire the whole MMD community.

The first oral systems arise from the progress in automatic speech recognition. To operate correctly, they focus on well-defined tasks, like SHRDLU and GUS do. For example, the Nuance company develops various specialized systems, often for telephone dialogue for clients such as banks [COH 04]. As for the first multimodal systems, i.e. systems that match speech recognition with gesture recording which at first corresponded to simple clicks of a mouse, they appear in a famous article [BOL 80], which shows that multimodality is much more efficient than just speech to refer to objects, as long as the MMD system involves a visual scene. A new side of the MMD field was then opened, new questions were asked on ergonomics, on the spontaneity of multimodal dialogue, on interactions between MMI and MMD, and in general on all the inputs and limits of multimodality, see a summary in [OVI 99]. Among these questions, the following opened new perspectives: if an MMD system is able to carry out automatic interpretation taking multimodality into account, should it not carry out automatic generation also taking multimodality into account? With demonstrators, for example in the field of air control, we have begun to explore this issue of output multimodality and outline its own field of research, that of Intelligent MultiMedia Presentation Systems (IMMPS, see Chapter 9).

The 1980s are thus full of questions. After the first systems fascinated and helped clarify the methodology and limits of MMD, they gave way to natural dialogue in natural language with new goals such a spontaneous speech processing, gesture recording and use of interaction devices with all that it implies: contextual management, adaptation to the display device and adaptation to the user.

The 1990s kept on the same path by broadening the panel of expected functionalities in an MMD system. This decade corresponded first to the entrance of the digital age, the consequences of which were primarily a renewal in theoretical and experimental research on spontaneous oral language, and second to the introduction of programming techniques based on important calculations, especially probabilistic calculations, which were costly in computer resources. Research on oral language was until then hindered by technical constraints, but the digital world greatly helped promote the rise of oral analysis software, the multiplication of studies and finally a change in point of view on oral language, which acquired the status of full-fledged study subject and was not just a poor parent of the written language, or even a poor child full of mistakes [BLA 10]. The consequences for MMD are that the work is not only based on grammars and rules stemming from the written language; little by little the specificities of oral speech are integrated: corrections, repetitions, inserted clauses, as we will see in section 5.1.2. On the other hand, the use of speech input creates new issues for MMD systems, with, for example, the need for the user to use a key or pedal at the same time as he/she speaks (push-to-talk), to let the system know the beginning and end of his/her utterance. As for programming techniques, they are enriched by advances in statistical approaches that integrate probabilities calculated from a corpus that, as Jurafsky and Martin, [JUR 09, p. 892] underline starts the probabilistic processing of speech acts, and brings a supplement to MMD system realizations, which goes beyond the quality of previous research prototypes. Efforts were also made to enrich the automatic understanding methods, with joint approaches that combine both *bottom-up techniques* (starting from the utterance, the system carries out various analyses to identify the underlying meaning and intention) and *top-down techniques* (starting from the possible plans and intentions, the system carries out various analyses to determine which intention satisfies the utterance). These efforts involve research on the representation of plans and reasoning, which presupposes that the system manages to reason on the beliefs of the user [VIL 05, p. 6]. We then see models of the belief, desire, intention (BDI) type appear.

Among the systems in the 1990s, the Trains system [ALL 95] is exemplary since it tries to find solutions to a vast panel of challenges around automatic understanding, of a dialogue with joint initiatives (not solely commanded by the system or the user), of representation and reasoning on time, actions and events. The task falls into the domain of transportation, but

unlike GUS or the example we gave in the introduction, it involves various modes of transportation and thus manages the connections between these modes, the planning issues, the optimizations (journey length calculations), the potential conflicts, etc.

In France, systems and publications are multiplying [BIL 92, GUY 93, DUE 94, LUZ 95, SAB 97, GRI 00] and we will remember as an example the Dialors system by D. Luzzati, which focuses once more on train ticket reservation. The methodology starts here again with an in-depth corpus study, in this case a corpus coming from Société Nationale des Chemins de Fer Français (National Society of French Railways SNCF) recordings, a corpus that has also been the focus of various publications. The Dialors system has an analyzer called Alors whose function is to turn utterances into an internal representation to the system, and a dialogue manager, Dialogue, who decides, depending on the representation, on the action to be carried out: request clarification, answer the user's query after consulting the train timetable database. The second component has the role of implementing the dialogue model suggested by the author, a model that distinguishes the *governing dialogue*, i.e. the main dialogue reflecting the task's progression, from other potential *incendental dialogues*, i.e. the clarification requests and other transient sub-dialogues, which do not influence the task's progression but allow the interlocutors to understand each other. This dialogue structure allows the system to carry out fine analysis and also to assess in real time the task's progression, without requiring the implementation of a more complex model such as the hierarchical model of the school of Geneva [ROU 85], mentioned earlier as an approach to discourse analysis.

