POMPEU FABRA UNIVERSITY

TRIGGERING LINGUISTIC REPRESENTATION OVER SINE-WAVE SPEECH

Master's Thesis

July 2015

Author: Jasna Čakarun

Mentor: Prof. Juan Manuel Toro Soto

MSc in Brain and Cognition

Pompeu Fabra University, Barcelona

TABLE OF CONTENTS

Abstract 3
Introduction 4
Method 10
Condition 1: Language 10
Condition 2: Non-Language 13
Results 15
Discussion 17
References 20
Appendix 21

LIST OF FIGURES

Fig.1. Example of audio-visual clip used in Condition 1 11
Fig.2. Graphic outline of the experimental procedure 133
Fig.3. Example of audio-visual clip used in Condition 2 14
Fig.4. Mean of correct answers per participant in Condition 1 vs. Condition 2 16

Abstract

Unlike animals, humans do not process speech merely as a linear stream of varied acoustic stimuli. What distinguishes us is linguistic representation, the tendency to process auditory language input as a hierarchy of linguistic categories with individual functions. An example of this restriction is the functional difference between consonants and vowels, wherein we tend use the former only for lexical retrieval, and the latter only for structural generalization. We tested this in a linguistic and a non-linguistic context, where participants had to extract a structural rule over consonants. The results expectedly confirmed the functional distinction in the linguistic context, in which participants were unsuccessful in rule-learning, but the same was observed also in participants in the non-linguistic context. The possible implication of this might be that the acoustic properties of the stimuli in the non-linguistic context nevertheless cued linguistic representation, which restricted the functional role of the consonants.

Introduction

There is a comic strip by Gary Larson that was published in The Far Side series, which tackles the question of cognition in a particularly amusing way. The first panel, entitled *What We Say to Dogs*, features a man scolding his dog at length for digging through the rubbish, and an apparently unfazed dog. The second panel, entitled *What They Hear*, reveals that the dog has heard nothing but a series of *blah blah blah* with the occasional exception of his name, *Ginger*. To the owner's chagrin, the language he generated to express his frustration appears to have been processed by the dog as a string of acoustic features without content or hierarchy, much like a human would process the sounds of a washing machine. This suggests that there might be an important difference in how our species deals with language in comparison to animals, although the issue of how we deal with non-language is just as important. We based our research on this contrast, and focussed on the human ability to learn structural rules in auditory input, which has also been observed in several other species. By examining the rate of success in rule-learning in both a linguistic and in a non-linguistic context (when we perceive speech vs. when we perceive non-speech), we aimed to provide further evidence for the existence of a uniquely human linguistic representation.

Before we concentrate on rule-learning within and without language, let us first define linguistic representation, the crucial term for our discussion. We understand linguistic representation as the human tendency, and ability, to process certain auditory information as language and thus ascribe it categorical linguistic properties. Once an incoming stream of auditory stimuli triggers linguistic representation, we mentally place it into a context of language and process it as speech rather than a sequence of random sounds. Linguistic representation is triggered by the universally recognizable features of speech, from phonetic to semantic, which we readily detect in auditory input. These features are encoded as linguistic categories in a hierarchical abstract system, which appears to be deeply rooted in our brain: it determines what we detect as speech, and how we process its constituents. Whether this hierarchy of linguistic categories is innate or acquired through language input has been subject of much debate,[1] but it is not the aim of this thesis to provide evidence for either theory. We will focus specifically on the categories of vowels and consonants, and their roles in the context of language (i.e. as linguistic representations), and outside it.

Both in terms of quality and quantity, vowels and consonants differ in several aspects. When it comes to phonetic quality, one of the most significant points of divergence between the two categories is their relative distinctive power (Nespor, Peña, & Mehler, 2003). Distinctive power is,

[1] We are referring to the well-known and long-standing debate between the supporters of linguistic nativism with N. Chomsky and S. Pinker as its most prominent champions, and the advocates of linguistic empiricism, e.g. G. Sampson and M. Tomasello.

both in consonants and in vowels, closely linked to their acoustic variability. Consonants, which constitute the morphological roots of words and carry lexical information, generally show minimal changes in their phonetic properties, and thus have better distinctive power than vowels. These, on the other hand, have little distinctive power as they often alter their acoustic attributes by undergoing harmonization, stress-dependent reduction and similar changes. Nespor, Peña, and Mehler (2003) extend this critical divergence in phonetic quality to an analogous divergence in linguistic function: while the relatively unvarying consonants contain lexical information that enables word retrieval, the highly changeable vowels are the carriers of prosody, which provides cues for determining syntactic constituents and the grammatical regularities that govern them.
In terms of difference in quantity, consonants generally outnumber vowels, with only a handful of language exceptions. It has been postulated, however, that the consonant-to-vowel ratio does not affect the functional distinction between the two categories: several studies have in fact demonstrated that in French, which has a nearly equal number of consonants and vowels, subjects were as unsuccessful in identifying lexical items over vowels than they were in gauging structural regularities over consonants (Bonatti, Peña, Nespor, & Mehler, 2007).

