HEY *SIRI*, WHAT CAN I TELL ABOUT SANCHO PANZA IN MY PRESENTATION? INVESTIGATING *SIRI* AS A VIRTUAL ASSISTANT IN A LEARNING CONTEXT.

B. Arend[1]

University of Luxembourg (LUXEMBOURG)

Abstract

In this paper, we address challenges arising from the use of the conversational interface *Siri* during a literacy activity in a learning context.

On one hand, we investigate potentialities of *Siri* as a virtual assistant and knowledge navigator for task accomplishment. Thus, we attempt to suggest possible lines of mobilizing *Siri* to afford and improve both students' task elaboration and second language performing during a literacy activity. In human-*Siri* interactions, a wide range of operations are carried out through closely intertwined oral and written instances of natural language (vocal and visual accounts of 'understanding').

On the other hand, we raise some crucial, critical questions with regard to concrete situations involving *Siri* as a 'learning' assistant. How do students deal with the unpredictability of the interactions with the virtual conversational agent? How do students handle interaction modalities such as identifying and 'well' pronouncing the right commands and the choice of words that activate the diverse features? Our paper seeks to provide Conversation Analysis (CA) [1] informed insights into human-*Siri* communication in a learning context.

Keywords: Conversational interface *Siri*, conversation analysis, learning assistant, challenges

1 INTRODUCTION

Our paper aims at discussing challenges arising from the use of the conversational interface[2] *Siri*. In a social scientist perspective[3], informed by conversational analysis [1], we attempt to suggest some possible lines of investigating interactional 'matters' users face when talking to/with *Siri*, more specifically in learning contexts. Following a brief introduction, we provide thoughts on the relevance of Conversation Analysis in order to study human interactions with the conversational interface *Siri*. The next section then provides examples of human-*Siri* interactions during a literacy activity by sketching out conceptual and rather practical emerging challenges. Thus, our investigations are supposed to shed light on two complementary aspects of *Siri* mobilization: (How) can *Siri* act as a learning assistant? Can *Siri* be considered as a communication partner?

1.1 Talking to *Siri*: General Issues

We note that *Siri* is a so-called virtual assistant and knowledge navigator with a voice-controlled natural language interface that uses sequential inference and contextual awareness to help perform personal tasks [2, 3]. As McTear et al. [3] point out, "with recent advances in spoken language technology, artificial intelligence, and conversational interface design, coupled with the emergence of smart device", it is now possible to use voice to perform tasks on a device (sending a message, setting an alarm, making a research, …). Thus, users can address spoken commands and questions and have audible and written reply from *Siri*[4]. Many tasks would require multiple steps to complete using touch, scrolling, and text input, but they can now be achieved with a single spoken command [3].

[1] In collaboration with **Henri Hansen**, www.comitoergosum.com (website in progress).

[2] "We use the term conversational interface to refer to the technology that supports conversational interaction with virtual personal assistants (VPAs) by means of speech and other modalities" (McTear et al., 2016, 11).

[3] In a CA approach, see *Methodology* section 2.

[4] *Siri* enables users of Apple iPhone 4S and later and newer iPad and iPodTouch devices to speak natural language commands in order to operate the mobile device and its apps (Rehal, 2016, 2021). *Siri* is integrated with Apple services like iMessage, Calendars, Safari browser, among other external services used to consult information and thus be able to perform tasks as to make an appointment on the agenda, send a text message among other possibilities.

Voice input is indeed often the most appropriate mode of interaction, especially on small devices where the physical limitations of the real estate of the device make typing and tapping more difficult. The users' speech utterances act as organizers directed at operating the mobile device and its apps i.e. generating information or performing tasks[5].

2 METHODOLOGY: SOME CONVERSATION ANALYSIS INFORMED INSIGHTS INTO HUMAN-SIRI COMMUNICATION 'PROCEDURE'

When focusing on human-*Siri* communication, certain questions arise. A priority issue is how talking to *Siri*? Can we talk to *Siri* in the same way we talk to one another? Effective conversational interfaces are supposed enabling people to interact with smart devices in a natural way, like engaging in conversation with a person. So, is the dyadic interaction between the human participant and *Siri* something that might reasonably be called a 'conversation'? [4]. Which are our expectations when talking to/with *Siri*?

In order to address these questions, we will provide some enlightening insights into human-*Siri* communication 'procedure' in terms of Conversation Analysis (CA) [1].