1.1.3. *Current systems: multiplicity of fields and techniques*

Our overview started in the 1950s and now reaches the 2000s. It is harder to use hindsight on this period that includes the current systems, especially since the work has multiplied and the number of techniques has increased. In general, beyond the improvement of all the models of the 1990s [JUR 09, p. 892], here is what appeared in the 2000s:

– the application of computer techniques of machine learning to MMD to relocate part of the different settings onto the big corpus processing or onto an improvement of the performances as the system is used [RIE 11];

– the tremendous efforts of standardization: W3C, ISO, TEI, DAMSL, etc.;

– the increase in system assessment methodology (see Chapter 10);

– the multiplication of communication modalities with the machine, and thus of the models and techniques of multimodal dialogue: force-feedback gesture or haptic gesture, gestures and postures caught on camera, taking into account the eye direction, lip reading, etc. [LÓP 05];

– the implementation of links with other scientific domains, such as robotics (see [GAR 11, Chapter 10]), and other fields of NLP, for example machine translation within MMD systems, being able to go from one language to another [LÓP 05];

– the increase in *toolkits* for a quick prototyping of MMD systems, for example the well-known VoiceXML, a standardized language for relatively simple, from a linguistic point of view, voice applications;

– the integration of MMD in wide intercommunication platforms, whether we are referring to ambient intelligence or other aspects, for example linked to software architectures [ISS 05];

– the rise of the embodied conversational agents (ECA) field that takes into account the emotions;

– the rise in the QAS field.

About this last item, for example, the point of integrating dialogue abilities to a QAS is to allow it to carry out exchanges to specify bit by bit the query [VIL 05, p. 48]. From a (single) QAS that is content with finding the result to a query, like the database managers do or like IBM Watson that is still limited by the rules of a game show, we move on to a system of questions and answers (plural), in which the dialogue allows for clarifications, precision, and especially follow-up questions on the same subject: “does this journey go through Meudon?”, “is this the shortest journey to get to Paris?”, “when does it leave?”.

An example for such a system using MMD and QAS is the Ritel project [VAN 07, ROS 08]. The system’s architecture highlights question management, with modules devoted to topic detection, user return management, dialogue history management, question routing, implicit confirmation management and additional query management. The project’s goals clearly highlight the QAS performances as much as the MMD performances, and the project is therefore a significant step for open-domain

MMD systems that started to emerge in the 2000s. As an example, to compare with the figures mentioned previously in this chapter, Ritel's vocabulary has 65,000 words, which approximately matches the number of entries in a language dictionary. Another example from the 2000s, the Amitiés system [HAR 06], a closed-domain MMD system provides us with a way to compare previous systems of the same type as well as an open-domain system such as Ritel. Amitiés was designed from an in-depth corpus study of about 1,000 dialogues all belonging to the financial domain on which one of the tasks is focused. The figures corresponding to this material are as follows: 30,000 sentences for approximately 8,000 words of vocabulary. This is much more than GUS could do, but is still very far from the 65,000 words of a language.

Finally, in the 2000s (and in the following decade), as we saw at the very beginning of this chapter, the first general public MMD systems have appeared, incorporated to various Websites, electronic diaries, geolocation systems and other personal digital assistants. Even if the quality is not there yet, we can imagine that it will help encourage the scientific community's efforts.

