Referential iconicity in music and speech

Verónica Giraldo

Musical meaning is multifaceted. It is highly sensory and yet often abstract; able to cross cultural boundaries and yet embedded in specific traditions. For the most part music as a semiotic system is characterized by non-referential meaning (Monelle, 1991). Nevertheless, in so-called programmatic music, musical themes are intended to refer to worldly objects and events on the basis of iconic (and indexical) grounds. Such non-arbitrariness has been extensively documented in the case of speech as well (Ahlner and Zlatev, 2010; Sonesson 2013; Imai and Kita, 2014). In an experimental study, I investigated how referential iconicity in speech operates in comparison to music, considering the factors (a) primary/secondary iconicity and (b) linguistic/cultural background. In the experiment 21 Swedish and 21 Chinese native speakers had to match musical fragments from Prokofief's *Peter and the Wolf* and spoken word-forms to objects, represented by schematic pictures. The experiment was designed to have two conditions to operationalize higher degree of primary and secondary iconicity, respectively. The results showed that there was no significant difference between the overall results for music and linguistic tasks, indicating that the cognitive-semiotic processes involved are not limited to a single cognitive domain or semiotic system. Both groups performed significantly above chance in both conditions, which serves as a clear indicator that interpreting referential music in music and speech sounds is *not* purely a case of secondary iconicity.

1. Introduction

Musicologists, linguists and cognitive scientists have analyzed music from various perspectives (e.g. Piaget, 1971; Coker, 1972; Lerdahl and Jackendoff, 1983; Nattiez, 1987, Monelle, 1991). However, there is little research concerning the ability of musical themes to refer to objects and events in the world on the basis of resemblance, or what I refer to as *referential iconicity*. This is hardly surprising, given that iconicity in speech has only recently become a popular topic (e.g. Ahlner and Zlatev, 2010; Dingemanese, 2012; Vigliocco, Perniss and Vinson, 2014; Ibarretxe-Antuñano, 2017). In this paper, I compare referential iconicity in music and speech, using concepts and methods from cognitive semiotics, the transdisciplinary field "focused in the multifaceted phenomenon of meaning" (Zlatev, 2015, p.1043).

Sonesson (1997) distinguished between *primary* and *secondary iconicity*. The first is so transparent that an interpreter can understand the iconic sign on this basis, while the second is discernible only after one is told what the expression means. Further, he proposed that iconicity in language and in music "most of the time seems to be secondary" (Sonesson 2009, p.51). However, though the resemblance between speech sounds and meaning maybe subtle, it is perceivable by both adults and children (Imai and Kita, 2014), even when it occurs across sensory modalities (Ahlner and Zlatev, 2010), *before* participants are told what the expressions in question mean. The same appears to be the case also in Western *programmatic music*, which "…carries some extra-musical meaning, some 'program' of literary idea, legend, scenic description, or personal drama. It is contrasted with so-called absolute, or abstract, music, in which artistic

interest is supposedly confined to abstract constructions in sound" (Encyclopedia Britannica, 1998). In other words, programmatic music aims to create a correlation between the music themes and something in the world, while non-programmatic music does not.

A popular method that has been used for studying iconicity in speech has shown that participants can "guess" the meaning of nonce words (e.g. *bouba* or *kiki*), by matching them to visual stimuli (shapes with soft or sharp edges). But, can such mappings also be found between musical themes and objects, especially for participants who are not familiar with the musical tradition in question? In Section 2, I present the relevant theoretical background and spell out the research questions in detail. In Section 3, the methodology, alongside the stimuli selection and choice of participants in the experiment are presented, finalizing with specific hypotheses. Section 4 reports the results which provide a general discussion, before concluding in Section 5.

2. Theoretical background

2.1. Signs and semiotic grounds

Within the transdisciplinary field of cognitive semiotics, combining concepts and methods from linguistics, semiotics and cognitive science (Zlatev, 2015; Zlatev, Sonesson and Konderak, 2016), researchers use different definition of the sign. Many refer to the classical definition given by Peirce below, with emphasis on the "somebody", i.e. the *interpreter* as a conscious agent who, as either producer or comprehender, establishes the relation between representamen and object, as illustrated in Figure 1.

> [a] sign, or *representamen*, is something which stands to somebody for something in some respect or capacity … The sign stands for something, its *object*. It stands for that object, not in all respects, but in reference to a sort of idea which I have sometimes called the *ground* of the representamen (Peirce, 2003, p.106).
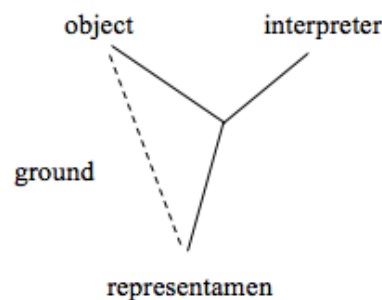

Figure 1. Graphic illustration of the sign and its interacting components
(adapted from Ahlner and Zlatev, 2010, p.314)

Different kinds of signs can be classified on the basis of the *predominant* kind of ground between representamen and object. For *iconic signs (icons),* this is similarity (e.g. a drawing of a cat). In the case of *indexical signs (indices),* the ground is (a) space-time contiguity or (b) part-whole relationships (e.g. a dog leash for a dog). For *(symbolic signs) symbols*, the semiotic ground identified is (arguably) conventionality*,* understood as common knowledge (e.g. RED is known to mean 'stop'). It is important to highlight, again, that signs typically combine all three grounds, or as stated by Jakobson (1965, p.26):

> It is not the presence or absence of similarity or contiguity between signans and signatum, nor the habitual [conventional] connection between both constituents underlies the division of

signs into icons, indices and symbols, but barely the predominance of one of these factors over the others.[1]

Signs do not occur in isolation but relate to other signs in more or less complex relations to form *mms*, allowing the expression of composite contents such as stories (Louhema, 2018; Stampoulidis, 2019). Language is the best-known sign system, but gestures and depiction are two other universal systems for human meaning making (Zlatev, 2019), with iconicity playing an important semiotic ground in all cases. This clearly contradicts the Saussurean understanding of the linguistic sign, and the claim of its fundamental *arbitrariness* (Saussure, 1959[1916]), understood to mean that there is no motivated linkage between the two entities that compose the linguistic sign (see Zlatev, 2014). From a cognitive semiotic perspective, on the other hand, very few linguistic signs are entirely unmotivated or, in other words, have a ground that is purely conventional.

2.2. Sound symbolism and types of iconicity

The thesis that language is a motivated semiotic system, with iconicity and indexicality playing central roles, has been theorized in alternative terms. One of these is *sound symbolism* that can be defined as "a motivated, non-arbitrary relation between the sound patterns and the meaning of words" (Johansson and Zlatev, 2013, p.3). An obvious example of this is *onomatopoeia,* but the phenomenon goes far beyond expressions (in English) like *bam, splash, woof* and *meow.* Such expressions are only the most commonplace example of *ideophones,* defined by Dingemanse (2012, pp.654-655) as "marked words depictive of sensory imagery found in many of the world's languages". Studies such as those of Dingemanse (2012) and Ibarretxe (2017) have contributed to the establishment of the study of ideophones in linguistics. Dingemanse (2012) provides a review of research on ideophones, as well as the tentative *implicational hierarchy* (ibid., p.663) for ideophones shown in (1).

(1) SOUND < MOVEMENT < VISUAL PATTERNS < OTHER SENSORY PERCEPTIONS < INNER FEELINGS AND COGNITIVE STATES

This means, for example, that if ideophones for movement are found in a language, that language will also have ideophones for sound. This means that the most common type of ideophones should be sound-to-sound ideophones, or onomatopoeias. This implies unimodality, or mappings that stay within the same sensory modality (in this case, hearing). The rest of the subcases naturally imply that they are *cross-modal*, or mappings that cross sensory modalities, e.g. from sound to movement (kinesthesia) or sound to (visual or tactile) shape (Ahlner and Zlatev, 2010). This aspect is relevant for music, where both unimodal and cross-modal mappings occur.

A possible explanation for some aspects of sound symbolism is the so-called *frequency code,* where high frequencies are said to be associated with small things, whereas low frequencies are associated with big things (Ohala, 1994, 1997). Johansson and Zlatev (2013) investigated different possible motivations for spatial demonstratives in the world's languages, arguing that the vowels and consonants in demonstrative pronouns could have a connection to proximity or distance, with a sample of 101 languages. In English, for example, the front closed vowel [i] in *this,* would map to closeness, whereas the open back vowel [a] in *that* should map to semantic distance. The study found that in 56% of the cases the prediction based on the frequency code was supported, in 22% of the cases showed a reversed pattern and in 22% there was no difference.