In experiments investigating speech processing, vowels and consonants have indeed been shown to perform markedly different functions. Toro, Nespor, Mehler, and Bonatti (2008) demonstrated that during word segmentation from an artificial continuous speech stream, experimental subjects delegated distinct tasks to the two categories: they used only consonants to extract words out of the auditory input, whereas its structural regularities were computed exclusively over vowels. The authors linked this functional distinction between vowels and consonants to their different linguistic categorization. Thus it might follow that different acoustic properties lead to different categorization, and this in turn establishes divergence in function. The latter is reflected in speech processing as the use of different operations, as evidenced by Bonatti, Peña, Nespor, & Mehler (2005). The authors identified an important categorical divergence in speech-processing: while transitional probabilities were computed automatically over nonadjacent consonants, the same operation was not observed in vowels, which appeared to be exploited only for discovering structural patterns. The explanation for this is based on the idea that the ability to generalize an abstract structure, such as a constituent in the hierarchy of syntax, cannot be reduced to a calculation of statistical dependencies (Hochmann, Benavides-Varela, Nespor & Mehler, 2011). The two categories therefore drive different computations, and these are in turn associated with different levels of processing: lexical in the case of consonants, and grammatical (structural) in the case of vowels.

The above studies have thus shown that subjects tend to use relations among vowels to extract generalizations in a speech stream, but at the same time disregard the same relations among

consonants, which are in turn used for lexical identification. This idea is known as the Symmetry Hypothesis, and has been tested several times. Having consonants instead of vowels as carriers of an abstract rule (e.g. ABA or AAA) resulted in failure in making a structural generalization, even when units were separated by short pauses (Toro, Nespor, Mehler, & Bonatti, 2008). This gave new support to the theory that consonants and vowels are linked to particular linguistic functions for which they carry specific information, and are thus not interchangeable, multi-purpose elements. This idea has coexisted with the hypothesis that the two categories might perform different linguistic functions because of the difference in their acoustic salience, not because of their inherently different quality. If this is the case, the category we favour in speech processing will invariably be the one with the most acoustic prominence. To test this hypothesis, Toro, Shukla, Nespor, and Endress (2008) carried out an experiment in which they manipulated the salience of the vowels and consonants in the stimuli that followed a specific structural rule, which subjects had to extract. The authors found that subjects could not successfully discern the underlying structural rule over consonants, not even after these were made much more acoustically prominent as compared to vowels, or when sonorants were used, or when vowels were omitted altogether.[2] This was taken as evidence that the previously observed functional difference between vowels and consonants was not a mere by-product of acoustic or perceptual differences between the two categories.

The tendency to employ vowels and consonants for different functions in speech processing appears to emerge very early in life. Pons and Toro (2010) have shown that the preference to extract a structural rule over vowels rather than over consonants is evident as early as at 11 months. The rates of success measured in the tested infants corresponded entirely to those measured in adults, upon which the authors concluded that the functional distinction between the two categories does not require years of language input before it can be put to practical use. Testing for the same tendency in 12-month-olds, Hochmann, Benavides-Varela, Nespor and Mehler (2011) provided further evidence for the above "division of labour" between vowels and consonants.[3] At approximately one year, when infants start building the lexicon, they are thus fully able to exploit this functional difference, simultaneously acquiring new words and discovering syntactic relations. It should be noted, however, that 11 months mark an important transition: around that age, infants appear to switch from a vowel

[2] When vowels were removed from the structure, the results in fact showed marginal evidence of structural generalization over the remaining consonants, for which the authors proposed two explanations. The observed effect might have emerged because the consonants took on the roles of the missing vowels, functioning as the syllabic nuclei, which could be the carriers of structural generalization. Alternatively, the participants might have started perceiving the all-consonantal structures as non-language, and thus showed some ability to extract a rule over them in the absence of linguistic representation.

[3] Hochmann et al. replicated the experimental structure seen in Pons and Toro (2010), but used different consonants and vowels in the familiarization and the test phase.

bias to a consonant bias in speech processing. The early tendency to pay more attention to vowels enables them to establish a basis for language learning: using prosodic cues, they determine syntactic boundaries, discover the syllabic repertoire and recognize several other key properties of the architecture. After the eleventh month, when the structural foundation has been laid, infants start with word acquisition by moving the focus of their attention from vocalic to consonantal sounds. This generates a new kind of analysis based on the quality of consonants and their statistical interrelations, which leads infants to develop the ability to recognize and acquire lexical items with remarkable speed and ease.

Infant studies thus suggest that linguistic representation, although based on a set of abstract categories, does not require mature cognition in order to emerge. Instead, it appears to be present since early infancy and persist throughout life, barring the impact of certain cognitive impairments. The universality of language and our apparent predisposition toward it has given rise to the idea that it is an exclusively human faculty, and therefore clearly different from what we designate as "language" in other animals. This distinction is based on our ability of highly complex abstract thought, which has not been shown in any other animal. The human language faculty hinges upon this ability, and the emergence of linguistic representation has been interpreted as a reflection of an elaborate mental system of linguistic categories and rules, which other species appear to lack. There is a considerable body of research on how animals process the structure of language input: these studies, which are mostly conducted on birds and monkeys, focus on the distinctions as well as on the potential similarities between speech perception between humans and other animals. Through different approaches, they address the central question whether animals can operate with abstract categories, and if so, to what extent the categories observed in animals are comparable to the ones characterizing language representation in humans (Yip, 2006).