1.2. A list of possible abilities for a current system

At the level of general public systems, as we mentioned, we are still far from a natural dialogue in natural language. A few tests of systems, called *voice-controlled* or *voice-recognition* systems, allow us to quickly verify this. For example, the geolocation systems and cell phones are still at a keyword detection level: city names for the first and recipient names for the second. We are still very far from the automatic understanding of utterances such as "I want to go to Grenoble by bypassing Lyon and avoiding the highway between Saint-Etienne and Lyon", in which the user mentions a point of passage and different preferences for two parts of the journey all at once (a much quicker request to say than to program directly into the system, if at all possible). We must still admit that from examples such as these, voice control is not often adapted to the computer system user: it is often noisy, we are never sure of being understood properly, and we are always convinced of being more efficient by directly manipulating the system with a classic MMI. Contrary to what various researchers claimed in the 1980s, one cannot say that because there are more and more computers and more and more data accessible that the MMD will impose itself as a new communication mode. As Vilnat [VIL 05, p. 5] states,

the question should instead be to know for which tasks it would be useful to implement a dialogue rather than any other kind of interaction technique: the major hindrance is the low interest of users in using an MMD system.

At the level of research prototypes, the natural dialogue in natural language becomes feasible, at least within the framework of a targeted task. This is the case for the closed-domain dialogue and also for some open-domain demonstrators such as IBM Watson. One should however note that the recent endeavors have focused on broadening the systems' abilities rather than developing NLP aspects. We will see this in the three parts matching the three characteristics of a cognitive system: input processing (section 1.2.1), the system's internal analyses (section 1.2.2) and output management (section 1.2.3).

1.2.1. *Recording devices and their use*

Chapter 2 of [LÓP 05] draws an exhaustive list of the multimodal MMD systems with processes carried out on inputs. Without drawing up such a list again, let us quickly mention the following recordings: speech recording; lip-reading the user to help or even replace speech recognition (noisy environment, disabled user, whispering); user recognition; face location and tracking, as well as mouth or eye tracking, and thus eye direction (both to monitor attention in relation to the dialogue and to help resolve a reference to an object in the scene); facial emotion recording; pointing gesture recording, especially those of the hand, and more general kinds of gesture made with the hands or the body. Moreover, we have already mentioned the force-feedback gesture in the case of a haptic interaction: this is a device that manages both the recording of the hand's position and the generation of a potential resistance toward the user. The point is to couple this device to an immersion in a virtual environment, the user seeing a graphical representation of his/her hand manipulating objects in the virtual scene. In this context, the force feedback makes complete sense: it simulates a touch perception that completes the visual perception.

There is no system that can carry all this out simultaneously and in real time, but it is an interesting challenge for the more technophile members of the MMD research community. We can see there are many possibilities and the computer challenges are vast: the processes matching these types of recording include many issues falling within the scope of artificial vision,

signal processing, mathematical modeling adapted to the representation of configurations and trajectories, all that with constraints of execution speed, precision and abstraction in representations that the system can efficiently manipulate, so the researcher can confront these representations with those stemming from automatic utterance understanding. As Bellalem and Romary [BEL 96] show us, for example, for gesture trajectories carried out on a touch screen, a representation of a gesture under the shape of a sequence of several hundreds of positions is simply unmanageable. It is necessary to abstract regularities and significant instants from it to reach, for example, a curve that can be described in four or five parameters. If this curve is then used to help resolve a reference to an object, it will be possible to confront it with a representation (also simplified) of the visual scene and the objects that appear in it.

Some processes require specific recording devices, with the immediate examples of a microphone for processing speech and of the keyboard for processing writing. Other processes can be carried out in various manners, from the most troublesome to the most transparent. An example of *troublesome recording* is the pointing glove that the user had to put on so the system can record the position and configuration of his/her hand or the glove with an exoskeleton required for force feedback. The increasingly common example of *transparent recording* is the camera or coupled camera system that allows the user the freedom to carry out various processes simultaneously, for example tracking his/her face and detecting the configuration of his/her hand.

Automatic speech recognition is a field in itself, and its use in MMD creates additional issues [JUR 09]. The idea is to go from an audio signal to a transcription according to a code which is more or less close to written language and requires various data sources, including the following: an acoustic model, a list of words in the given language, a dictionary of pronunciations and, the source of almost essential data to increase performances, a *language model*. This model is built from statistical corpus analyses. By bringing the notion of context (one, two or three previous words), it allows the system to calculate the probabilities and retain the most probable hypotheses for the word (or other unit) it is currently recognizing. In the framework of a speech dictation, the language model is built from calculations carried out on texts taken from literature or the written press. We maximize the size of these texts so as to refine the language modeling in

terms of possible word sequences. Within an MMD system framework, building this model requires us to aptly choose the corpus used for statistical calculations: it is not necessarily relevant to only keep oral dialogue transcriptions, but having dialogues close to those expected by the system is a definite advantage, even if various language models have to be managed. Moreover, alternating the user and the system's interventions brings us an additional limitation: the probabilities for a user's utterance depend on what the system just said. The language models thus have to take into account the state of the dialogue, and become more and more difficult to manage.