But are such correlations *functional* for language users, rather than just "evolutionary relics"? (see Imai and Kita, 2014, for discussion). A popular paradigm of matching non-words to objects (represented visually), suggests a positive answer to this question (Köhler, 1929;

---

[1] In this citation, *signans* corresponds to representamen, and the *signatum* to the object.

Ramachandran and Hubbard, 2001; Maurer, Pathman and Mondloch, 2006; Imai et al., 2008; Ahlner and Zlatev, 2010; Cuskley, Simner and Kirby, 2017). Participants are typically provided with two shapes (one sharp, one rounded), and told that one was called e.g. *bouba,* and the other e.g. *kiki.* Following this, they are asked to state which of the shapes was *bouba,* and which one was *kiki.* Ramachandran and Hubbard (2001) found that in 95% of the cases people matched *bouba* to the rounded figure and the sharp one to *kiki.* Ahlner and Zlatev (2010) followed up on this study, by making different combinations and contrasts of vowels and consonants, where made-up words with the vowel /i/ and consonants /p/, /t/, /k/, were expected to have "sharper" features, while made-up words with the vowel /u/ and consonants /m/, /l/, /n/, would have "softer" features. They found that both vowels and consonants contributed to the appropriate mapping to sharp and round figures, and proposed that haptic sense, in the vocal tract, mediated between the visual and auditory modalities. More specifically, the authors explicitly linked such findings with iconicity. But what *kind* of iconicity?

Peirce (1974[1931]) distinguished between three different types of iconicity: *imagistic*, *diagrammatic* and *metaphoric* (Jakobson, 1965; Devylder, 2018). Imagistic iconicity implies a nearly full match between the representamen and the object, as in a photograph, or a realistic sound imitation. Devylder (2018) presents a very straightforward example of a diagrammatic icon, shown in Figure 2: it is the *relations* between the signs that resemble the relations between the eyes and mouth of a human face (ibid, p.322), not the images of these signs.

Figure 2. Diagrammatic icon of a human face

Sonesson (1997) provided another important subdivision of iconic signs in terms of *primary* and *secondary iconicity*. The first states that the "perception of an iconic ground obtaining between two things is one of the reasons for positing the existence of a sign function joining two things together as expression and content" (Sonesson, 1997, p.741). In other words, it is the perception of the similarity between *object* and *representamen* that allows for the understanding of the sign. For example, an observer can understand the image in Figure 3 as a picture of a lama, given that (s)he is familiar with what a lama in the real world looks like. Secondary iconicity, on the other hand, implies that the "knowledge about the existence of a sign function between two things functioning as expression and content is one of the reasons for the perception of an iconic ground between these same things" (ibid.). Here, the interpreter needs to know what a particular iconic sign represents in order to perceive the similarity in question. For example, only given that we are told that Figure 4 represents a boa constrictor that has swallowed an elephant (Saint-Exupéry, 1943), it is possible (for most of us) to see the similarity in question.
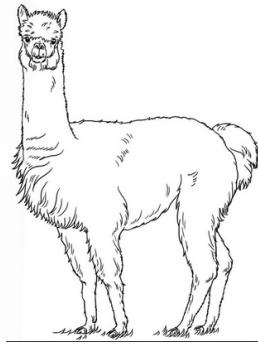
Figure 3. Example of primary iconicity: the similarity of the image to lama leads to understanding the sign.

Figure 4. Example of secondary iconicity: being told that this is, for example, "a boa having swallowed an elephant", we can see the similarity to such an event.

Ahlner and Zlatev (2010) proposed that to solve the typical "bouba-kiki" task (see above), participants need first secondary iconicity, i.e. to be told that what they are given are two signs. But since they are not told which expression is which object, they also need primary iconicity, to be able to use the most "natural" matching between sound forms and shapes. Analogously to the way that the different grounds may combine in a specific sign, the interpretive process may involve a combination of both secondary and primary iconicity, in a sequence of three steps, shown in Figure 5. In Step 1, the instructions inform the participant (the interpreter) that there is a relationship between the each representamen (R1 and R2) and the objects (O1 and O2). In Step 2, the participant "discerns the composite analogous ground" (ibid, p.319) between the objects and the representamina. This is a secondary iconicity element, since the participant knows already there is a sign relation and, on this basis, looks for the ground. Finally, in Step 3 the participant posits a specific sign relation (e.g. R1-O1 and R2-O2), thus creating two specific signs, and this is the primary iconicity element. From this we can gather that there is a combination of primary and secondary iconicity in solving the task, with a considerable role for secondary iconicity.
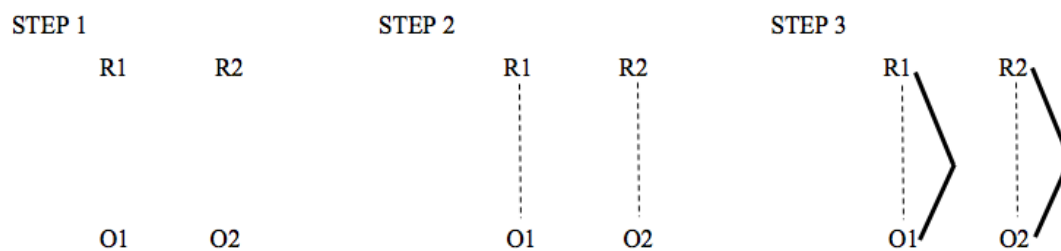
Figure 5. Matching representamina to objects by finding a composite analogous ground in more-contrastive tasks (Ahlner and Zlatev 2010, p.319).

2.3. Music as a semiotic system

As stated in the introduction, comparisons between linguistic and musical aspects have been extensively explored. Under the understanding that music is systematic in its structure, many researchers have delved into the study of the "syntactic" structure of music, such as Bernstein (1976) and Keiler (1978, 1981). Lerdahl and Jackendoff (1983) set out to explore the "grammar of music" within a generative linguistics theoretical framework (e.g. Chomsky, 1965). Treating the structure of music, or language, in purely syntactic terms, however, misses the obvious fact that both are crucially systems of *meaning*, even if the meaning conveyed in music may differ from meaning in language. As pointed out by Monelle (1991, p.30), it is natural to turn to semiotics when analyzing music as a particular kind of semiotic system.

Taking into account that one of the dominant functions of language is to state propositions about the world, it can be understood that the main difference between language and music as semiotic systems is that music has predominantly *non-referential* meaning, while linguistic meaning is to a large extent *referential* (*denotational*). Many argue that that this type of meaning cannot be obtained from music since no specific semantic reference can be ascribed to musical terms (Kivy, 2002, 2007), but there are instances in music that question this claim, namely in the cases of programmatic music, where referential iconicity is present. More generally, it can be argued that:

> … lacking specificity of semantic reference is not the same as being utterly devoid of referential power. Instrumental music lacks specific semantic *content,* but it can at times suggest semantic *concepts.* Furthermore, it can do this with some consistency in terms of the concepts activated in the minds of listeners within a culture (Patel, 2008, p.328).

A study by Hacohen and Wagner (1997) grant support for this. The authors first provided a group of listeners with a selection of nine Wagnerian leitmotifs, alongside seven semantic scales.[2] The participants were asked to rate each leitmotif on the seven scales, finding that the leitmotifs were systematically arranged in three clusters: a friendly, a violent and a dreary cluster. On this basis, a second group of participants, based on the same selection of nine leitmotifs, were first asked to imagine the piece as a musical theme in a movie, and then asked to provide a name

---

[2] Leitmotifs can be understood as "compact musical units designed to suggest extra musical meaning such as a particular character, situation or idea […] Leitmotifs provide an opportunity to study the semantic properties of music because they are designed to have referential qualities" (Patel, 2008, p.328). This notion is mostly discussed with reference to Wagner's compositions, but they are not limited to his music. The seven semantic scales used were: 1. Joy/Sadness 2. Hope/Despair 3. Natural/Supernatural 4. Strength/Weakness 5. Impetuosity/Restraint 6. Dignity/Humility 7. Kindness/Cruelty.

for the movie that would go along with the music. The results showed that there was an overlapping consistency in the titles given to each leitmotif.