2.1 Building on assumptions about the interactional partner's expectancies

Conversation analysis studies the methods human participants orient to when they organize social action through talk [1, 5, 6, 7]. In other words, CA is concerned with how people achieve courses of action in and through talk and how they make their respective understanding of the actions accountable to each other. Thus, conversation analytic research states that humans always adjust their actions to a specific recipient. Sacks et al. [1] refer to "recipient design" as "a multitude of respects in which the talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the particular other(s) who are the co-participants". That means that by building on assumptions about the interactional partner's knowledge and expectancies, participants adjust their turns to the recipient, thus constituting a continuously modified 'partner model' [8].

In this line of thought, we note that humans usually adjust their questions or commands to *Siri* by 'easily dealing' with *Siri's* speech recognition system, all the more since *Siri* gives hints about how to be addressed.

Example 1:

Siri: *Some things you can ask me:*

Find This Week in Tech podcast in iTunes store

Play Bea's voicemail

Find a good pizzeria nearby

Check my email

Make the screen brighter.

Furthermore, we assume that *Siri* can build on assumptions about the interactional 'partner' by 'interpreting' his/her words in a context related way, e.g. by localizing his/her geographical position. In the below displayed interactions, *Siri* makes accountable that it is able to rely on relevant apps as well as to understand context in order to be responsive[6], i.e. perform the requested reply. Moreover, *Siri's* response, its accounts of understanding are carried out in multiple modalities. Besides a spoken answer, *Siri* simultaneously provides the transcription of the oral utterance as well as access (where appropriate) to relevant websites.

In *Example 2*, in its reply, *Siri* takes into account the human participant's local(ized) position and by that way his/her perspective.

[5] For the anecdote, in 'Dominate the day', we see Dwayne Johnson relying on his personal assistant *Siri* in different situations, from travelling to space to hailing a car. 'The Rock' and Apple unveiled a mini-movie via a worldwide launch in July 2017 teaming the action star with the tech company's voice assistant *Siri*.

[6] Siri's responsivity will be addressed in an upcoming paper; upcoming website comitoergosum.

Example 2:

Human: *Where is the nearest pizzeria?* ((oral utterance and written transcription))

Siri: *Here's what I found* ((oral and written reply))

[website and geographical coordinates]

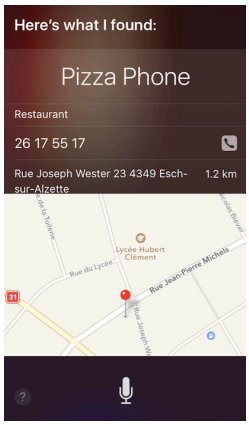

Figure 1

Note also that the human participant's question is quite narrowed through the use of the grammatical superlative "*nearest"* and that *Siri* can visibly (audibly) deal with this kind of linguistic accuracy. (By the way, the participant indeed can confirm *Siri's* 'mission' as accomplished as well as his expectations fulfilled (*Figure 1*).

In the following example (*Example 3*), we point out that it is important to consider whether an utterance such as "*is there a pizzeria near here"* is intended as a question or as request for directions [3].

Example 3:

Human: *Is there a pizzeria near here?*

Siri: *OK, Béatrice. Here's what I found:*

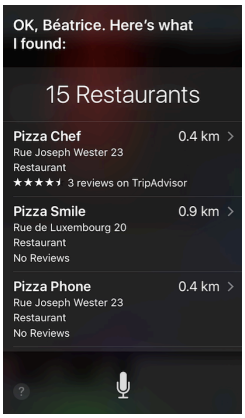

Figure 2

Siri's answer "*OK, Béatrice. Here's what I found"* (*Figure 2*) is supported by a list of 15 pizza restaurants located quite near to the user's geographical situation, in each case with distance measurement in *km*. The answering performance makes accountable that *Siri* is able to treat a deictic local reference ("*here"*) by considering the human participant's perspective: "*here"* is associated with the user's local position.