There is an additional issue compared to vocal dictation: if the result consists of a written text matching what has been said, the speech recognition module result in an MMD system can be much more detailed. First, it can include various recognition hypotheses, so that the following modules make a choice depending on their own expectations. When an utterance includes an unknown word, i.e. a sequence of phonemes that do not match any of the words in the lexicon, the recognition module has a choice between various solutions: either bring it back to one of the words of the lexicon, even if the pronunciations are vastly different, or try to transcribe the sequence of phonemes with a potential spelling depending on the languages. While these two solutions might be acceptable for speech dictation, the second, for example, perfectly adapted to transcribing surnames that the system does not recognize, it is not the case for MMD: not only does the recognition module have to indicate that it is an unknown word, but it also has to transmit a code describing the word's pronunciation, so that the system can add it to its vocabulary and pronounce it in turn, if only to ask the user what it means. To get the job done, each recognized word is given a confidence score, and the syntactic or semantic analyzer uses these confidence scores and its own preferences to find (rather than have imposed) the most plausible transcription of the utterance.

An additional aspect with consequences on the nature of the result transmitted to the other modules of the MMD system is found in the prosody. Whether one is talking of the role of the recognition module or of another specific module, it is useful for the written transcription of the utterance to be accompanied by coding, by a transcription of the prosody. We will see in Chapters 5, 6 and 7, that prosody helps in semantic analysis (by providing focalization clues), in solving references when a gesture is used jointly with a referential expression and in identifying speech acts, by providing a tone

outline that allows us to privilege one hypothesis over another. We thus expect various indications from the prosodic analysis module: locating the focalization accents, temporal breakdown of the utterance, word by word, to match words and gestures in a multimodal dialogue, and a coding of the intonation's main characteristics. More in-depth analysis, with, for example, the detection of periods, requires additional indications, but for now falls more in the domain of subsequent oral corpus analysis than the domain of real-time analysis for the MMD. This is actually a criticism that can be used against many of the current systems: they do not use prosody, even though it is an essential component of oral language. Initiatives such as that of [EDL 05], who presents Nailon, an automatic prosody analysis system able to detect in real-time various prosodic characteristics of an utterance in MMD, are important.

One last aspect in which automatic speech recognition module has a role to play is speaking turn management. The MMD systems have long remained limited to an alternating operation of interventions, the system never interrupting the user and only starting to speak once the user has finished his/her utterance. More than that, we saw in section 1.1.2 with the push-to-talk button or pedal that it was on the user to let the machine know the beginning and the end of his/her intervention. We are now able to expect that an MMD system will let the user express himself at any point, with no constraints, and it is up to the machine to detect the beginning and the end of the interventions. This is actually one of the functions of the Nailon system, which uses prosodic clues of fundamental frequency and rhythm to automatically detect the end of a user's intervention.

1.2.2. *Analysis and reasoning abilities*

Once the signals have been received at the system's input and transcribed into appropriate representation, many analyses and reasonings will be carried out so that the system can understand the meaning of the user's utterance, his/her intent and, thus, the answer to give him/her. The analyses fall in the domain of automatic understanding of natural language, that is of NLP, and cover the following aspects: word identification (lexical analysis) so as to find their meaning (lexical semantics) stored in the system according to a well-defined formalism; the identification of the sentence's structure and the grammatical functions of the identified components (syntactic analysis): the construction of the sentence's semantics by combining the meaning of words

and following their syntactic relations (propositional semantics); the allocation of referents to the first and second person pronouns, to referential expressions in general, and to anaphora in particular (pragmatic analysis sometimes called *first-level* pragmatics); the identification of the implicit and the context attached to the utterance (*second-level* pragmatics); and the determination of speech acts so the system can understand the nature of the user's intervention (*third-level* pragmatics). Beyond the simple transcription of the *literal meaning* of an utterance, here we enter into the field of the determination of its *contextual meaning*. As for NLP, the implemented methods and algorithms have evolved for all these analyses. Where at one time, the symbolic approaches stemming from the AI were the only ones around, we now see statistical approaches, and they are at all levels of the list given above. These approaches have proved their efficiency in many fields of NLP, and have sometimes completely replaced symbolic approaches. In the MMD domain, it is the hybridization of symbolic and statistical approaches that provides us with the most promising results.