It should be noted that music can be referential to other genres and pieces of music or meaning within the musical realm. The characteristics that delimit each music genre are very often delimited by time epochs. In this sense, a salsa piano player in the 1970s could actually play a romantic piece (dating from the XIX century) adapting the music to the rhythms and styles characteristic of salsa.[3] This corresponds to the notion of *dialogicity* in language (Bakhtin, 1981): the continuous connections between new and previous utterances. For example, a character's voice in a literary novel is involved in a dialogue not only with other voices within the same text, but also with utterances in different works. But can music also represent physical worldly objects? Some examples of the type of music that does so are (a) *Le Carnaval des Animaux* (1886) by Camille Saint-Saäens*,* a humorous suite representing different animals (b) works by Rimsky-Korsakov such as *Flight of the Bumblebee* (1889-1900), an orchestral interlude evoking the rapid movements of a bumblebee and *Scheherazade* (1888), a symphonic suite based on the *One Thousand and One Nights* (1839), where some characters, such as Scheherazade are represented by their own melody and instrument (c) Vivaldi's well-known *Four Seasons* (1721), a four-part violin concerto, each representing a season of the year and (d) *Peter and the Wolf* (1936) by Prokofiev.

A common question posed regarding such programmatic music is whether it can convey referential meaning *across different cultures*. The music pieces (a-d) belong to the *Western classical music* tradition. This tradition, encompassing a broad number of musical periods and styles that vary according to their historical period, ranges from the year 500 AD to the present. A dominant genre originally in Europe, it has expanded world-wide due to colonization and cultural influence. For example, regardless of classical music being remarkably different from a Colombian *bambuco,* its influence is reflected through musical notation and instrumentation.[4] This type of influence can be further perceived through a large range of folk music genres across the world, such as jazz, salsa, and a number of traditional folk music across Europe and the Americas.

At the same time, even though many music cultures around the world have been influenced by European classical music, this does not mean that they are considered a part of classical music. Nor does it mean that members of the cultures that create and perform this type of music feel that classical music reflects their musical culture. These traditions have drastically different rhythms and intonations, which have been shaped by the local people, their culture, and the historical epoch. These are characteristics of the complex process of musical evolution. As pointed out by Monelle (1991, p.276):

> Music carries a social context: it starts and develops amongst a whole community. The aspects of intonation are tied to a historical, social and cultural period. These define the means of musical expression and the selection and interconnection of musical elements.

Especially relevant in the present context is Chinese classical music, which differs from Western in many respects (Shen, 1991; Tien, 2015). In Western music, tonality is a fundamental arrangement for the composition of most classical (and non-classical) music.[5] Tonality is basically the systematic arrangement of musical pitches or chords, grounded in a hierarchical structure, where different pitches possess larger stabilities than others. This can be reflected in

---

[3] For example, *Sonido Bestial* (1971) by Richie Ray and Bobby Cruz.

[4] The *bambuco* is a traditional music originating from the Andes region in Colombia that has been inspired by the rhythms of waltzes and polskas, and its main instruments are the piano, vocals, guitar, and a series of variations of guitars namely *tiple*, *bandola* and *requinto*.

[5] Note, however, that not all Western music is necessary tonal, such as the subgenre of *atonal music* that emerged at the beginning of the XX century.

the use of scales, which is a construction of seven tones arranged according to their resonance, where the most stable pitch is called the tonic, and gives rise to different keys. There is a variety of scales, fluctuating on the resonance distance between the tones, but this pattern provides a fundamental structure for Western classical music. This, however, does not apply to Chinese classical music, which according to Shen (2000, p.22) focuses on "harmonies" rather than "melodies". Further studies have explored the aesthetic differences between Chinese and Western classical music attributing some of their ground differences to cultural backgrounds, specifically from Chinese's traditions that have been maintained for hundreds of years (Chen, 2015). Following the understanding that there are intrinsic cultural differences between Chinese and Western classical music, it is possible to connect this to how members of both cultural groups perceive music, where studies such as Drake and El Heni (2003) validated that acculturation, or the "learning by immersion in the auditory world around us" (Drake and El Heni, 2001, p.22), influences how members of different cultures perceive music. In a similar study, Lynch and Eilers (1992) found that music perception in both adults and children was culture specific.

Even though the examples mentioned here are only some out of the many characteristics that differentiate Western classical music and Chinese classical music, this indicates that the Western harmonic and theoretical system cannot be fully applied to Chinese classical music, and that this could even be reflected in the way we perceive and listen to the music.

2.4. Summary and research questions

In this section I described features of language and music as non-arbitrary semiotic systems, and discussed different kinds of iconicity, with focus on the distinction between primary and secondary iconicity. I discussed the presence of referential iconicity in at least some kinds of classical music in the Western classical tradition, which while highly influential is not universal. I discussed in particular how it differs from the Chinese tradition. On the basis of this background, I may present the following research questions:

- RQ1. Will iconic themes in programmatic music and speech display predominantly primary or secondary iconicity?

- RQ2. Will there be differences in how participants belonging to different musical traditions perceive iconic themes from a particular culture?

3. Methods

To address the subject of referential iconicity in speech and music empirically, an experiment was conducted where Swedish and Chinese participants were asked to match different representamina (musical and linguistic) to a number of objects shown on a computer screen.

3.1. Materials

The following materials were used in the study, with representamina always in the auditory modality, and objects in the visual modality: images or written words denoting the respective objects in the world.

3.1.1. Musical representamina

The music stimuli used for the experiment were six excerpts of the musical piece *Peter and the Wolf*, composed in 1936 by Sergei Prokofiev, is an "orchestral tale for children" (Rifkin, 2018). A narrator tells the story of the boy Peter, who lives with his grandfather in the woods and sneaks

out of his yard to play in the meadow, against his grandfather's will. Peter is accompanied by a duck, a bird and a cat, before they encounter a dangerous wolf. As it is a symphonic tale, the verbal narration is complemented by the orchestra. However, in the introduction of the piece, the narrator discloses that each character in the story is represented by a different instrument in the orchestra, followed by a short introductory melody of each character performed by their respective instrument. This piece was chosen, as it is an opportune example of programmatic music, with clear representations of objects existing in the world. The arrangement of the piece, which starts with introductory melodies to each character, was another pertinent factor for employing this as the main musical material, given that it allowed for a clear representation of each character, and the duration of each melody was appropriate for the duration of the experiment. Also, the fact that each character is represented by a specific instrument helped address the notion of frequency code.

The musical stimuli consisted of the six introductory melodies corresponding to six of the characters presented in the piece: Hunters, Bird, Cat, Duck, Wolf and Grandfather. Further details of each melody are presented in Table 1. The average duration of the six melodies was 14.6 seconds. Peter's theme was not used as musical representamen, as its melody differs from the others, given that it is performed by the whole strings section (violins, violas, cellos and double basses) of the orchestra, as opposed to only one instrument. It was, however, used in the process of participant recruitment, see Section 3.3.

Table 1. Details of the musical stimuli

Character	Instrument	Pitch Range	Duration	Iconicity
Bird	Flute	E4-G6	13 seconds	Unimodal
Duck	Oboe	C4-D5	16 seconds	Cross-modal
Cat	Clarinet	G3-F4	10 seconds	Cross-modal
Hunter	Timpani	E2-C3	7 seconds	Unimodal
Wolf	French Horns	G2-F4	19 seconds	Cross-modal
Grandfather	Bassoon	B1-G3	23 seconds	Cross-modal

The musical stimuli were in accordance with the frequency code motivation for sound symbolism (see section 2.2.), according to which pitch corresponds to the size of an object. Since each character was represented by a different musical instrument (as shown in Table 1) the pitch range of the latter can be shown using so-called scientific pitch notation (SPN), which combines musical pitch with musical notes (Young, 1939). A visualization is presented in Figure 6.

Figure 6. Scientific Pitch Notation in octaves of the note C.