These issues were addressed by de la Mora and Toro (2013), who focussed on the key feature of linguistic representation, the functional distinction between linguistic categories, and observed for it in appropriately trained rats. The tested categories were vowels and consonants, and the experimental aim was the same as in the analogous human studies: to establish whether rats, too, are unable to extract the structural rule of a speech stream when the rule is presented over consonants, as compared to vowels. The results indicated that the animals were equally successful in both, and did not show the human tendency to exploit only vowels when recognizing structural generalizations. The authors proposed that this divergence stems from the crucial difference between how humans and animals process language: while in humans the acoustic difference between consonants and vowels serves to define these as linguistic categories with different

functions, animals perceive them merely as acoustic qualities of the input.[4] The separation of consonantal and vocalic sounds into two categories points to the constraining influence of linguistic representation, which is observed only in humans. Animals, on the other hand, do not possess a complex mental system of language, and rely solely on acoustic differences when learning simple structural regularities.

Our deep-seated linguistic system thus greatly influences how we process the sounds that surround us. It primes us to scan auditory input for specific cues that identify it as language, which we segment into discrete categories and establish hierarchical relations between them. Because we are cognitively equipped to do so, we become skilled in processing and producing language soon after we are born, and it becomes, with its limitations, one of the important media through which we understand our environment. The fact that people require very few linguistic properties to be present in acoustic input in order to perceive it as speech points to how entrenched our species is in language, and how crucial it is for our communication. Animals exhibit different means of interaction, and perceive, as shown above, all acoustic stimuli as non-language, i.e. as input unconstrained by top-down bias. This distinction raises an important issue: can humans do the same for non-linguistic acoustic input? We are constantly exposed to it, from sounds produced by electronic devices, to those generated by machines and automobiles. We perceive and process these stimuli; however, given the pervasive influence of language representation, can humans process such non-linguistic input without breaking it into categories with different functions? We decided to answer this question by comparing the processing of natural and sine-wave speech.

Sine-wave speech is produced by a modification of natural speech. It involves the replacement of three or four formant frequencies of the natural speech input with pure tones showing a sinusoidal pattern (Remez, Rubin, Pisoni, & Carrell, 1981). Because of this radical change, the resulting sine-wave signal lacks period and formant structure, and thus does not provide the cues traditionally viewed as the defining properties of speech: it does not convey manner and place of articulation, consonant voicing, colour or stress. Sine-wave stimuli are in fact mostly described as “computer sounds,” “space signals” and similar non-human acoustic expressions. Interestingly, however, when explicitly labelled as language or given within a linguistic context, they tend to be perceived as language, and the modified structure of the original natural speech can be recovered. This was demonstrated with simple context manipulation (Remez, Rubin, Pisoni, & Carrell, 1981), and through

[4] It should be noted, however, that absence of an abstract hierarchy in auditory input does not necessarily mean that all sounds are perceived as equal. Their acoustic differences may serve some animals to discriminate between various sounds and the ways they can be employed. For example, Newport, Hauser, Spaepen, and Aslin (2004) have shown that tamarin monkeys display a pattern opposite to ours: they can compute transitional probabilities over vowels, but not over consonants.

audio-visual matching aimed to trigger the McGurk effect (Tuomainen, Andersen, Tiippana, & Sams, 2005; Vroomen & Bart, 2009). The last two studies have shown that audio-visual integration of sine-wave stimuli and consequently the McGurk effect were evident only in the subjects that were tested in the "speech mode" (i.e. those that were explicitly told the stimuli were language). No such effects were observed in the opposite "non-speech mode," where the sine-wave stimuli were labelled simply as "auditory stimuli." This led the authors to the conclusion that the speech mode, which in our terminology corresponds to linguistic representation, selectively enhances only the acoustic (and visual) stimuli that are relevant for phonetic perception (Tuomainen, Andersen, Tiippana, & Sams, 2005). This supports the idea of the constraining effect of the linguistic system in the context of language, and provides a good example of the use of sine-wave speech in creating a context of non-language where linguistic restrictions do not apply.

Our own experimental design was similar. In order to examine the influence of linguistic representation in sine-wave speech processing, we created two conditions: language, and non-language. In both, the sine-wave stimuli were presented as the auditory part of a series of audio-visual clips. Since we were not testing audio-visual integration or perceptual accommodation shown in the above studies as the McGurk effect, we had the sine-wave stimuli perfectly match the visual content. In the language condition, we created a linguistic context by playing the sine-wave stimuli over visual clips showing a mouth uttering them (these were in fact utterances of the original, natural stimuli before the conversion). In the non-language condition, we created a non-linguistic context by playing the same sine-wave stimuli over visual clips showing a hand pressing buttons and flipping switches on a large control panel, where the hand motions were in synchrony with the sounds. There were 36 different sine-wave stimuli presented in the familiarization phase of each condition, and 24 more in the test phase. The stimuli were trisyllabic nonsense words whose structural rule was given over consonants. After a period of exposure to the stimuli consisting of three consonants and three vowels, the subjects were tested for rule-recognition in stimuli that included a new set of consonants, but followed the same pattern.