Starting with a semantic representation that is faithful to the utterance, we then reach an enriched representation through one or more implicit or explicit messages that the utterance carries forth. This enriched representation is what the system will confront, internally, to previously manipulated representation as the dialogue advances. It is also due to the information it contains that the system will be able to abstract the structure of the dialogue and compare it to structures that are considered for this task. This approach was imagined back in the 1990s [REI 85] but its computer-based implementation within a real MMD system framework only happened much later, and is still going on today. The system can thus carry out an assessment of the task's satisfaction, identify the deficiencies and decide what its next intervention will be. All this proceeds from the reasoning that it implements so as to process the user's utterance in the more relevant manner, taking into account what has already been done, what the utterance brings to the dialogue, and what still needs to be achieved to satisfy the task. Here we are situated not in linguistics and pragmatics but in modern AI: the themes approached are those of knowledge representation and especially following formalisms stemming from logic, in order to allow automatic deductions and those of expert systems and multicriteria decisions. Actually, as for the analyses mentioned in the previous paragraph, we are faced with a hybridization of various approaches. As an example, the approaches based on the expressive power of a well-defined logic, and its consistency with

natural language, have explored different types of logics: propositional logics, modal logics, temporal logics, description logics and hybrid logics.

1.2.3. *System reaction types and their manifestation*

Once the system has decided what action to carry out as a reaction to the user's utterance, it still needs to materialize this action. In the case of a system which is only written or only oral, it must generate an utterance in natural language. Natural language generation is a research field in itself [REI 00] that includes various aspects such as sentence construction and the determination of referential expressions. We find here the same issues as those involved in automatic understanding, but in an inverted manner. Even if they have some linguistic resources in common, the generation methods and algorithms are specific and are not a simple overturn of their understanding equivalents. In the case of an oral system, the last step carried out by the system is text to speech, that is pronouncing the chosen utterance. We can also find here the concerns of prosody: to look real, the utterance must be pronounced with intonation, rhythm and even focalization, and all these must be perfectly in keeping with the system's communication intent.

In the case of a multimodal system, for example when an avatar graphically represents the machine, the issue includes the gesture generation and their temporal synchronization with the words of the generated verbal message. When gesture is possible, the issues of generation, and especially that of determining referential expressions, reach a new dimension: speech and gesture complement each other, and the system must choose which part of the message to allocate to each aspect. Moreover, the graphical design of the avatar itself is a field that generates important questions about the realistic aspect of the avatar's physical appearance, its gaze and its movements – eye, eyebrow, lip movements when an utterance is verbalized – of the head, if nodding, and more generally of the body. The indications given to the user through these movements play a role in the man–machine communication, and it is essential for the various movements that have the same purpose, for example those of the eyes and the eyebrows that indicate the avatar's attention level, to be correctly synchronized. The emotions are also transmitted through gestures and are also a field of research on to themselves, which requires studies on the typologies of emotions, their relevance in MMD and the way they should be rendered, not only visually, but also in speech. Finally, in the case of a system including an MMI and manipulating a vast quantity of data,

the issue also includes that of the graphical presentation of information (such as the Immps mentioned earlier). It can happen that the MMI itself generates earcons. The earcon generation and the speech generation must then be carried out in a relevant manner, i.e. without superimposing them, which might hinder the user's perception.

1.3. The current challenges

Can a machine think? The Turing test and Grail-style quest for the talking machine is still as fascinating as it ever was, but the current limits to the MMD systems mean that the question is not set in these terms any more. In a more pragmatic approach, the questions are set in terms of limits in the abilities of the machine to model and naturally process natural languages, to represent and reason on logical representations and to process and integrate various and varied signals. Commercialized versions prove this everyday: an MMD system only works well within a limited application framework, that is in a sufficiently limited framework for the maximum interaction possibilities to have been imagined upstream, during the design phase. Contrary to what attempts such as ELIZA had us believe, nothing is magical, and no matter what technologies are implemented, whether or not they are symbolic, statistical, or involving machine learning or not. Everything must be anticipated, and this represents an amount of work proportional to the considered abilities for the system.

