HEY *SIRI*, YOU ARE CHALLENGING THE INTERFACE BETWEEN THE ORAL AND THE WRITTEN. SOME BASIC REFLECTIONS ON VOICE-CONTROLLED NATURAL LANGUAGE HUMAN-*SIRI* TALK

Béatrice Arend, Pierre Fixmer

University of Luxembourg (LUXEMBOURG)

Abstract

The conversational interface *Siri* enables users to perform tasks on a smart device through natural language voice commands (such as sending a message, setting an alarm, getting targeted information). In addition to a vocal reply, *Siri* simultaneously provides a transcription of the user's and its 'own' oral speech (as well as access, where appropriate, to relevant websites). Moreover, the written utterances are delivered with respect to the lexical and grammatical spelling rules of the operating language, and punctuation marks are made in accordance with syntactical structure. Through the visuospatial display, language performing and hence understanding are laid out before the user's eyes and become a potential object of reflection. As the user gets situated visual access to his vocal command i.e. to *Siri's* '*doing* understanding' of the command, he has the opportunity to monitor and to assess this understanding as well as to make assumptions about next steps (repair, repetition, different pronunciation). The written text 'seen' as an object of investigation invites the user to initiate (or not) a new vocal command. We consider *Siri* as a tool to elicit and enhance language performing as well as to trigger reflection on the written word.

From a social scientist perspective, we will point out how human-*Siri* talk raises conceptual challenges relevant to the interface between the oral and the written language. Furthermore, our paper seeks to take a closer look at the potential of the conversational interface *Siri* to support *knowing a natural language.*

Keywords: *Siri*, conversational interface, social scientist perspective, *knowing* a natural language, conceptual challenges.

1 INTRODUCTION

Investigating the dynamics between the oral and the written in human-*Siri* communication seems still to be a rather unexplored research topic. Relying on a sociocultural perspective on language and learning [1, 2, 3, 4], our paper aims at launching a discussion on how the use of the conversational interface *Siri* prompts reflection on the complex dynamics between oral and written language. We will sketch out some conceptual issues arising from *Siri's* transcription-performance. Since human-*Siri* 'conversation' is instantiating in time *and* space, we will focus our attention on the visuospatial dimension of talking to/with Siri.

We will not argue by exploring the mechanism of writing *versus* oral language [5]; our purpose is to touch on some 'pivotal' aspects related to the interactional phenomenon of human-*Siri* talk actualizing in (synchronized) oral and written instances.

In the history of research on oral and written language, sociocultural theory inspired scientists 'handle' written language as a tool which requires voluntary and intentional efforts; the written text is considered as addressing a distant 'other' in quite different conditions from oral communication. So, what about *Siri's* writing performance? Even if the 'voice first device'[1] applies to human-*Siri* interaction modus, the written word is omnipresent. *Siri's* ability to provide the respective transcripts synchronously with the user's vocal 'intents' or with 'its own' oral utterances undoubtedly deserves investigations on dynamic interrelations between oral and written language, in terms of "awareness and control" [4].

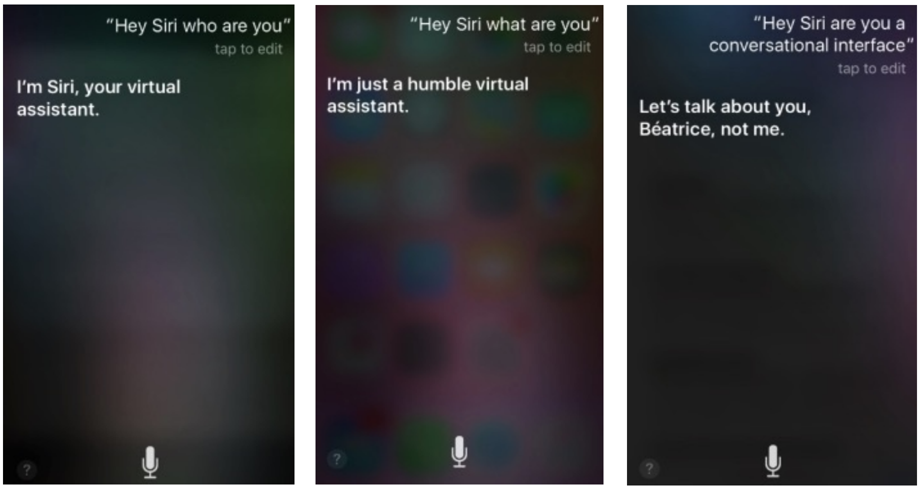
In the following, with regard to *Siri's* speaking and writing abilities [6, 7, 8, 9], we address some fundamental issues related to using *Siri* at the interface between the oral and the written, in order to

[1] i.e. 'the spoken word first device'

trigger (and pursue) exploratory discussions about how talking to *Siri* relies on and can act upon *knowing*[2] natural language.