Our aim was to compare the rate of success in rule-learning over consonants both in a linguistic and in a non-linguistic context. We formulated our hypothesis based on the experimental results and theoretical conclusions from the studies discussed above: if sine-wave speech stimuli are presented as sounds generated by a machine, the linguistic representation will not be triggered. In such case, subject will not be constrained by linguistic categories and their prescribed functions, and will be able, much like de la Mora and Toro's rats (2010), to extract structural rules from the acoustic input both over vowels and over consonants. Thus we expected participants to successfully learn the

consonant-implemented rule in the non-language condition, and participants in the language condition to fail.

Method

We will present the two conditions, and all the relevant information pertaining to each, in two separate sections.

CONDITION 1: LANGUAGE

In this condition, we created a linguistic context in which the participants were exposed to a series of sine-wave stimuli featured in audio-visual clips, and were finally tested to gauge whether they had implicitly learned the rule given over the consonants in the stimuli.

Participants

The participants were 16 (8 male, 8 female) students from Pompeu Fabra University. Most were native speakers of Spanish or Catalan, or bilingual, and all were competent speakers of English, as was stated in their personal data. All participants were selected from the subject database of the Centre for Brain and Cognition at Pompeu Fabra Universty, and were compensated with 5,00 euros for their participation.

Materials

The stimuli were 60 trisyllabic nonsense words (henceforth *words*) converted into sine-wave stimuli. The familiarization phase featured 36 different words containing three consonants, K, L, and S, and three vowels, E, I, and O. The consonants were combined following a structural in which the first two had to be the same, and the third one different (XXY, e.g. KEKILO). Two vowels were inserted between the first and the second consonant, and between the second and the third consonant (X_X_Y). The vowels did not follow a rule but were balanced so that every consonant triad had the same six vowel combinations (e.g. LOLEKI, LOLIKE, LELOKI, LELIKO, LILOKE, LILEKO).

The stimuli in the test phase were composed of two consonants, P and M, and the same three vowels, E, I, and O. While the vowels had the same distribution as those in the familiarization phase, the consonants were arranged by the above rule (XXY) only in half of the stimuli. These were designed as the correct answers for the test phase, while the other half represented the incorrect answers. In these, the same consonants were used, but followed the rule XYX (generating, for example, the incorrect PEMIPO instead of the correct PEPIMO).

In both the familiarization and the test phase, the words were arranged into a sequence in which each word was different from the preceding and the following one as much as possible. We avoided linear repetition of the same consonantal or vocalic combination, e.g. by having SISELO followed by LOLEK, rather than by KIKESO or by SISOLE.

The complete list of the experimental words is featured in the Appendix.

The words were read by a female volunteer, who was recorded with an HD camera. The recording was a long audio-visual clip, which was edited in Adobe Premiere Pro: it was cut into single-word segments, and cropped to show only the lower part of the speaker's face. Finally, the audio was removed from the clips in order to undergo conversion into sine-wave speech. This was done through the Praat software (created by P. Boersma and D. Weenink, and freely available at http://www.fon.hum.uva.nl/praat/). Once converted, the sine-wave audio stimuli were superimposed onto their original clips, which generated 60 audio-visual clips of a female mouth articulating the nonsense words in robotic-sounding sine-wave speech.

Fig.1. A still from an example audio-visual clip used in Condition 1. The moving mouth articulating the sine-wave stimuli served as a trigger for linguistic representation, which would result in the participants processing the auditory stimuli as language.

Procedure

The experiment was conducted in a soundproof booth at the laboratory of the Centre for Brain and Cognition at Pompeu Fabra University. It was run in PsyScope installed on a Macintosh computer. The participants listened to the auditory part of the experiment through headphones, and used the keyboard to press the marked relevant keys (1 and 3 for choosing between clips in the test phase, and the space bar for starting the two parts of the experiment). After the participants were explained the stages of the experiment, they were ready to begin. At the start and immediately after the

familiarization phase of the experiment, they were presented with the same explanation in written form.

In the first part, the familiarization phase, the participants viewed a series of 72 audio-visual clips showing a female mouth uttering the nonsense words. This phase included two repetitions of the 36 words featuring the consonants K, L, and S, which were created as described above. All of the sine-wave audio stimuli in the clips followed the XXY consonantal rule. Every clip was followed by a 300 ms pause during which the participants saw a white screen and heard silence. The participants' task was to watch and listen attentively.

The last clip in the familiarization phase was followed by written instructions to the second part, the test phase. This time, the audio-visual clips were presented in pairs, and the test was designed as a two-alternative forced choice task. The participants had to watch and listen attentively, and choose one clip in each pair as the answer to the question given in the instructions before the start of the test phase: "Which of the two clips is more similar to the clips in part 1?" The same question applied to all pairs of clips. There were 12 pairs altogether, containing 24 clips in total, of which half were the correct alternatives featuring words that followed the XXY consonantal rule, and half were the incorrect alternatives featuring words that followed the XYX consonantal rule. The words featured the consonants P and M, and were created as described above. The correct answers through the 12 pairs of clips were arranged so as to avoid a predictable alternation or repetition (the correct answers followed thusly: 1, 3, 3, 1, 3, 1, 1, 3, 3, 1, 3, 1).