Moreover, when the user gives subsequently her next voice command "*show me the way to the second one", Siri* replies quite quickly:

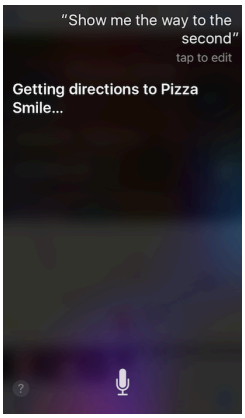

Figure 3

After having announced the launch of the process (*Figure 3*), *Siri* provides access to the built-in *Maps* application guiding the participant, from her starting position (at home), to the pizzeria represented as second item on the before provided list (*Figure 2*). Thus, by relying on the immediately previous interaction, *Siri* was coping with the participant's 'second' command as a part of a chain of actions and shows that it is possible to ask follow-up questions. Voice commands can be tied to prior ones by *Siri*[7]. Here, by the means of *doing* establishing the backwards oriented link, the conversational interface suggests that it takes the participant's concerns into consideration in their interactional unfolding, although establishing relevant participant oriented connections between follow-up interactions still remains a big challenge for *Siri* [3, 4]. Note that 'keeping track' in a conversation appears to be a very important issue for improving conversational interfaces since systems are expected to be more human like.

2.2 Examples of turn-taking in human-*Siri* communication

According to CA [1], conversation is unfolding in time referring to what has been said before and simultaneously initiate expectations about relevant next utterances. Speech utterances do not occur in isolation but relate to each other in a number of ways [3].

More specifically, in a CA perspective, conversation is considered as organized in co-occurring structural units, i.e. sequentially organized. Sequences are ordered series of turns through which participants accomplish and coordinate an interactional activity. The most common type of sequences are dyadic adjacency pairs uttered by two different speakers who produce one turn each. Typical instances of adjacency pairs are question-answer, greeting-greeting, request-acceptance/rejection. The first pair part (FPP) always requires a certain type of second turn (SPP); there is an expectation that the second participant will produce a response that displays the understanding of the first pair part. In this way, each turn provides both an opportunity for the participants to monitor and to display their understanding.

In this light, we can assume that human-*Siri* communication is built on unfolding series of turns. Either *Siri* initiates 'conversation' by inviting the human participant to perform a voice command or the human utters the first pair part. The FPP is multimodally uttered: The *oral* invitation/command is transcribed *in situ* and displayed as *written* text. Through the transcription of the human's voice command, simultaneously presented, *Siri* provides a situated visual account of its understanding, i.e. of its speech recognition ability. At the same time, the visuospatial instantiation of *Siri's doing* comprehending natural spoken language allows the human recipient monitoring *Siri's* understanding as well as making assumptions about the relevance of *Siri's* next turn (SPP). As mentioned above, *Siri's* reply is carried out in multiple modalities. Oral and written language are closely interrelated components of the system, a characteristic feature which might be of interest for mobilizing *Siri* in educational contexts[8]. But, herein lies one crucial difference to human communication! Recent CA informed research in human communication indeed underlines the complex coordinating dynamics of verbal (speech) and *various* non-verbal utterances (gaze, gesture, body posture) co-constituting a face-to-face conversation [9]. In addition to that, we must not forget that people "convey much more

[7] With restrictions [3, 4]

[8] See section 3

than just meanings of the words spoken. Their speech also conveys their emotional state and aspects of their personality" [3]. Human participants' conduct display understandings of the physical world going beyond their current interaction [10].

Moreover, CA considers conversation in terms of turn allocation at TRP (transition relevance places) [10]. Mostly, turn transition (speaker change) is accomplished smoothly at TRP, and such places are accountably projected. Participants display 'recipiency'; they show to each other that they are ready to listen or that they will go to talk. According to Heath [11], human participants display 'availability' and organize a pre-initiating activity providing an enabling environment for communication. Furthermore, the following 'rules' of turn-taking are relevant: Only one person talks at a time. Overlap of speech is common but brief. Participants proceed to the next turn with very little gap. Longer gaps and silence should be avoided; when they occur, they are meaningful and are most of the time perceived as trouble [5, 6, 7].

We note that *Siri* displays availability (*Figure 4*); in addition to a written utterance eliciting the human participant to perform a voice command, there is an oscillating line[9] (on the bottom of the screen) signalizing *Siri's* 'recipiency'[10]. A moving line is also an account of *Siri's doing processing* (i.e. *in process/progress*) (*Figure 5*). With regard to the *avoid the gap* rule, this 'procedure' is relevant for human-*Siri* communication as a means of orienting the inferences made by humans when there is a delay in *Siri's* response[11]. McTear [3] argues that "a brief period of silence might lead the user to infer that the system has not understood or is unable to generate a response. An advanced conversational agent that emulates the processes of human conversation would also have to be able to model conversational phenomena such as these in order to engage naturally in conversation".