The main challenge of MMD remains, as it was in Pierrel's time [PIE 87], the multidisciplinary design of comprehensive systems that allow the user to express himself/herself spontaneously as he/she would with a human interlocutor, and this for a variety of applications, so as to offer systems that are accessible to anyone, in all everyday situations. More precisely, we will develop four sets of challenges: theoretical challenges, challenges concerning the span of expected abilities in a system, technical challenges concerning system design and technical challenges trying to help system development.

1.3.1. *Adapting and integrating existing theories*

According to Cole [COL 98, p. 191], recent strides have not included the development of new theories but focused on the extension and integration of existing theories. Thus, we can find many hybrid approaches that use the expressive power of more than one existing theory. This observation that we

mentioned earlier when talking of the increasing closeness between symbolic approaches and statistical approaches is still true today. In linguistics, we mentioned the prosodic, lexical, syntactic, semantic, and pragmatic analyses; what was long considered to be a succession of analyses carried out one after the other is now approached in a completely different manner: a part of the results of the prosodic analysis is used for the semantic analysis and a part of the results of the syntactic analysis is used for the pragmatic analysis, and the latter is not monolithic but involves various aspects that are almost independent from each other. One challenge consists of completely reviewing the classic breakdown into natural language analysis levels, and better integrating the analyses that have common goals. In MMD, the goals are a list that depends on the targeted system but includes at least detecting the end of the user's utterance; representing its meaning in a logical manner, or at least, as a data structure that is usable by the considered algorithms; resolving the references to the objects managed by the applications; identifying the implicit content carried that has not been explicitly said by the utterance; updating the dialogue history; etc. Each goal in this list is reached due to the help and collaboration of various analyses. For example, to automatically detect the end of the user's utterance, we need a prosodic analysis that indicates when the tone outline dips and thus provides a hint, and we need a syntactic analysis, which shows if the sequence of words captured until then is a grammatical sentence or not, and whether or not it needs additional words. Depending on the system's personality, especially its tendency to interrupt the user, we can even imagine that a semantic analysis provides an additional argument, as soon as a semantic result is obtained. If we remain within a cascading analysis operation, this type of mechanism is impossible. One of the challenges is thus to explore the collaborative analysis implementations. If we start the first analysis at the end of the user's utterance, then we lose any possibility for the system to interact in real time. One of the challenges is thus to carry out analyses at all times, almost one for each word uttered by the user. If we consider a module to be a black box that gives a result at one time and within a single data structure, then the prosodic analysis should not materialize itself in a single module but in various models: one for the determination of the tone outline, one for the detection of prominences, one for the rhythm, etc. A modular breakdown that follows the breakdown into linguistic analysis levels cannot be justified any more, and the application of linguistic theories to MMD is still the focus of research. In multimodal dialogue, the integration of theories is all the more crucial: the gestures are

linked to prosodic aspects, ergonomic aspects, etc. As we will see in Chapter 2, collaboration between fields is essential.

Finally, to end this list of theoretical challenges, let us underline the importance of methodology, with which to carry out experimentations and to create and use reference corpus for the MMD. This challenge is linked to resources and is key not only for the study of dialogues covering a specific task, but also to carry out machine learning algorithms or to generate data such as lexicons, grammars and language models for the oral dialogue as well as the multimodal dialogue. In this case, one of the challenges is in a better integration of these resources. As an example, the Ozone project we have mentioned allowed us to reflect on the concept of meta-grammar (or meta-model), with the goal of instancing from a joint base of linguistic grammar and statistical language model.

1.3.2. *Diversifying systems' abilities*

The technical challenges linked to the abilities of an MMD systems are the NLP, AI, ECA, QAS and MMI tasks, and many more. In general, all the components we have mentioned could be improved, with a greater scope of phenomena taken into account and a greater finesse in their processing. Cole [COL 98] highlights various linguistic aspects such as exploring the nature of discourse segments and the discourse relations, as well as the need for additional mechanisms to manage key phenomena such as the highlighting of information in a linguistic message. All these challenges focus on the same goal: increasing the coverage, the fluidity and the realistic aspect of the dialogue. To make it more clear, the goal might be to achieve a natural dialogue system or even a natural multilogue system in natural language [KNO 08], which will be multimodal, multilingual, multitasking, multi-roled, multi-thread, multi-user and, of course, capable of learning etc.