In each pair, the two clips were presented with a 300 ms pause between them, during which the participants saw a white screen and heard silence. As the second clip ended, they could answer the test question by pressing 1 to signify that the first presented clip was more similar to those in the familiarization, or 3 to signify that the second presented clip was more similar to those presented in the familiarization. PsyScope recorder reaction times from the end of the second clip, i.e. the time each participant took to answer the question for each pair of clips.

After the participants had chosen a clip in pair 12, the experiment ended. It lasted approximately ten minutes in its entirety.

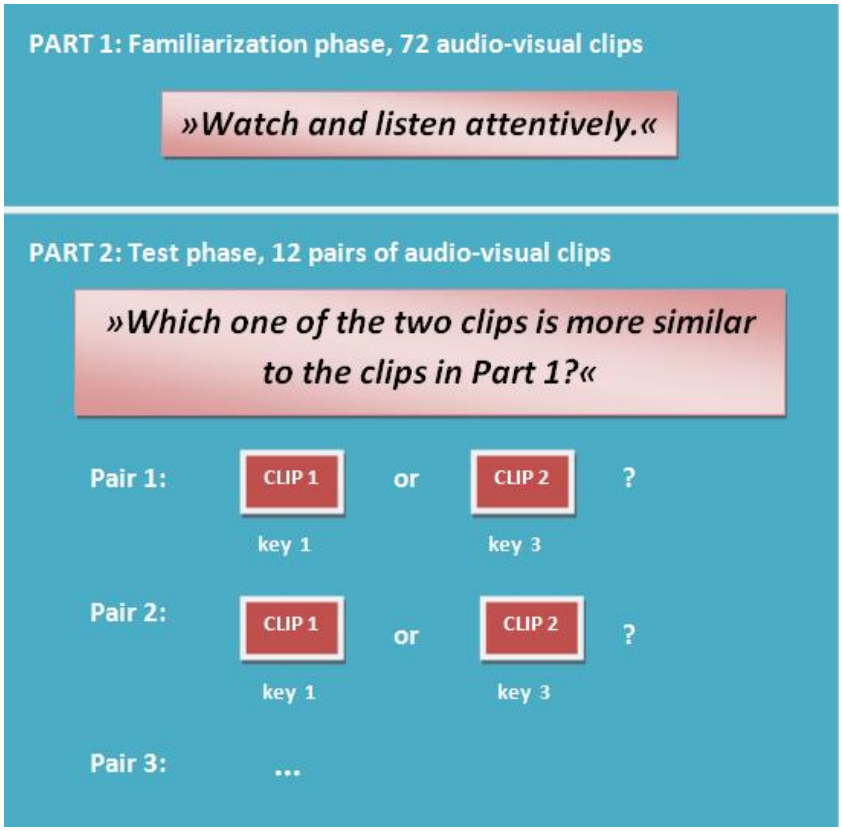

Fig.2. A graphic outline of the experimental procedure. There was no repetition in the 24 clips presented in the test phase: the repeated labels "Clip 1" and "Clip 2" always referred only to the clips' order of appearance in each new pair.

CONDITION 2: NON-LANGUAGE

In this condition, the only change was made in the context, which was non-linguistic. The participants in Condition 2 were, like their peers in Condition 1, exposed to a series of sine-wave stimuli, and finally tested to gauge whether they had implicitly learned the rule given over the consonants in the stimuli.

Participants

Condition 2 included the same number, gender ratio and linguistic background of participants as in Condition 1.

Materials

The words and the subsequent sine-wave audio stimuli in Condition 2 were the same as those in Condition 1, both in the familiarization phase and in the test phase.

The recording, cutting, editing and converting into sine-wave speech was not necessary as we used the audio stimuli designed for Condition 1. The only difference was in the visual clips on which we superimposed the sine-wave audio stimuli. In order to create a non-linguistic context and avoid triggering linguistic representation, we presented the sine-wave speech as generated by a large control panel, not unlike one seen on spaceships from sci-fi films. Studies investigating sine-wave speech often reported the subjects' feedback on their perception of the stimuli, which often included associations with spaceships, mission control panels, alien ships, and similar imagery. Searching on YouTube, we found an appropriate clip, which featured a hand pressing a series of buttons and flipping switches on a large illuminating panel, alternating between two sides and three positions (side one, side two; top row, bottom row, middle).