The moving waveform (as well as the microphone icon or the colorful circle) (*Figure 6*) are visual cues, within the realms of technical possibilities, for taking into account turn allocation phenomena, e.g. for signalizing TRP. We should note here that *Siri* cannot continuously hear: speech overlap is not possible. The limited ability to listen is a crucial difference to human interaction. Human participants can do both at any point in the interaction, monitoring their conversational partners' utterances *and* respond to all actions. In terms of turn distribution, when talking to *Siri*, the human participant has to adapt to *Siri's* system specific *'economy' of opportunities to speak* [1].

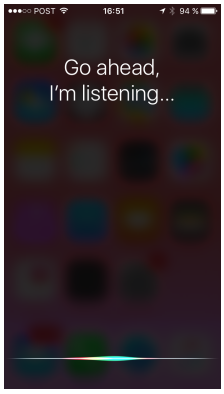

Figure 4: Siri is listening

Figure 5 Siri is 'searching' for information

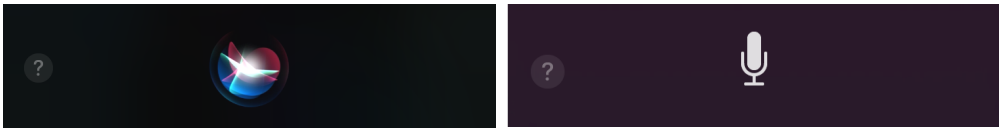

Figure 6: Siri is talking/has talked; the user has to touch the icon to be heard again

[9] representing the waveform of the ambient soundscape

[10] The voice command directed to *Siri* by the human participant is also represented by this kind of line.

[11] We underline here that delays are very short!

3 INVESTIGATING *SIRI* AS A VIRTUAL ASSISTANT DURING A LITERACY ACTIVITY IN A LEARNING CONTEXT

Virtual learning assistants are supposed to provide support for task elaboration and to foster information/knowledge building[12]. Could *Siri* be considered as a virtual 'learning' assistant facilitating task performing in a learning context?

In this section, we present some aspects of a conversation analysis driven case study investigating how a student is mobilizing *Siri* in order to realize a (re)search activity on famous servant and master characters in literature. We will point out both potentialities of *Siri* to contribute to task accomplishment and to elicit second language performing as well as 'limits' of the conversational interface.

3.1 Setting the Scene

A secondary school English teacher wanted to explore the feasibility of integrating *Siri* into his English Second Language classroom[13]. The participants are students attending a third year in a Luxemburgish Secondary School. English is learned as a second language additionally to German and French. At the time of collecting the data, the students had their second year of English Second Language class. The teacher asked his students to engage in a *Siri* supported homework. Thus, he gave the students the assignments to gather information about fictional characters interacting with each other as *servant and master* such as (among others) Sam and Frodo (*The Lord of the Rings*), Sancho Panza and Don Quixote (*Don Quixote*), Friday and Robinson (*Robinson Crusoe*) by mobilizing *Siri*. While doing homework, the students had to document the interactional communication process (by making screenshots of the smart device[14]). Then, they had to present their findings to each other in the Second Language classroom.

3.2 Mobilizing *Siri*: Challenging 'the Interface between the Written and the Oral'

As the teacher's pedagogical issue focuses on the students' oral *and* written language performances (in English), we should give special regard to the complex dynamics between oral and written language related to *Siri*. In section 2, we sketch out the relevance of simultaneously 'produced' vocal and visual written utterances as closely intertwined in human-*Siri* communication, for understanding and monitoring understanding. *Siri's* transcription-performance (consisting in synchronically transforming the participants' oral commands as well as *Siri's* oral replies in written language) is a key feature. Thus, from a conceptual point of view, we should touch upon some specific issues concerning 'the interface between the written and the oral'[15] [12].

According to Goody [12], writing makes it possible to scrutinize discourse in a different kind of way by giving oral communication a semi-permanent form. Moreover, writing increases the potentialities of criticism because writing lays out discourse before one's eyes in a different kind of way. At the same time, it changes the nature of communication beyond that of face-to-face contact. In this line of thought, Olson [13] argues that writing "is the technology for making utterances and thoughts real" and that written signs are as much parts of the external world as any other physical object. Kalantzis and Cope [14] underline that speaking is unfolding in time while writing is in addition visibly materializing in space.