Here we can see the ten possibilities for the note C, where there is a progression from low to high frequencies, from left to right. The distance between each C is of an octave, meaning that it is the same note, but with double its frequency. Between every C in the figure, there are six notes, namely D, E, F, G, A and B. In Table 1, when referring to, for example, D3, it means that this is the first D placed between C3 and C4. The symbols 𝄢 and 𝄞 are used to indicate the pitch of written notes. The notes marked after the former symbol represent notes with lower pitch, whereas the notes marked after the latter represent notes with higher pitch. The pitch range provided in the third column in Table 1 denotes the pitch range of the instrument in each melody (which were gathered from the music scores of each individual melody), and not the instrument in general. Cells with higher color pigmentation represent instruments with lower frequency, while those with lighter pigmentation represent those instruments that have higher frequencies. For the warmup task, a 22 second excerpt of the piece *Dance of Dragons* composed by Ramin Djawadi (2015) for the popular HBO series Game of Thrones, was used.

3.1.2. Ideophone representamina

For the first linguistic task, a set of six *ideophones,* in a language that was unfamiliar to all participants (Basque) was used. These were chosen from Ibarretxe's (2017) wide compilation of conventionalized Basque ideophones, on the basis of four criteria: (a) the ideophone had to be reduplicated; (b) it denotes an action; (c) two Basque native speakers confirmed that the ideophones were unimodal (i.e. resembled the sounds made by the denoted actions), and recognizable (some ideophones in the compilation were outdated or very specific to geographical locations); and (d) each ideophone did not correspond to any word in the languages of the participants, Chinese and Swedish. All of the ideophones were recorded by a female native speaker of Basque (age 22) and are shown in (2-7).

(2) *Grik-grak* 'to crackle'
(3) *Pil-pil* 'sound of boiling water'
(4) *Draka-draka* 'horse galloping'
(5) *Zirris-zarraz* 'sound of sawing'
(6) *Zorro-zorro* 'snoring'
(7) *Trinkili-trankala* 'move noisily, with difficulty'

3.1.3. Fictive word representamina

For the second linguistic task, six two-syllabic CVCV non-words were created. For this, the following criteria provided by Ahlner and Zlatev (2010, p.324) were employed: (a) consonants were either voiceless obstruents [tʃ, t, k], or voiced sonorants [m, l, n]; (b) vowels were either front close unrounded [i, e] or back open [a, u]. This provided a basis for creating non-words that sounded either "soft" and "round", or "sharp" and "pointy", based on intuition and "the 'synesthetic' properties that have been ascribed to sounds in previous studies" (ibid., pp.323-324). All of the non-words were audio recorded by the same speaker that recorded the ideophones, and all recordings had a duration of 1 second.

As with the ideophones, it was made sure that these forms did not match actual words in Swedish or Chinese. For example, having [mu] as a first syllable was avoided as it resembled the word *wood* in Chinese. In order to make sure of this, two L1 Chinese speakers and two L1 Swedish speakers (who did not participate in the main study) were asked to listen to the recordings of all words to make sure none of these had any meaning in either language. A list of all fictive words is provided in Table 2.

Table 2. Spoken stimuli for fictive words

"sharp" and "pointy"	"soft" and "round"
[keti]	[nalu]
[kitʃɪ]	[lulu]
[tike]	[lamu]

Two extra non-words [kling] and [klang] were presented in the warm-up task, where they were to be matched with the two policemen in the well-known Swedish children's books series *Pippi Longstocking* (Kling is tall and slim and Klang is short and bulkier). These were likely to be familiar for the Swedish participants and less so for the Chinese, but since the purpose of the warm-up was to give each participant an understanding of the task, this difference was not judged to be problematic.

3.1.4. Image objects

The image-objects to be matched with music representamina were generic outlines, all without any coloring, of each respective animal or person. Given that each participant was not to see the representamen of each character more than once (see below), a set of six foils was employed. This group was designed to have a proportionate number of humans and animals. As a result, images of a squirrel, a cow, a pig, a bear, a ballerina and a Mexican luchador, (all following the same characteristics described above) were used. For the warmup, the target image was that of a dragon, and the foils consisted of a platypus, a donkey and a racoon. All of the objects were found through an image search in Google. Example of two of the images are shown in Figure 7.
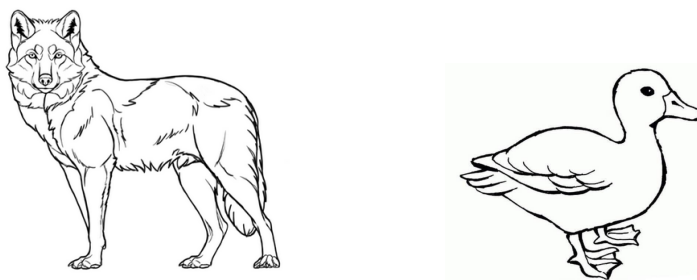
Figure 7. Examples of two of the images chosen to represent their respective characters

The image-objects to be matched with the fictive words described in the previous sub-section consisted of six figures which were hand-drawn and later scanned. Three of the shapes had sharp, angular inflections, whilst the other three had wavy, round contours, as shown in Figure 8. Additionally, two foils were used, consisting of two elemental geometrical figures, a square and a circle. The full compilation of the images used for both music and fictive words tasks, including the warm-ups, is shown in Appendix I.
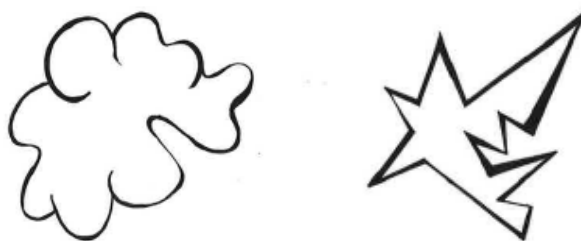
Figure 8. Examples of two of the shapes created for the sound symbolism tasks

3.1.5. Written word objects

Since the image-objects could have been culturally and individually biased, in their role of representations of the ultimate objects, half of the trials for the music tasks consisted of written words where, instead of the participant being presented with the generic images described above, they were presented with a word in their respective language, written in a large font.[6] These were used as controls in the two music tasks (see below). The word-objects for the ideophone tasks consisted of the translations of the examples (2-7) to both Chinese and Swedish, performed by native speakers. The full compilation of all written words for both music and ideophone tasks in both Swedish and Chinese is shown in Appendix II.

3.2. Experimental design

Each participant was to perform the following four tasks, using the materials described in the previous section.

> T1. Match music representamen with images as objects
>
> T2. Match music representamen with written words as objects
>
> T3. Match fictive word representamen with images as objects
>
> T4. Match ideophone representamen with written words as objects

Each of these tasks had two conditions: a *less-contrastive* and a *more-contrastive*. In the more-contrastive condition, the participant was presented with two representamina, and two objects, which were determined by the content of each task. For example, if the task was a more-contrastive music task with images as objects, then the representamina consisted of two melodies (e.g. 'wolf' and 'duck') and the objects consisted of the images of the duck and the wolf. If the task was a more-contrastive ideophones with words as objects, the representamina consisted of two ideophones (e.g. *Grik-grak* and *Zorro-zorro*) and the objects consisted of their respective translations to either Swedish or Chinese. Figure 9 illustrates these two examples as presented in the experiment.

---

[6] This raises the question about possible effects of orthography on solving the task, as discussed by Cuskley, Simner and Kirby (2017). As this was, however, not systematically controlled in the present study, it remains a question for future research.

Figure 9. Examples of the more-contrastive condition (T1) and the more-contrastive condition for the ideophone task (T4) presented in Swedish

In the less-contrastive condition, the participant was presented with one representamen and four objects, which, as in the more-contrastive condition, were determined by the content of each task. In a less-contrastive condition for music with written words as objects (T2), for example, the representamina consisted of one melody (e.g. 'cat') and the objects consisted of the word for the cat, which was the target object, and the words for three of the foils (i.e. pig, ballerina and fighter) in either Swedish or Chinese. If the task was fictive words with shapes as objects (T3), the representamina consisted of one fictive word (e.g. [lamu]) and the objects consisted of one round, soft-like shape, which was the target object, and the three foils, which consisted of one sharp shape, a square and a circle. Figure 10 illustrates these examples as presented in the experiment.
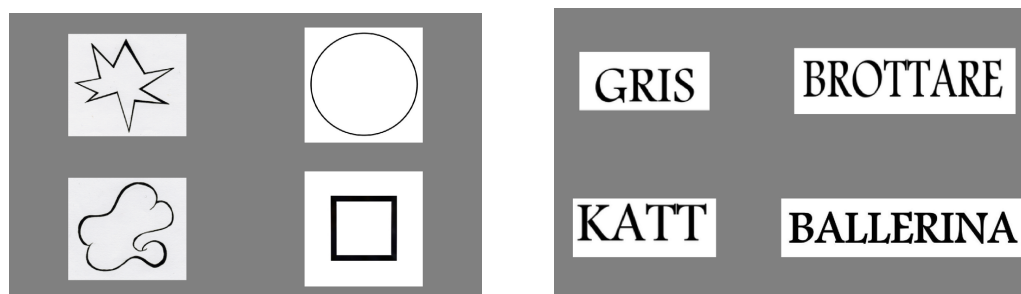
Figure 10. Examples of the less-contrastive condition (T2) presented in Swedish and the less-contrastive condition for fictive words (T3).