The clip was then edited in Adobe Premiere Pro to fit the sine-wave stimuli. Because the latter were based on 60 trisyllabic words, the space panel clip had to be cut into segments showing *three* button pushes or switch flips. Given this restriction, we obtained 18 different clips: 9 were cut from the original and then reversed to produce 9 new clips (e.g. where in the original clip the hand moved left to right, we created a new clip by reversing the movement, and thus obtained a new clip where the hand moved right to left). Finally, the sine-wave audio stimuli were superimposed on them, which required some of the visual clips to be slightly sped up. Because we had 16 unique visual clips and 60 unique audio clips, we had to superimpose several different sine-wave stimuli on a single repeated clip. To prevent the participants from being distracted by repetition, we arranged them so that each clip was followed by another that was highly contrastive (e.g. right-to-left switches in the top row of the first panel side were followed by left-to-right switches in the bottom row of the second panel side). The order of the audio-visual clips was the same as in Condition 1, which was designed to avoid sequences of vocalic or consonantal repetitions.

Fig.3. A still from an example audio-visual clip used in Condition 2. The footage of the moving hand flipping the illuminating switches created a non-linguistic context, in which the sine-wave stimuli were merely sounds produced by the control panel.

Procedure

The experimental procedure in Condition 2 was the same as in Condition 1, with the exception of the spaceship panel clips which replaced the moving mouth, and were used to create a non-linguistic context.

A minor change was also made in the duration of the pauses between clips both in the familiarization phase and in the test phase of Condition 2. The reason behind it was the limited flexibility of the visual clips of the panel: because the hand mostly flipped a row of four switches, we were able to cut two sequences of three out of the four flips, but we had to invariably cut the clip as soon as the hand moved toward the fourth switch (or as soon as it moved from the toward the second after having flipped the first one). Since the sine-wave audio stimuli could not be sped up or slowed down, we had to adapt the visual clips to them, which resulted in audio-visual clips that were shorter than those with the moving mouth. This was due to the inclusion of several hundred ms of silence in the audio-visual clips in Condition 1: the latter showed the initially closed mouth, the moving mouth, and the relaxing into s neutral, closed position at the end. This was not the case in the audio-visual clips with the panel, where there was no initial or closing silence due to the sharp cut before the hand moved to the fourth switch. Specifically, while the audio-visual stimuli featuring the moving mouth lasted around 2 s (Condition 1), the audio-visual stimuli featuring the panel lasted 1 s or less (Condition 2). This resulted in shorter intervals of silence between panel clips and consequently a faster tempo of presentation, which we compensated with increasing the duration of the pauses in to 1300 ms. This balanced the intervals of silence between the two conditions, albeit at the expense of a slightly longer white screen during the pauses in Condition 2 (although this was, based on the participants' feedback, not distracting).

Results

The data consisted of the participants' answers to the question in the test phase (binary choices of 1 or 3), and were analyzed in Matlab. We measured the success in rule-learning across participants in both conditions, comparing each against chance performance, and against the other. The results of the analysis are presented separately for each condition.

In Condition 1, where the sine-wave stimuli were processed in a linguistic context, the participants did not perform above chance level (i.e. 50% correctness in answering, which is in 6 out of the 12 test trials). The exact statistical values were the following: $t(15) = 0.92$, $p = 0.38$, $M = 6.37$, $SD = 1.62$. This

means that in the linguistic context, the participants failed in learning the structural rule that was given over the consonants of the sine-wave speech stimuli.

In Condition 2, the sine-wave stimuli were processed in a non-linguistic context. As in the linguistic context in Condition 1, the participants did not perform above chance level. The exact statistical values were the following: $t(15) = 1.53$, $p = 0.14$, $M = 6.75$, $SD = 1.94$. This means that in the non-linguistic context, too, the participants failed in learning the structural rule that was given over the consonants of the sine-wave speech stimuli. The mean number of correct answers per participant was marginally higher than in Condition 1, although still not significant, especially given the variance.

Comparing performance between the two conditions, we found that the difference in the mean number of correct answers per participant was negligible (an average of 6.38 correct answers in Condition 1, and 6.75 in Condition 2). Neither condition showed performance above chance level, which means that context did not influence the success in rule-learning to the extent that it would set the first condition apart from the second. The exact data of the statistical analysis of the mutual comparison of the conditions are the following: $t(30) = -0.59$, $p = 0.56$, $SD = 1.79$.