We should note here that we consider natural language use (in its oral and written occurrences) as a social phenomenon, in accordance with our epistemological stance on language and learning [10, 11]. Furthermore, *knowing* is considered in its social nature and closely related to the intersubjective dynamics of communicative processes. In this vein, we look at *natural language* as articulated in speaking, reading and writing (abilities).

2 HEY SIRI, WHO/WHAT ARE YOU?



"Hey Siri who are you"
tap to edit
I'm Siri, your virtual assistant.

"Hey Siri what are you"
tap to edit
I'm just a humble virtual assistant.

"Hey Siri are you a conversational interface"
tap to edit
Let's talk about you, Béatrice, not me.

Figure 1.

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 [7, 9]. As McTear et al. [7] 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*[3]. 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 [7]. *Siri* enables the user to do things with (orally uttered) words[4]. 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]. In that sense, the power of the spoken word seems undeniable.

Thus, according to our purpose, with regard to both the oral and the written word, our investigations in this section will focus on 'voice control' and 'natural language' i.e. on *Siri's* key characteristics allowing users to engage in a diverse range of operations through natural language voice commands.

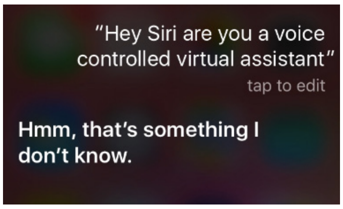
[2] We prefer the term *knowing* rather than the term *knowledge*. *Knowing* captures our preoccupation with investigating how natural language is used and locally managed (most of the time effectively and skillfully) in the process of human-*Siri* talk.

[3] The name 'Siri' is actually an acronym; it stands for 'Speech Interpretation and Recognition Interface'. *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). *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.

[4] In the sense of J. L. Austin (1962) *How to do things with words*.

[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*.

2.1 ‘Voice-controlled’ human-*Siri* talk: ‘Augmented’ voicing!?



“Hey Siri are you a voice controlled virtual assistant”

Hmm, that’s something I don’t know.

Figure 2.

As we mentioned before, *Siri* is a built-in, voice-controlled virtual assistant available for Apple users. The idea is that users talk to *Siri* as they would to a human assistant and *Siri* aims to help them get things done, whether that be making a dinner reservation or sending a message. Through (*Siri*) addressed voice commands the user can consult information or perform tasks. The user’s voice gets things done or moving ahead, *Siri* makes it (them) possible: emails reach the addressees, booking is made and so forth. According to Austin [12], we may say that, from the user’s perspective, “to utter the sentence is to do it (…) the issuing of the utterance is the performing of an action, it is not thought of as just saying something”.

Moreover, besides mobilizing *Siri* through his/her voice, the user can actually *see* ‘the voice-control’. The display screen of the smart device provides visual access to the performance flow. Here, we will particularly point at the written word as the in situ materializing ‘token’ of performative voicing. The spoken words as well as their impact are laid out before the user’s eyes through simultaneous transcription or relevant websites. In this voicing[6] process, *Siri* can be considered as a tool ‘augmenting’ the user’s spoken word which is visibly acting (operating).

It is precisely here that we should pay closer attention to the concept of *voice*. In reference to a bakhtinian view revised by Linell [13], the concept of *voice* involves at least three dimensions, “material or physical embodiment, personal signature, and perspectives on topics and issues”. Furthermore, a person’s voice is considered as his speaking consciousness having a will or desire behind it and having its own timbre or overtones. These properties contribute to sense-making in communication and reflect the values behind the consciousness which speaks. Thus, the term *voice* can stand as a metaphor for “namely an expressed opinion, view or perspective, something that a person would typically say and stand for” [13]. Insofar as we consider *voicing* as a social phenomenon related to and enacted by *human* beings, the assumption that personal voice and consciousness are thoroughly relational seems clear and relevant. But, what about *Siri’s* voice?