Intuitively, the less-contrastive condition is harder than the more-contrastive one. Theoretically, it can be argued that it requires a higher proportion of primary iconicity than secondary iconicity (see section 2.4), as the participant cannot use a simple exclusive inference in making the representamen-object matching (*à la* "this rather sounds like this image, so the other sound must be the other image").

In the more-contrastive condition, the combination of primary and secondary iconicity is as in Figure 5 (Section 2.2). In the less-contrastive condition, the process of semiosis consists of the same three steps, but differs in the number of candidate objects per representamen. This also consists of a combination of primary and secondary iconicity, but here the degree of secondary iconicity is not as high as it is in the more-contrastive condition, given that the participant has more choice possibilities. This process is shown in Figure 11.

STEP 1 STEP 2 STEP 3

R1 R1 R1

O1 O2 O3 O4   O1 O2 O3 O4   O1 O2 O3 O4
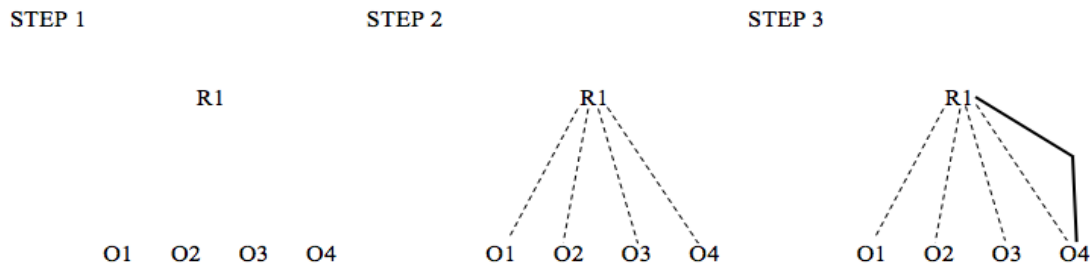
Figure 11. Matching representamina to objects by finding the most congruent iconic ground in the less-contrastive condition (adapted from Ahlner and Zlatev, 2010)

The experiment was designed and carried out in PsychoPy. The aim was that the participant was to receive all instructions from the program itself, so as to minimize the interaction between the participant and the researcher during the experiment. All instructions were presented in both speech and writing, in either Swedish or Chinese. These were translated and recorded by the same L1 speakers that produced the translations for the materials described in section 3.1.

Prior to both warmups and the actual experiment, a set of general instructions was presented, explaining that the participant was to match musical melodies and made-up words to objects shown on the screen. The instructions specified that they were to be presented with images in pairs or in groups of four, and that when presented with a pair of objects, they would listen to two melodies or made-up words, and when presented with a group of four objects, they would only listen to one melody or made-up word.

To minimize confusion, a set of instructions came before each task. Before less-contrastive conditions for T1, T2 and T4, the instructions stated that the following melody/word matched *one* of the four objects on the screen. Once the participant had read and listened to all instructions, they were presented with the four objects and the representamina. After the music, or the ideophone was finished playing, the instructions (in only auditory format) asked the participant to click on the object that best matched the representamen, which was presented one more time. Once the participant had listened to the music or the ideophone the second time, they were to click on their object of choice. In every task, once the participant made their choice and clicked on the object, the program automatically showed the following task, so they did not get to change their answers.

Before the more-contrastive condition for T1, T2 and T4, the instructions stated that each of the two melodies (or made-up words), matched each of the objects on the screen. Once the participant had read and listened to all instructions, they were presented with the two objects. In these tasks it was imperative to make a clear distinction between both representamina, so for this purpose, the instruction enounced 'recording 1' and 'recording 2' before each representamen was played. After the participant had listened to both melodies, she or he was instructed to click on one of the two representamina, which was then played a second time. In this case, it was taken for granted that, once the participant had clicked on their object of choice, it meant that they had matched the remaining representamina to the remaining object.

For the both conditions in T3, no written instructions were presented. Instead, in the less-contrastive condition, the four objects were shown as the audio instructions informed the participant that one of the figures was called *X* and subsequently asked the participant to click on the figured called *X*. In the more-contrastive condition, the two objects were shown as the audio instructions stated that one figure was called *X* and the other was called *Y,* subsequently asking the participant to click on the image called *Y*. Similarly, it was taken for granted that the clicking on an object called *Y,* meant that they had assigned the name *X* to the remaining figure (i.e. the exclusive inference mentioned above). The position of the objects in each task was randomized every time the experiment was run, as well as the order of the tasks, meaning that no participant would see the stimuli in the same order, thus minimizing chances for associations based on the order the material was presented.

An important aspect that contributed to the design of the experiment was the music stimuli, as it was essential that each participant listened to all six melodies, which was accomplished by presenting the two conditions per type of task. For the fictive words task, they were designed so that if the target figure in the *more-contrastive* condition was one with "soft" shapes, then the target figure in the *less-contrastive* condition would be one with "sharp" shapes.

Three different versions of the experiment were prepared, with different combinations of stimuli and conditions for the different tasks were designed. The visual presentation of all three tasks is presented in Appendix III.

3.3. Participants

To address the second research question concerning possible cultural differences in the perception of music, twenty-one L1 Swedish speakers, and twenty-one L1 Chinese speakers were recruited, using online advertising and personal contacts.[7] It was imperative that none of the participant had previously heard the piece *Peter and the Wolf,* as they would thus know the image or word that matched the melody. For this purpose, all of the potential participants were provided with Peter's melody, which is the most recognizable of all. Those who did not recognize the music were eligible for participation.[8] The only other criterion was that they were native speakers of Swedish or Mandarin Chinese, and that none spoke Basque.

The Chinese group consisted of 21 participants (8 male and 13 female), with an average age of 29. All of them were from mainland China and moved abroad after the age of 18. The Swedish group also consisted of 21 participants, out of which 8 were male and 13 were female, and the average age was 25. All the participants performed the experiment in Sweden, either in Malmö or Lund, and each session took between twenty and thirty minutes. All participants signed a consent form, where they were presented with general instructions and were informed that a part of the session would be audio-recorded. Upon completion, each participant received a cinema ticket as reward for their participation.

3.4. Procedure

Oral interaction with the participants was performed in English. Prior to each session, the researcher asked the participants to state their age, and the languages they spoke (to make sure they did not speak Basque). Participants were then informed that they would start with a warmup session, where they would have the chance to ask any questions or state doubts if necessary. The researcher also clarified that all instructions would be presented in their language. Once the two warm-up tasks were done, the researcher asked again if they had any questions or doubts before continuing. Following this, each participant was presented with the four tasks (in two conditions each) in random order. As stated before, the program provided all instructions, so the researcher did not interact with the participant at all while they were taking the experiment. Instead, my task was to keep track of their associations, for the purpose of the interview that was to follow. The experiment lasted around 8 minutes.

Following this, the participant was interviewed concerning the reasoning behind their choices. The participant was shown a slide-show version of the experiment that they had just run, in their respective language. The procedure here was simple: each slide provided them with the

---

[7] It is relevant to mention that both Mandarin Chinese, and to some degree Swedish, are *tonal* languages. This was not the reason for selecting these particular languages for the study, but could, as an anonymous reviewer noted, have a potential effect on the results. Thus, this is one more factor to consider in future studies.

[8] It is worth mentioning that during the recruitment process, there was a higher number of Swedes that recognized the music, whilst none of the Chinese (candidate) participants did so.

respective representamina to the objects shown. Once they had heard all the representamina one more time, the researcher reminded them of the object they had clicked on and proceeded to ask why they had chosen that specific one. In the case for the more-contrastive conditions, the researcher asked if they would have matched the remaining representamen to the remaining object, and if so, why. In the end, the researcher asked they participant if they had any questions, and provided them with a brief explanation of the purpose of the experiment. Upon completion, they were rewarded with a cinema ticket.