With regard to these thoughts, we argue that *Siri* displays the user's as well as 'its own' discourse 'before the participant's eyes' through transcribing the oral utterances. Thus, the spoken word becomes a visible object. We can observe that participants often proceed to repair (they reformulate their question/command) because *Siri's* transcription does not match with their 'intent'. The 'misunderstanding' (the 'failure' of the speech recognition system) is instantiating in a visible 'object' i.e. in written signs displayed on the screen. The visualization of speech utterances enables the human participant to stand back from them and examine them in a more abstract way [12].

[12] See also McTear et al. (2016), section 18.3.5.

[13] The data for this paper are drawn from a larger sample of protocols.

[14] In our present study we refer to an array of these screenshots.

[15] Investigating the dynamics between the oral and the written in human-*Siri* communication seems still to be a rather unexplored research topic. Here, we briefly summarize some key ideas. We will look into this relevant issue in a forthcoming paper.

In our study, we observed that the students were particularly faced with challenges of pronunciation; (many times) repairing was indeed related to pronunciation. Proper pronunciation (especially of names) in a second (foreign) language is not evident. Or the students initiated self-repair through reformulating the voice command while focusing on pronunciation, either *Siri* interacted with the users to seek clarification and elicit a new spoken command. Here, on one hand, *Siri* could be considered as a 'learning assistant' as the participants had to investigate and to improve their pronunciation to be understood by *Siri* [16]. But, on the other hand, repairing pronunciation was often a laborious trial-error process as will be evidenced in the examples described below.

We note that in CA, *repair* is itself a mechanism of conversation. Repair is used to ensure that the interaction does not freeze in its place when trouble arises, (…) and that the turn and sequence and activity can progress to possible completion" [10]. In terms of sequential organization of talk, we will show in the following that repair leads to expanded sequences. In the case shown below, we see how the student's recurrent repairs of pronunciation related to *Siri's* speech-recognition 'failing' cause 'annoying' delays in sequence completion i.e. in task accomplishment.

3.3 "Who is Sancho Panza?"

One of the students of the English Second Language classroom (Roberto) opts for the fictional dyad *Sancho Panza and Don Quixote*. As Roberto's mother tongue is Spanish[17], he is interested in gaining knowledge about the famous characters of the Spanish novel.

In the following, we see that his search is a quite tricky matter.

In terms of adjacency pairs (question-answer), the first sequence is smoothly accomplished[18] (*Figure 7*):

Roberto: *Who is Sancho Panza?*

Siri: *OK. I found this:*

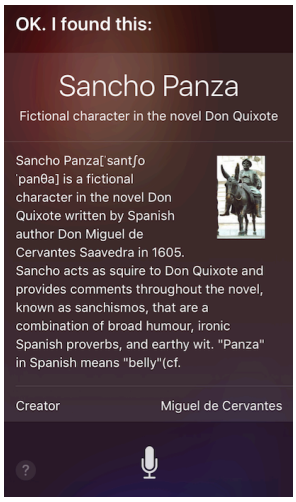

Figure 7

Then, matters will be somewhat different.

Roberto goes ahead to a next voice command by seeking for information about *Don Quixote*. When vocalizing his question "*Who is Don Quixote?*", his pronunciation of the name is quite similar to the Spanish one. And the result is the following (*Figure 8*):

[16] See examples below.

[17] From Roberto's perspective, Luxemburgish, German, French and English are second languages.

[18] Roberto had changed *Siri's* language, he had operated a switch from French to English.

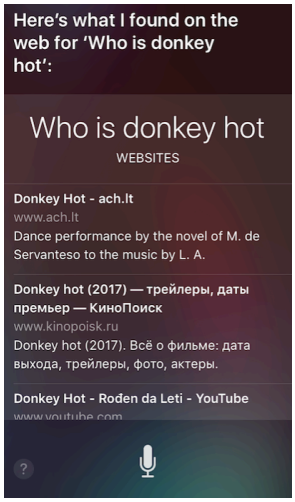

Figure 8

In view of the screen display (*Figure 8*), Roberto infers that his pronunciation was not appropriate in English and proceeds more or less randomly to repair. He tries to vocalize '*Don Quixote*' in a manner he finds consistent with the written word[19]. *Siri's* answer (Figure 9) however gives again clear accounts of 'misunderstanding': the written word '*Dom quickset*' does not match with Roberto's orally expressed 'intent'.