Human users might have the impression to talk to/with another human when mobilizing and listening to *Siri*: *Siri’s* voice[7] does not only sound humanlike, we actually hear a human’s voice! But, we should cast serious doubt on the virtual assistant’s ‘speaking consciousness’. *Siri* is ‘only’ a built-in *Speech Interpretation and Recognition Interface*, and voice actors (female and male) lend their respective voice to *Siri*. This is notably the case for Susan Bennett who provides the first female American English voice for *Siri*: during a TV-interview, she talked about “the first time I heard my voice as *Siri*”[8]. Nevertheless, human users greatly appreciate the humanizing and in some ways entertaining voice timbres of the interface.

We note also that, in order to assist the user, *Siri* refers to countless ‘voices’ in the available database, among others to the voices of multiple diverse websites and Apple services. *Siri* responds to voice commands by providing a wide range of more or less pertinent written words. And, besides the issue of authorship and reliability of sources, drawing on the voices of the worldwide web often goes together with unpredictability[9].

Even though *Siri* is a sophisticated voice-activated tool, an efficient voice-controlled virtual assistant, the user should not underestimate the risk that the called voices might not be ‘under control’.

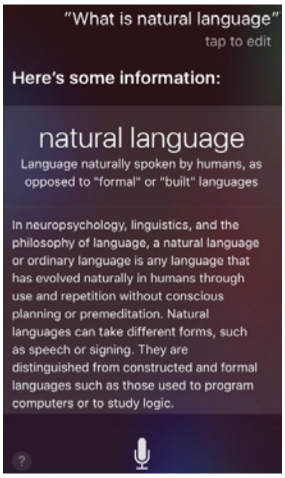
[6] ‘doing things with words’ process

[7] actually we should say, *Siri’s voice**s***

[8] https://www.youtube.com/watch?v=bL1tSgKrcT0

[9] See Arend (2018)

2.2 ‘Natural language’ human-*Siri* talk: Speaking, transcribing and reading.



“What is natural language”

tap to edit

Here’s some information:

natural language

Language naturally spoken by humans, as opposed to "formal" or "built" languages

In neuropsychology, linguistics, and the philosophy of language, a natural language or ordinary language is any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation. Natural languages can take different forms, such as speech or signing. They are distinguished from constructed and formal languages such as those used to program computers or to study logic.

Figure 3.

Users can talk to *Siri* in a natural way[10]. ‘Natural language’ is one of *Siri’s* highly valued key characteristics: an account of advanced technology in conversation as well as sometimes a source of amusement for users. In this subsection, we sketch out how *knowing* a natural language and talking to *Siri* can be considered as interrelated in terms of speaking, transcribing and reading.

According to Tomasello [3], natural language is composed of lexical and syntactic symbols shaped by social-communicative practice. People display *knowing* a natural language through *talking*, as well as through *reading* and *writing* the symbols in the context of social events. The development of natural language skills and of communicative practice as a whole largely draws on feedback about communicative efficacy. The feedback can be used to make further inferences about the meaning and the significance of words.

Human-*Siri* communication is considered as efficient when *Siri* is providing accurate support. To that end, *Siri* basically requires an appropriate oral language performance from the user. Sometimes however, users are faced with challenges of proper pronunciation to be ‘understood’ [6]. In the case of that scenario, *Siri’s* transcription performance, consisting in synchronically transforming in written language the user’s oral utterances as well as *Siri’s* oral replies, appears to be a key property in terms of feedback. Through the visuospatial display of the transcribed word, the user gets synchronized visual access, feedback, to his oral language performing. Thus, he has the opportunity to monitor and to assess his vocal command i.e. *Siri’s* ‘understanding’ of the command instantiated in transcription. Furthermore, he can make assumptions about next steps, about how ‘adapting’ his voice. The user encounters his pronunciation skills as a visible object and creates links between the oral and the written.

According to Vygotsky [4], with writing, language becomes transparent as an object of thought. Writing brings awareness to speech. In this line, Olson [2] argues that writing turns language from a means of communication into an object to think about. He claims that, additionally to speaking, reading and writing provide a new consciousness of language; most verbalizable thoughts and intents are about objects and events. “But others, more important to my argument, are thoughts about the language itself” [2]. In that sense, in human-*Siri* talk, the user’s transcribed voice command becomes an object of thought in its written form. Moreover, the transcribed word is also a visible account of *Siri’s* speech recognition ability and thus becomes an object of technology-focused inquiry and assessment with regard to communicative efficacy.