3.6. Summary and hypotheses

I started by presenting the materials employed, followed by the design of the experiment and the participants that partook in the experiment, finalizing with the procedure. On this basis, the four hypotheses were formulated:

- H1: Throughout Tasks 1-4, more-contrastive conditions will show more successful matching than less-contrastive conditions.
- H2: For the music tasks (T1 and T2), Swedish L1 speakers will show more successful matching than L1 Chinese speakers.
- H3: For the fictive words and ideophones tasks (T3, T4) no difference in the performance between the Swedish and Chinese participants is expected.
- H4: L1 participants of both languages are expected to perform better in the fictive words tasks (T3) than in the ideophones tasks (T4).

The first hypothesis is based on the assumption that in both programmatic music and speech, a combination of primary and secondary iconicity is responsible for succeeding in the tasks. However, the role of secondary iconicity is greater in the more-contrastive than the less-contrastive condition. The second hypothesis is based on the assumption that due to their greater familiarity with Western classical music Swedish participants would find it easier to perceive referential iconicity in the music tasks. The third hypothesis, on the other hand, expects no differences concerning the fictive words tasks (T3), as the type of iconicity presented in these tasks can be understood as reflecting universal sound symbolism (Imai and Kita, 2014). Given that the ideophones were in Basque, which was equally unfamiliar to both the Swedish and Chinese participants, no difference was expected here either. The fourth hypothesis predicted that all participants were expected to perform better in the fictive words tasks (T3) than in the ideophones tasks (T4), given that the first reflects universal human capacities in speech comprehension (Imai and Kita, 2014), while the latter involves word-forms that occur in an actual natural language, and are thus less exaggerated and more conventionalized (Dingemanse, 2018).

4. Results

All of the results were analyzed statistically, where each hypothesis was tested using mixed-effects logistic regression with random intercepts for participant and item. From the 42 participants who took part in the experiment, a total of 336 answers were gathered for all tasks (T1-T4), with 84 answers per task. These results are shown in Table 3.

Table 3. A summary of the results

Condition	Total number of choices	Incorrect	Correct
Linguistic	168	63 (37.5%)	**105** (62.5%)
Music	168	**72** (42.9%)	**96** (57.1%)
HYPOTHESIS 1			
Overall **Less-Contrastive**	168	**99** (58.9%)	**69** (41.1%)
Overall **More-Contrastive**	168	**36** (21.5%)	**132** (78.5%)
Less-Contrastive Music tasks (T1-T2)	84	**55** (65.5%)	**29** (34.5%)
More-Contrastive Music tasks (T1-T2)	84	**17** (20.1%)	**69** (79.9%)
Less-Contrastive Linguistic Tasks (T3-T4)	84	**44** (52.4%)	**40** (47.6%)
More-Contrastive Linguistic Tasks (T3-T4)	84	**19** (22.7%)	**65** (77.3%)
HYPOTHESIS 2			
Chinese Participants Music Tasks (T1-T2)	84	**38** (45.3%)	**46** (54.7%)
Swedish Participants Music Tasks (T1-T2)	84	**34** (40.5)	**50** (59.5%)
HYPOTHESIS 3			
Chinese Participants Fictive Words Task (T3)	42	**17** (40.5 %)	**25** (59.5%)
Swedish Participants Fictive Words Task (T3)	42	**11** (26.2%)	**31** (73.8%)
Chinese Participants Ideophones Task (T4)	42	**18** (42.9%)	**24** (57.1%)
Swedish Participants Ideophones Task (T4)	42	**17** (40.5%)	**25** (59.5%)
HYPOTHESIS 4			
Fictive Words (T3)	84	**28** (33.4%)	**56** (66.6%)
Ideophones (T4)	84	**35** (41.7%)	**49** (58.3%)

When looking at the overall results obtained for both linguistic and music tasks, a binomial test showed that the proportions of correct answers in music and linguistic tasks were both above chance (0.375), $p= 0.000$ (music tasks); $p= 0.000$ (linguistic tasks). The regression analysis further showed that the effect of condition on performance was non-significant: $\beta= -0.3494$, $z= -0.674$, $p= 0.5005$, indicating that, even though the participants performed slightly better at the linguistic tasks, there was no statistically significant difference between the two kinds of tasks. This can be further visualized in Figure 12, where the proportions of accurate answers for linguistic and music tasks are presented.
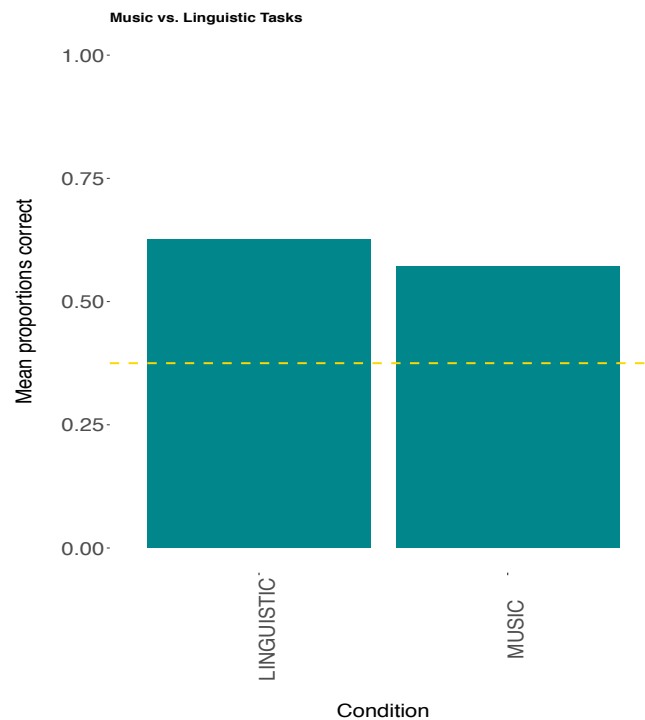
Figure 12. Proportions of accurate answers for both music and linguistic tasks. The golden dotted line represents the chance level

4.1. More-contrastive vs. less-contrastive conditions (H1)

In order to test the first hypothesis, concerning higher success rates for the more-contrastive than the less-contrastive condition, the total amount of answers for all participants were counted. Out of the total 336 replies gathered, there were 168 per condition. The effect of condition on performance was, as expected, highly significant: β= 2.0500, z= 6.684, p= 0.000, and thus supporting the first hypothesis. Furthermore, an exact binomial test indicated that the proportions of correct answers for both more and less-contrastive tasks were greater than chance: for the more contrastive tasks, this would be 0.5, as the participant had a 50-50 chance of making the correct association, where the p-value obtained was of p= 0.000. For the less-contrastive tasks, the participant had four options, meaning that the chance was 0.25, and p= 0.000. Looking at the music tasks (T1 and T2) separately, as expected, music tasks obtained much better results in the more-contrastive condition than in the less contrastive condition: β= 2.1052, z= 5.650, p= 0.000. The exact binomial test showed that both conditions were above chance significance (less-contrastive (0.25): p= 0.03248; more-contrastive (0.50): p= 0.000). The results obtained for the linguistic tasks (T3 and T4) were similar, with significant differences between the conditions: β= 1.8896, z= 3.391, p= 0.000695 and above chance significance (less-contrastive (0.25): p=0.000; more-contrastive (0.5): p= 0.000). These results are presented in Figure 13.
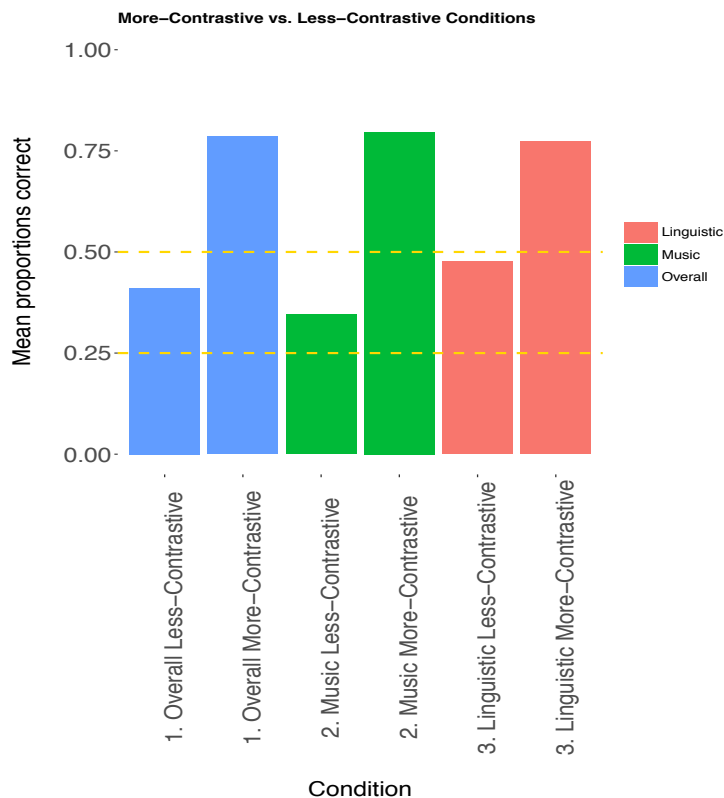
Figure 13. Proportions of accurate answers for less- and more-contrastive conditions for all tasks (T1-T4). The golden dotted line represents the chance levels (50% for more contrastive and 25% for less contrastive conditions).