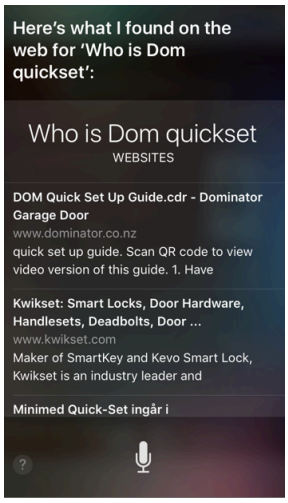

Figure 9

Instead of moving ahead in a trial-error process[20], Roberto adopts a different 'strategy' in order to (re)launch the search. Thus, he 'submits' a 'newly designed' question "*Who is Sancho Panza's master*"[21]. Here, the student infers that *Siri* is able to 'understand' the noun phrase "*Sancho Panza's master*"[22] in its grammatically and semiotically intertwined functioning. Roberto builds on 'the grammatical skills' of the conversational interface as well as on *Siri's* previously displayed resources regarding *the squire Sancho Panza* (*Figure 7*). The student is expecting that, besides (beyond) categorizing the possessive determiner (*'s*) as such, *Siri* will operationalize it by recognizing the semiotic relationship between two items. But once again, the operation fails. We can say that *Siri* gives accounts of 'casting doubt upon' Roberto's pronunciation skills in English (*Figure 10*).

[19] We should point out here that Roberto's vocalization, in consistency with the written word, builds on his multilingual resources (here, especially on *German* language). Note that in Luxemburgish primary schools, the language of literacy is *German*; children engage in reading and writing activities in *German*.

[20] which, at a first glance might seem entertaining but is actually not particularly motivating

[21] Our transcription is in accordance with Roberto's 'intent'!

[22] A noun phrase is a small group of words standing together as a conceptual unit, typically forming a component of a clause.

Subsequently, Roberto's next trial (Figure 11), within the same logical way of proceeding is more 'successful'[23]. *Eureka*! When opening one of the indicated websites (*Figure 11'*), Roberto finds the desired information focusing on *Don Quixote*.

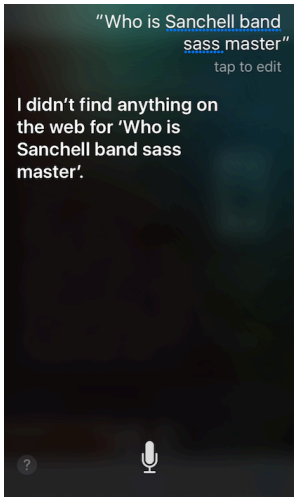

Figure 10

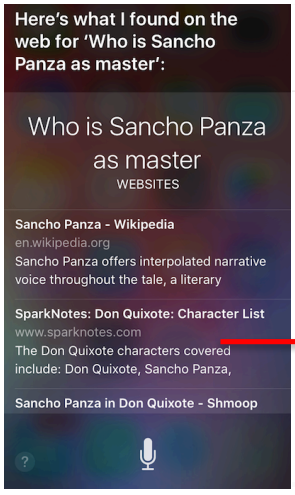

Figure 11

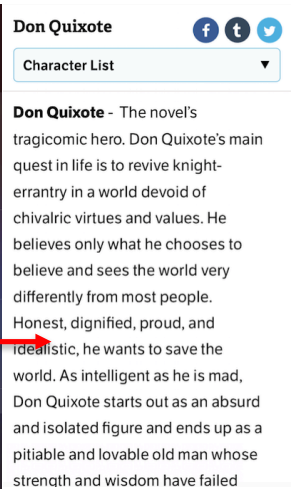

Figure 11'

When presenting his findings to the other students, Roberto points out that he had expected *Siri* to provide a similarly structured 'one-step response' (*Figure 7*) to *both* questions ("*Who is Sancho Panza?, Who is Don Quixote?"*). Through repeated repairs, Roberto mainly acted upon the fact that at some moment a problem had arisen in sequence progression. As an English non-native (novice) speaker, he had to cope with 'accurate' pronunciation. *Siri* however did not provide any repair-assistance regarding appropriate pronunciation. In this context, a possible route to improve *Siri's* performance could be to work on 'implementing multilingualism in action'. *Siri* should be able to identify a proper noun like *Don Quixote* even if the user is a non-native speaker in the 'language in action'.