Referring to the assumption that *knowing* a natural language includes “awareness of the language, the ability to think about language” [2], we assert that mobilizing *Siri* can be considered as engaging in *knowing a natural language*. The user is asked to enter into a dynamic speaking-reading process at the interface between the oral and the written. While addressing *Siri*, he is simultaneously the addressee of his/her own words displayed in transcribed form and is faced with them in terms of

[10] *Siri* can currently rely on 21 built-in languages.

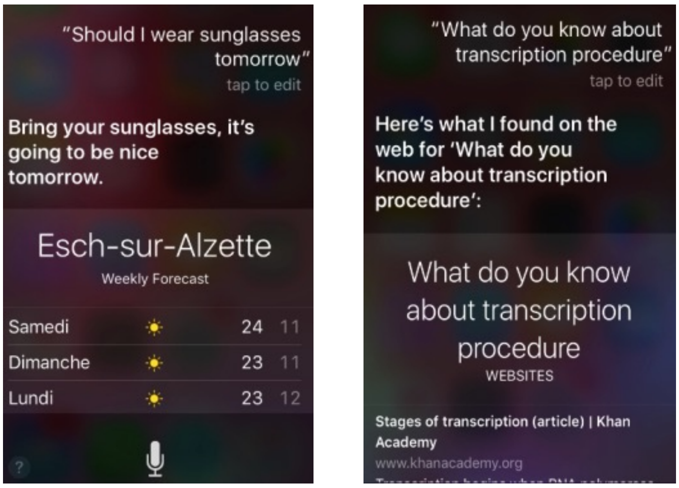
meaning, significance and task performing. Thus, ‘awareness of the language’ means also becoming aware that *providing an utterance is performing an action*. Conversation *is* action: “how utterances produced in conversation should be viewed as actions that the speakers carry out in order to achieve their goals and how addressees interpret these actions” [7].

In this complex process, the screen of the smart device provides a window into how users and *Siri* are *doing* knowing a natural language: into what users and *Siri* ‘know’ and how this knowledge is treated by the human participant (voicing, reading) as well as by the conversational interface (relying on a speech recognition system and on available data bases).

3 HUMAN-SIRI TALK: CROSSING THE BOUNDARIES BETWEEN SPEAKING AND WRITING.

In the previous sections, we point out to which extend oral words are performative in human-*Siri* talk. As also mentioned before, written language, more particularly transcription, is an inherent feature of *Siri’s* speech recognition apparatus. *Siri’s* transcribing performance has both “obvious utilitarian advantages, such as record keeping (…) as well as conceptual advantages through creating new objects such as words and sentences to think about” [14]. It is crucial to emphasize what this process of simultaneous transcription involves additionally to representing speech on a material two-dimensional surface, all the more since researchers in linguistics consider transcription as “a massive project of translation” [2]. The main issue is “what are the features of sound organized in time that are to be depicted by marks organized in a two-dimensional space?” [2]. It is generally agreed that transcription is a limited representation of speech. When transcribed, speech is usually deprived of most of its musical dimensions: stress, tones and tunes, rhythm, tempo [11]. Choices are made how to preserve this dynamic distributed stream of speech and above all the intended meaning. The orthography and punctuation, the lexical and syntactic structure of the written form are supposed to do what tone and emphasis could have done in speaking.

With regard to our concerns, we can assert that transcription assigns a spatially organized segmental structure to human-*Siri* talk. Voice commands are ‘translated’ i.e. grammaticalized and edited in structural units, the same applies to *Siri’s* answers (see screenshots).



“Should I wear sunglasses tomorrow”
tap to edit

Bring your sunglasses, it’s going to be nice tomorrow.

Esch-sur-Alzette
Weekly Forecast

Samedi	☀	24	11
Dimanche	☀	23	11
Lundi	☀	23	12

“What do you know about transcription procedure”
tap to edit

Here’s what I found on the web for ‘What do you know about transcription procedure’:

What do you know about transcription procedure
WEBSITES

Stages of transcription (article) | Khan Academy
www.khanacademy.org

Figure 4.

Thus, for instance, in the above provided screenshots, we can see that the user’s transcribed speech is marked by double style inverted commas. *Siri* applies here conventional quotation rules to signal direct speech i.e. the user’s voice. Single style inverted commas are used in *Siri’s* reply to repeat (quote) the user’s previously uttered voice command. Furthermore, the spelling is correct, the usage of capital and lowercase characters complies with the standards.