4.2. L1 Swedish speakers vs. L1 Chinese speakers in music tasks (H2)

This hypothesis focuses only on music tasks (T1-T2), where a total of 168 answers were collected. In order to analyze this data, the total amount of answers for the two tasks performed by L1 Swedish speakers was compared to the total amount of answers provided by L1 Chinese speakers, where each group provided 84 responses. The overall results for both language groups are presented in Figure 14. The regression analysis showed that the difference was not significant: *ß*= 0.1992, *z*= 0.631, *p*= 0.528 and the exact binomial test showed that proportions of correct answers for the music tasks for Swedish and Chinese participants was above chance (0.375), *p*= 0.0009627 (Chinese), *p*= 0.000 (Swedish). Thus, H2 was not supported.

Music Tasks L1 Swedish vs. L1 Chinese
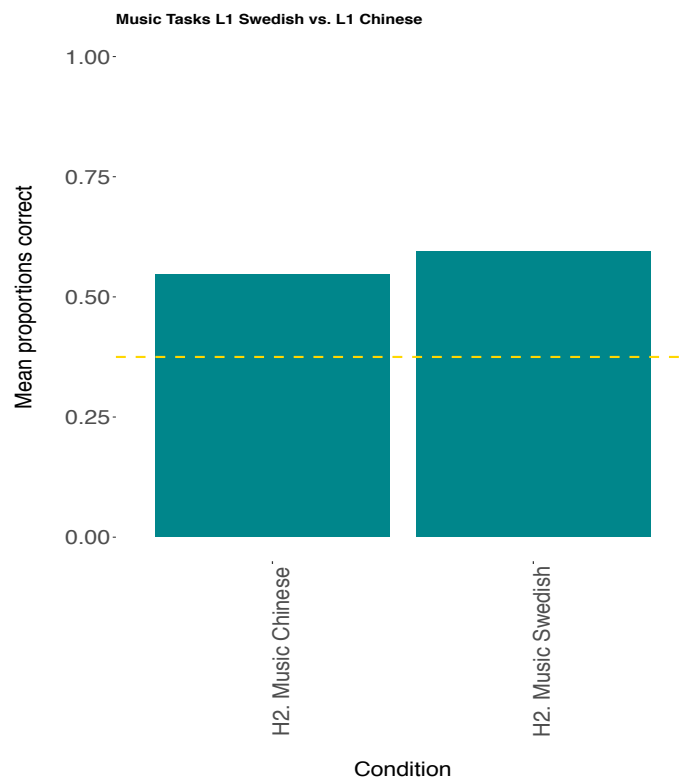
Figure 14. Proportions of accurate answers for L1 Swedish and L1 Chinese speakers in music tasks (T1-T2). The golden dotted line represents the chance level (0.375).

4.3. L1 Swedish speakers vs. L1 Chinese speakers (H3)

For the fictive words linguistic task (T3), a total of 84 answers were gathered. Similar to the previous hypothesis, the total amount of answers provided by participants of each group (42 answers per language group) were gathered and then compared. While the L1 Swedish performed somewhat better than L1 Chinese speakers, the regression analysis showed no significant difference between the two groups: *ß*= 0.7139, z= 1.447, *p*= 0.148. Also, the exact binomial test showed that proportions of correct answers for the fictive words task for both groups was above chance significance (0.375), *p*= 0.003085 (Chinese), *p*= 0.000 (Swedish).

      With respect to the ideophones task (T4), where a total of 84 answers were collected (42 per group), there was even less difference between the two language groups. According to the mixed-effects logistic regression, the difference was not significant (*ß*= 0.2269, z= 0.362, *p*= 0.717.) and the exact binomial test showed that proportions of correct answers for the ideophones task for Swedish and Chinese participants was above chance (0.375), *p*= 0.007564 (Chinese), *p*= 0.003085 (Swedish). Thus, concerning the matching of fictive words and ideophones, the hypothesis was supported. The results for this hypothesis are shown in Figure 15 below.
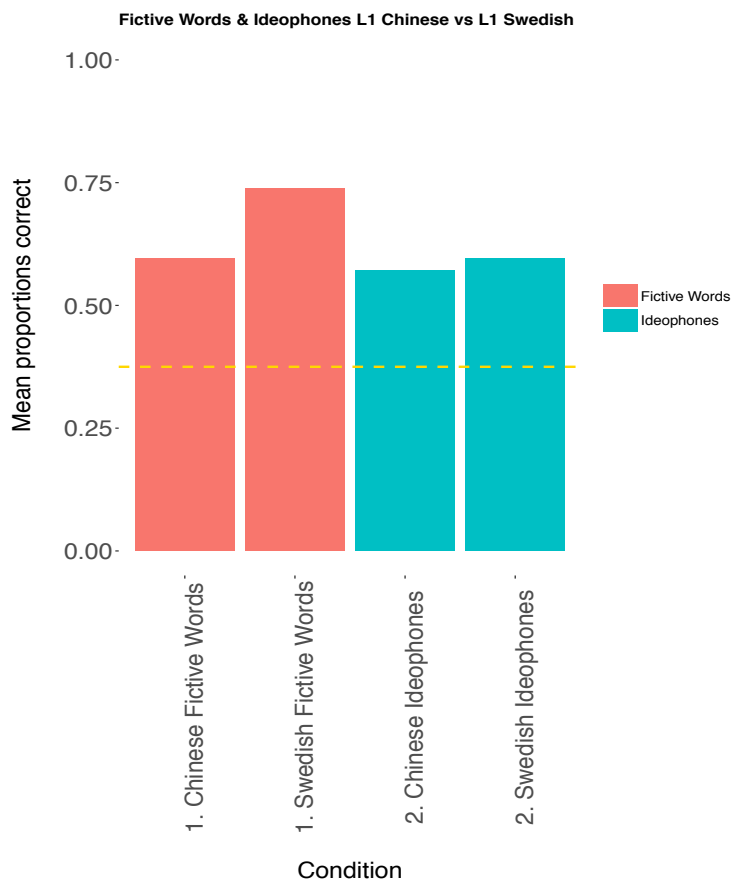
Figure 15. Proportions of Accurate answers for L1 Swedish and L1 Chinese speakers in the fictive words and ideophones tasks (T3-T4). The golden dotted line represents the chance level (0.375).

4.4. Fictive words vs. ideophones tasks (H4)

Similar to the previous hypothesis, this concerned both linguistic tasks (T3 and T4), with a total of 168 answers, 84 per task, but compared tasks rather than groups. It predicted that participants were to perform better in the fictive words task (T3) than in the ideophones task (T4). The results are shown in Figure 16.

While participants seemed to perform better in the fictive words than in the ideophones task, the regression analysis showed that this difference was not significant (*ß*= -0.6568, *z*=-0.700, *p*= 0.484). Furthermore, the exact binomial test showed that proportions of correct answers for the ideophones and fictive words tasks was above chance significance (0.375), *p*= 0.000 (Fictive Words), *p*= 0.000 (Ideophones). Thus, H4 was not supported.