4 CONCLUSION

Siri can be considered as a performant virtual assistant when mobilized by the human participants in order to activate and operationalize *built-in apps* (maps/navigation, mail, address book, …) or operate on *thematic third party data bases* (restaurants, hotels, ticket reservation…). But, "*we are wary of the status of any descriptions of and ascriptions* [to Siri] *as a conversational partner"* [4]. In particular, as concerns concrete situations in learning contexts, we raise some critical issues (section 3.3). Despite offering some strong key features which students can/could explore and exploit (section 3.2[24]), one of the main challenges, (especially for students), focuses on identifying the appropriate commands and the choice of words that activate each feature. What is the required context-related syntactic and lexical anatomy of a voice-command turn? Further, the user is supposed to have an 'accurate' pronunciation in the registered language (section 3.3).

Conversation analytic driven studies explaining conversational phenomena in terms of recipient design (turn-taking, addressee) can sketch out possible routes forward to improve conversational agents or at least touch on further conceptual challenges that face research in this area [4].

ACKNOWLEDGEMENTS

I would like to express my gratitude to Henri Hansen for his support.

[23] We point out here that *Siri* has visibly not understood/established the inferred grammatical and semiotic links.

[24] *Siri* affords conceptual reflection on 'the interface between the written and the oral'.

REFERENCES

[1] H. Sacks, E.A. Schegloff, & G. Jefferson, “A simplest systematics for the organization of turn taking in conversation”, in *Language*, vol. 50, no 4, pp. 596-735, 1974.

[2] S. Rehal, “Siri - The Intelligent Personal Assistant”, *International Journal of Advanced Research in Computer Engineering and Technology*, vol. 5, no. 6, pp. 2021–2024, 2016.

[3] M. McTear, Z. Callejas, D. Griol, *The Conversational Interface*: *Talking to Smart Devices.* Switzerland: Springer International Publishing, 2016.

[4] St. Reeves, "Some Conversational Challenges of Talking with Machines", *Workshop at the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, Portland, Oregon, USA, 2017.

[5] B. Arend, P. Sunnen, "Coping with Turn-Taking: Investigating Breakdowns in HRI from a CA Perspective", *Proceedings of ICSIT 2017*, pp.149-154, 2017a.

[6] B. Arend, P. Sunnen, "Investigating Breakdowns in HRI: A CA Guided Single Case Study of a Human-Robot Communication in a Museum Environment", *International Journal of Mechanical and Mechatronics Engineering*, vol. 11, no. 7, pp. 934-940, 2017b.

[7] B. Arend, P. Sunnen, "Multilingualism in Action: A Conversation Analytic View on how Children Are Re-Voicing a Story in a French Second Language Learning Lesson", *The IAFOR Journal of Language Learning, vol. 3, no. 2,* pp. 19-31, 2018.

[8] A. Deppermann, H. Blühdorn, "Negation als Verfahren des Adressatenzuschnitts: Verstehenssteuerung durch Interpretationsrestriktionen", *Deutsche Sprache: Zeitschrift für Theorie, Praxis, Dokumentation,* vol. 13, no. 1, pp. 6-30, 2013.

[9] J. Streeck, C. Goodwin, C. LeBaron, C. (eds.), *Embodied Interaction, Language and Body in the Material World*. Cambridge: Cambridge University Press, 2011.

[10] E.A. Schegloff, *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge: Cambridge University Press, 2007.

[11] Ch. Heath, “Talk and recipiency: sequential organization in speech and body movement”, in *Structures of Social Action* (M. Atkinson & J. Heritage eds.), pp. 247-265. Cambridge: Cambridge University Press, 1984.

[12] J. Goody, *The Interface Between the Written and the Oral.* Cambridge: Cambridge University Press, 1987.

[13] D.R. Olson, “History of Schools and Writing”, in *Handbook of Research on Writing* (Ch. Bazerman ed.), pp. 283-292. New York, London: Routledge, 2010.

[14] M. Kalantzis, B. Cope, *Literacies*. Cambridge: Cambridge University Press, 2012.

Corresponding authors: Béatrice Arend in collaboration with Henri Hansen

Email: beatrice.arend@uni.lu