The transcription modus in human-*Siri* talk should not be considered as reducing, but rather as an added value with regard to understanding and language awareness (see section 2.2). Moreover, as the spoken word *and* its transcription are simultaneously uttered, the written is animated by intonation and timbre. Actually, we hear and see a dynamic in real time occurring cross-over between the oral and the written[11].

In this vein, with regard to *knowing* a natural language, *Siri's* transcription function appears to be well suited for the purpose of supporting *knowing* writing (especially in a second language). The transcribed form of voice increases the user's potentialities of approaching literacy (reading and writing) as it gives a synchronously laid out access to oral discourse.

4 PERSPECTIVES

As we have pointed out in the previous sections, the user can 'wake up' *Siri* by his/her voice[12] and thus trigger human-*Siri* talk in order to perform a task or get targeted information. Besides being a rather efficient tool which can assist the user by 'rendering commanded services', *Siri* appears also to be a potential assistant in terms of *knowing* a natural language. On one hand, the user's voice 'input' is treated[13] and generates an audible as well as a synchronously displayed visible 'reply' (i.e., most of the time, delivery of the requested information or completion of the requested task)[14]. On the other hand, *Siri's* reply, more particularly the transcribed form of *Siri's* answer, gives feedback about the user's and *Siri's knowing* a natural language (see sections 2.2, 3). Here, we touch on 'pivotal' aspects related to the interface between the oral and the written. When mobilizing *Siri*, the human user is engaging in a dynamic *voicing-transcribing-reading* process. At the same time, he/she is both challenging and augmenting his/her representations about the oral and the written.

Even though, in view of the mentioned characteristics, we might think about how to use *Siri* in educational language learning contexts, we are wary of any ascription to *Siri* as a virtual *teaching-learning* assistant [6]. Of course, a child-addressed *Siri*-like tool could support young language learners to develop language awareness in a vygotskian sense as "what children have to learn is attention to the language" [4]. But, unpredictability and 'control' (see section 2.1) are still important and relevant matters in human-*Siri* talk and should in that case be given priority with regard to deontological issues. *Siri* is a performant tool in order to activate built-in apps or operate on thematic third party data bases as well as to enhance attention to and reflection on *the words themselves* [1]. But, *Siri* cannot[15] be considered as 'a substitute teacher' nor as a conversational partner [9].

This paper is supposed to launch further discussions about *Siri* and co. as well as about 'augmented voices' challenging the boundaries between the 'virtual' and the 'real' space[16].

REFERENCES

[1] J.S. Bruner, *In Search of Pedagogy. Volume II*. London, New York: Routledge, 2006.

[2] D.R. Olson, *The Mind on Paper.* Cambridge: Cambridge University Press, 2016.

[3] M. Tomasello, *Die Ursprünge der menschlichen Kommunikation*. Frankfurt am Main: Suhrkamp, 2009.

[4] L. Vygotsky, *Thought and Language*. Cambridge: MIT Press, 1986.

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

[6] B. Arend, "Hey *Siri*, what can I tell about Sancho Panza in my presentation? Investigating Siri as a virtual assistant in a learning context? ", *Proceedings of INTED 2018*, Spain, pp.854-863, 2018.

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

[11] Of course, we should not deny failures in *Siri's* speech recognition system (see Reeves, 2017; Arend, 2018)

[12] In this paper, we do not focus our investigations on the tactile dimension of using a smart device (touching, scrolling).

[13] Within *Siri's* abilities (relying on speech recognition system, context awareness, configured apps, other built-in services).

[14] Note that *Siri's* assistance can be provided in a step by step procedure.

[15] Not yet ?

[16] upcoming paper

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

[9] 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.

[10] M. Bakhtin, *Toward a Philosophy of the Act*. USA: University of Texas Press, 1993.

[11] P. Linell, *The Written Language Bias in Linguistics*. New York: Routledge, 2005.

[12] J.L. Austin, *How to Do Things with Words*. Oxford: Oxford University Press, 1962.

[13] P. Linell, *Rethinking Language, Mind, and World Dialogically*. USA: Information Age Publishing, Inc., 2009.

[14] D.R. Olson, M. Dascal, "Writing and the mind", *Pragmatics and Cognition*, vol. 21, no. 3, pp.425-pp.430, 2013.