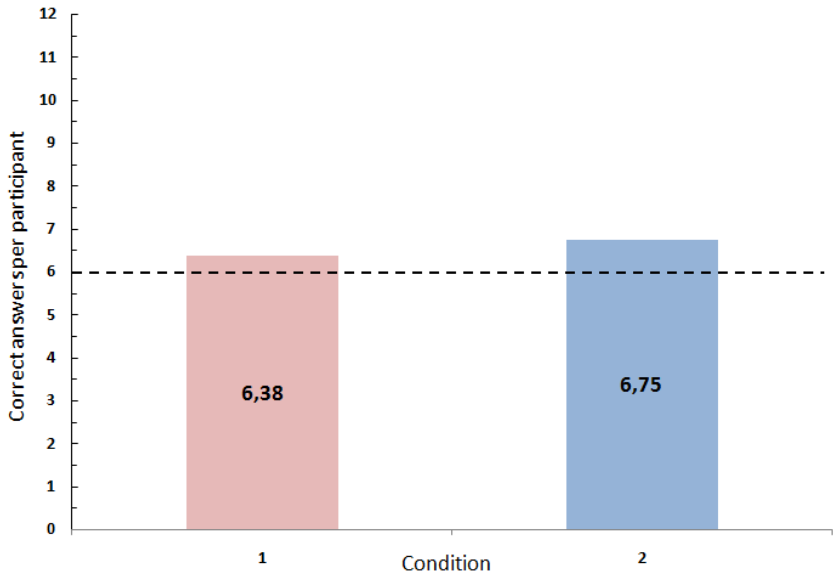

Fig.4. Mean number of correct answers per participant in Condition 1 (language) and Condition 2 (non-language). The difference in performance indicated by the number of correct answers (out of 12) is minimal between the two conditions. The dashed horizontal line marks the chance level at 50% (6 correct answers). The performance in both conditions is thus only just above random answering, and it does not show significant success in rule-learning regardless of the experimental context.

Discussion

Based on what we had learned about the different function of consonants and vowels in speech processing, we set out to determine whether the emergence of linguistic representation, which underlies this distinction, is context-dependent. Our aim was to test rule-learning over consonants, which has been shown to be unfavoured in linguistic contexts, but had not yet been examined in a non-linguistic context. Because the absence of linguistic representation should be the absence of linguistic categorical perception and functional restriction, we assumed that people would, in a language-free context, process non-linguistic auditory input much like animals do. We thus expected the participants in the non-linguistic context to be free of language-imposed restrictions, and thus be able to extract structural rules from sine-wave auditory stimuli both over vowels and over consonants. In contrast, we anticipated that the participants in the language context would fail due to the constraining effect of linguistic representation, which would be triggered by the sight of a mouth moving in synchrony with the sine-wave stimuli.

The results confirmed only part of our hypothesis. As predicted, the participants in the language condition (Condition 1) failed in reliably extracting the underlying structural rule over consonants. Their performance only just exceeded chance level, which was consistent with the results obtained in the studies discussed in the Introduction (among others, Bonatti, Peña, Nespor, & Mehler, 2005; Toro, Shukla, Nespor, & Endress, 2008). However, the same was the case in participants in the non-language condition: they, too, did not succeed in learning the rule given over consonants, even though the sine-wave auditory stimuli were presented in a non-linguistic context, as sounds generated by a control panel. We could interpret these data as evidence that people cannot process non-linguistic acoustic input as a rat or a tamarin would; before we do, however, we must re-examine our experiment to see whether its design could have somehow biased the final results.

The first aspect of the non-language condition that might have influenced the participants' performance were the audio stimuli. The sine-wave speech, which was intended to be reminiscent of a signalling spaceship control panel, was obtained through the conversion of the nonsense words as read by a female volunteer. The words were simple, short and articulated very clearly, and they contained the highly distinctive consonants S, K, and L in the familiarization phase (as well as P and M in the test phase). This generated sine-wave speech stimuli that might still have been perceived as language, albeit "mechanized," since they locally retained clearly distinct consonantal sounds, as well as the original cadence, which is a strong cue for linguistic representation. For a researcher, what is most intriguing about sine-wave speech is also the most problematic, as one cannot revert to hearing it as non-language once the phenomenon of its duality has been recognized. It was therefore difficult

to objectively gauge whether the stimuli could be plausibly presented as non-linguistic, or if they still carried recognizable language features. As evident from the participants' post-experiment informal feedback, only a few of them associated the control panel sounds with language.

Another concern regarding the non-linguistic sine-wave stimuli was their speed. As noted in Methods and Materials, the visual clips corresponding to the length of the acoustic stimuli were cut from a longer clip showing a hand flipping rows of four switches on two sides and on two levels of the panel. Because we needed only three flips to synchronize with the three syllables of the audio stimuli, we thus had to cut the original clip into segments that often started or ended quite abruptly. More importantly, because these clips did not include the short period of silence present in the mouth clips, the pauses between them had to be lengthened in order to match the rhythm of the presentation of the mouth clips. Despite these adjustments, many participants in Condition 2 remarked about the fast pace of the sequence of clips. Although the pauses had been balanced between both conditions, the movement of the hand from switch to switch had to match the transitions from one syllable to the next, and was thus still perceived as quite rapid. Compared to the motion of the hand, the movement of the mouth was much slower and more gradual.

If the speed of the moving hand in fact impeded rule-learning, it could have been because the participants were not able to process both the images and the sounds, and make a meaningful connection between them in that short a time. Similarly, viewing a very rapidly moving mouth would most likely not significantly facilitate the processing of the fast-paced speech it produced. Nevertheless, there is an important difference between the association of linguistic visual cues with language, and that of non-linguistic visual cues with non-language. Because of our cognitive predisposition for language and a lifetime of experience with it, we operate with a large memory base of associations between language and visual stimuli that cue or aid its processing. The sight of a moving mouth is a universal trigger of language representation, one we encounter regularly and recognize instantly. The association between a spaceship control panel and sine-wave audio stimuli, on the other hand, is arguably a novel connection. While the participants, seeing the illuminating panel, most likely did perceive the context as non-linguistic, the rarely experienced combination of the visual and the acoustic stimuli might have caused their slower processing. What is more, if the movement of the hand in the visual clips was indeed too rapid, the simultaneity of the audio-visual stimuli might have taxed short-term memory, and this may in turn have impeded normal processing of the auditory input. It is therefore conceivable that the participants in the non-linguistic context did not fail to learn a structural rule given over consonants because of an intrusive influence of linguistic representation: rather, it might simply have been the consequence of an attentional overload.