The question of *realism* is a great question, which starts with speed: a system taking 10 s to answer does not have a chance of achieving realism. If this criterion can be measured, however, there are some that cannot: how should we measure the realism of a synthetic voice, of sentence construction, of an ECA's gestures? The fact that some users reject an artificial voice is sometimes based on tiny details that are hard to measure, such as a minute defect in elocution rhythm. The perception of these minute defects can create

unease and disturb the user. The field of robotics or that of computer-generated images use the term of *uncanny valley* to describe this type of phenomenon. The issue is that we try and get closer to the human (to reality for computer-generated images) but there is still a small gap between what is achieved and what is aimed for. And this gap, as minute as it might be, is enough to be perceived, and to irritate. To counter this, some designers make the gap visible and forego the goal of getting close to the human. So some mechanical toys that look like animals do not have any fur. In MMD, for example, the Web service Ananova takes on the appearance of a gorgeous young lady … with green hair [HAR 05, p. 341].

Finally, a key challenge for the abilities of an MMD system is its robustness, that is its ability to manage its own shortfalls, at a linguistic analysis level for example, its own deficiencies and errors, and its ability to always bounce back, to help the dialogue progress no matter what the cost, by using the task to be solved, or not. This implies the design of modules able to operate with incomplete entries and have strategies to manage problems. This also implies ability to predict, from the first stages of design, tests and settings with real data, real conditions, rather than laboratory-controlled conditions.

1.3.3. *Rationalizing the design*

At the design level, there are multiple methodological and technical challenges. Once the list of understanding and generation abilities is determined, they have to be instanced and organized into modules, components or agents in an architecture, and the interaction languages between these elements, the evaluation methods and construction methods of necessary resources of integration have to be specified. The main challenge here is the rationalization of the architectures' engineering (see Chapter 4), and in general the rationalization of production flows, as in any professional technical field. Thus, Harris [HAR 04] focuses, Chapter 9, on a very precise description of a design team, with the different professions involved: a dialogue team leader; an interaction *architect*; a lexicographer in charge of the aspects linked to corpus; a *scriptwriter* in charge of anticipating the expected types of dialogues, but also the definition of the system's personality and its possible reactions; a *quality engineer*, without forgetting the ergonomics experts, technical experts as well as an expert in the field covered by the task. The task, and more generally the context of the dialogue, can require integration into another field of research. A first example is robotics, in which

we are starting to see systems integration abilities belonging to robotics and MMD abilities, preferentially in a multimodal manner [GOR 11]. A second example is that of MMI when we try to give them speech, while keeping, on the one hand, the possibilities of directly manipulating the HM and, on the other hand, the MMI advantages in terms of ergonomics, plasticity: adapting to the user, the terminal, the environment.

1.3.4. *Facilitating the implementation*

At the level of system development, the technical challenges are found in the facilitation of development processes. A first step on this path is the multiplication of toolkits devoted to MMD. VoiceXML is a basic example, but there are many other platforms devoted, for example, to helping design multimodal dialogue systems [LÓP 05]. A second step would be the implementation of a library offering a rich and performing panel of NLP tools and dialogue managers. This is an important challenge and was tentatively introduced by products such as Apache's OpenNLP for some aspect of written NLP. An *OpenDial* library would probably be useful and would help focus efforts elsewhere rather than on the components that all systems have in common. Finally, a third step in the same direction would be the materialization of a whole set of services linked to vocal recognition, text to speech, prosodic, syntactic and semantic analyses in a software layer such as middleware, or better yet, in a computer extension card, such as graphics cards for 3D visualization. This challenge, if it happens someday, would allow it to be exceptionally easy to develop a system: all processes would be carried out in hardware rather than software, which increases the speed, and it would really open the door to systems usable in real time. Obviously, this is not a simple challenge, and if we compare it to 3D, for which the graphics card works much more during the design than the final product, which needs overspecific and overdelicate processes, we could imagine that a *dialogue card*, at first, would accelerate and simplify the design of systems without carrying out the full development.

1.4. Conclusion

The quest for a machine able to understand human language and answer its user as well as a human interlocutor has gone on for more than 50 years. The issues arising from natural language processing have not allowed us to

achieve completely comprehensive systems yet. However, we can note a diversification of communication modes and aspects of the man–machine interaction. By relying on the theoretical, methodological and technical stages that have marked the history of the man–machine dialogue, this chapter has outlined the abilities considered for a man–machine dialogue system, and the current scientific limits and challenges in this field.