On the other hand, and the shortcomings of our experimental design notwithstanding, it is possible that humans in fact cannot process non-linguistic stimuli like other animals do, as an auditory stream characterized only by its acoustic properties. What distinguishes us is linguistic representation, which we perceive to be the hallmark of human language. The auditory stimuli that do not trigger it are considered non-language, which suggests the coexistence of two separate modes of processing auditory input: linguistic, which conforms to the deeply-rooted linguistic hierarchy, and non-linguistic, which circumvents these constraints altogether. The latter is not unlike the auditory processing observed in animals: but can *we* really do it? If the conclusions of our experiment reflect reality, and humans in fact apply abstract categorization even to non-language, can we still talk about non-linguistic contexts?

If the distinction does exist and we are capable of two kinds of acoustic processing, we might, until new research sheds further light on the issue, draw alternative explanations from our observation of daily life. Let us return to our introductory example: when the washing machine fills up with water and goes into the wash cycle, our attention is caught by the regular rhythmic sounds it makes as it spins. These sound are often likened to language, in that they resemble the repetition of a word, which feels reproducible through our phonetic repertoire, or even in writing. The sine-wave stimuli in our non-language condition were similar: short and simple, containing a clear consonantal – vocalic contrast, and following a regular stress pattern. The acoustic quality of the stimuli might thus have triggered linguistic representation even though the accompanying visual stimuli did not. Perhaps it is acoustic quality of the auditory input, rather than visual context, which dictates the manner in which the input will be processed. If this is the case, the presence of linguistic cues may override the influence of (at least certain) contextual factors, and whatever we *can* perceive as language, we *will* perceive as language.

References

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic Constraints on Statistical Computations. *Psychological Science*, *16*, 451-458.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2007). On Consonants, Vowels, Chickens, and Eggs. Psychological Science, 18, 924-925.

De la Mora, D. M., & Toro, J. M. (2013). Rule Learning Over Consonants and Vowels in a Non-Human Animal. *Cognition*, *126*, 307-312.

Hochmann, J. R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and Vowels: Different Roles in Early Language Acquisition. *Developmental Science*, *14*, 1445-1458.

Mehler, J., Peña, M., Nespor, M., & Bonatti, L. (2006). The "Soul" of Language Does Not Use Statistics: Reflections on Vowels and Consonants. *Cortex*, *42*, 846-854.

Nespor, M., Peña, M., & Mehler, J. (2003). On the Different Roles of Vowels and Consonants in Speech Processing and Language Acquisition. *Lingue e Linguaggio*, *2*, 203-229.

Pons, F., & Toro, J. M. (2010). Structural Generalizations Over Consonants and Vowels in 11-Month-Old Infants. *Cognition*, *116*, 361-367.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech Perception Without Traditional Speech Cues. *Science*, *212*, 947-950.

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding Words and Rules in a Speech Stream. *Psychological Science*, *19*, 137-144.

Toro, J. M., Shukla, M., Nespor, M., & Endress, A. D. (2008). The Quest for Generalizations Over Consonants: Asymmetries Between Consonants and Vowels Are Not the By-Product of Acoustic Differences. *Perception & Psychophysics*, *70*, 1515-1525.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2004). Audio-Visual Speech Perception Is Special. *Cognition*, *96*, B13-B22.

Vroomen, J., & Bart, M. (2008). Phonetic Recalibration Only Occurs in Speech Mode. *Cognition*, *110*, 254-259.

Yip, M. (2006). The Search for Phonology in Other Species. *Trends in Cognitive Sciences*, *10*, 442-446.

Appendix

We attach the full list of the words used in the two conditions of our experiment.

FAMILIARIZATION PHASE

KEKOLI SOSILE
LILOSE KOKESI
SESOKI LELISO
KOKELI SOSIKE
LELOSI KIKOLE
LOLEKI SISEKO
SESOLI
KEKILO
KOKISE
LILEKO
SOSELI
LOLIKE
KEKOSI
SESILO
LILOKE
KIKESO
LELIKO
SISELO
KEKISO
SOSEKI
KOKILE
LOLESI
SISOLE
LELOKI
KIKELO
SISOKE
LOLISE
SESIKO
LILESO
KIKOSE

TEST PHASE: "CORRECT ANSWERS"

PEPIMO
PEPOMI
PIPEMO
PIPOME
POPEMI
POPIME
MEMOPI
MEMIPO
MOMEPI
MOMIPE
MIMEPO
MIMOPE

TEST PHASE: "INCORRECT ANSWERS"

PEMOPI
PEMIPO
POMEPI
POMIPE
PIMEPO
PIMOPE
MEPOMI
MEPIMO
MOPEMI
MOPIME
MIPEMO
MIPOME