Fictive Words vs. Ideophones
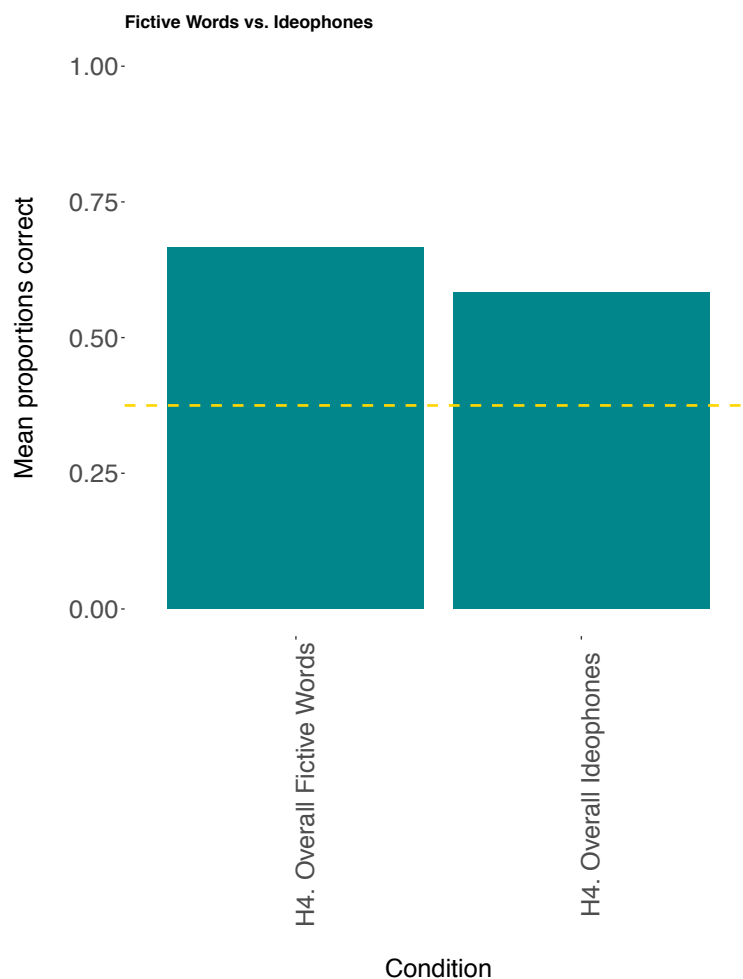
Figure 16. Proportions of accurate answers for fictive words and ideophones tasks (T3-T4). The golden dotted line represents the chance level (0.375).

4.5. Summary

Out of the four hypotheses, two were supported: H1 concerning the conditions more or less contrastive, and H4 concerning lack of differences between the Swedish and Chinese groups in the linguistic tasks. Importantly, however, the results for even those hypotheses that were not supported yield an important result: referential iconicity played an important role in solving the respective tasks, irrespective of cultural group (and thus familiarity with Western classical music).

The results for H1, which predicted a statistically significant difference between more-contrastive and less-contrastive conditions showed that, indeed, more-contrastive tasks were more transparent to participants of both language groups. H2 predicted an advantage for the Swedish group in the music tasks (since they involved fragments from Western classical music), but contrary to the expectations, both groups performed similarly. Concerning both linguistic tasks (fictive words and ideophones), H3 anticipated no difference between the groups, which was indeed the case. On the other hand, H4 predicted that the overall result of the fictive words task were to show a higher success rate than in the ideophone task, which was not the case, indicating that the conventional and language-specific character of the Basque ideophones did not appear to conceal their iconicity for the participants. Perhaps most importantly, the results showed that the participants of both language groups solved both music and linguistic tasks equally well, and that

there was no difference between the case where objects were represented by images (T1) and by words (T2), indicating that the use of (schematic) images was reliable.

5. Discussion

5.1 Primary and secondary iconicity

The process of interpreting referential iconicity in speech and music was expected to show a combination of both primary iconicity (from ground perception to sign recognition), and secondary iconicity (from sign knowledge to ground recognition). But the roles of each kind of process were expected to vary, depending on the type of task, and possibly on how transparent the representamen-object relationship is.

With respect to the type-of-task factor, the more-contrastive condition operationalized a higher role for secondary iconicity (Figure 9), as participants were told that they are presented with two signs, and only have to decide which of the two possible mappings is more "natural". The less-contrastive condition (Figure 10), on the other hand, was an operationalization of a greater role for primary iconicity, as participants have to choose between four possible objects for a single representamen, and make this decision on the basis of the perception of the "best" iconic ground. The hypothesis (H1) that predicted that participants of both language groups would be more successful in the more-contrastive conditions (for all four tasks) than in less-contrastive conditions, was clearly supported. This is in line with Sonesson's conjecture that (referential) iconicity in both speech and musical signs is predominantly of the secondary kind. Still, as even the less-contrastive condition gave rise to success rates that were significantly above chance. This is the clearest indication that the referential iconicity in question is not only secondary, but involves a *combination* with primary iconicity.

Another expected difference that could be seen as a reflection of different proportions of primary/secondary iconicity was however not supported: that the unconventional fictive words would be more transparent than the more conventionalized ideophones. Thus, this "non-difference" can be interpreted as more or less equal roles for both primary and secondary iconicity, at least in the case of the particular stimuli that were used in the experiment.

In sum, the results of the study support the hypothesis that in both speech and music signs, there is a combination of primary and secondary iconicity (Ahlner and Zlatev, 2010), but with a higher degree of the latter (Sonesson, 2009). As pointed out in Section 2, this is rather an example of Jakobson's (1965) understanding of the sign itself as not being exclusively based on a single kind of ground, but rather on a combination of grounds, where one is (usually) predominant.

5.2. Cultural differences and conventionality in music

Given the differences between Chinese musical traditions and Western musical traditions emphasized by Shen (1991, 2000), and considering that the Western classical music is much more prevalent and accessible in Sweden than China, led to the prediction that Swedish participants were to perform better in the two music tasks (T1 and T2). Interestingly, this hypothesis was not supported, as there was no significant difference between the success rates of the two language groups, and further that both were above chance. In addition, there were no significant differences between the two music tasks (linking to images or words denoting the same referents, respectively), indicating that the images used were not biased towards Swedish culture. This implies that familiarity with musical conventions is not a prerequisite for perceiving referential iconicity in programmatic music, in the broad sense of the term, as briefly discussed in the introduction.

The lack of cultural differences in the music tasks could be due to Prokofiev's specific choice of interpretation of each character in a way that is culture-general. Furthermore, the Chinese speakers that participated in the study could perhaps be seen as non-representative, as all lived in Sweden, and could have been familiarized with the Western musical tradition. Alternatively, it could be the case that despite there being a great deal of culture-specific conventionality, the way we understand and listen to music is not directly linked to each individual's culture, which could point to a universal aspects in music perception. Thus, it is important to highlight that the present findings apply to this *particular* music piece, and in order to choose between one of these interpretations, it would be necessary to investigate a more varied series of programmatic music, with more "mono-cultural" participants.

Further, despite the fact that this general hypothesis concerns the music tasks, it is pertinent to discuss the lack of cultural differences, also found in the linguistic tasks involving fictive words and ideophones. As was stated in Section 4.3, despite Swedish participants performing better than Chinese participants, the mixed-effects logistic regression analysis showed no significant difference between the two language groups. This finding stands in line with Imai and Kita's (2014, p.5) proposal of "universal sound symbolism", but extending this to the Basque ideophones used in the study as well, and not only the fictive words, as there were no significant differences between these tasks:

> There has been an assumption in the literature that sound symbolism is universal; if a certain sound–meaning correspondence is identified by speakers of one language, this should be generalizable to speakers of any other languages. This assumption has been supported by the fact that speakers of many different languages sense Köhler's shape sound symbolism [26] in the same way, as reviewed earlier (for English [32], Japanese [68,69], Himba [27], Kitwonge-Swahili bilinguals [70]).[9]

Nevertheless, it is necessary to mention that many Chinese participants seemed to have difficulties with the fictive words task (T3), as they questioned the formulation of the instructions (*One of these figures is called e.g. lulu and the other is called e.g. keti. Click on the image called e.g. lulu*) during the interviews. Their difficulty was in that they could not comprehend why they were to name an abstract shape. In particular, many participants picked the "blobby" figure despite listening to a word with "sharp" consonants and vowels, since it "looked more alive", as one participant stated. Interestingly, even though Swedish participants did not seem as confused by this task, some participants also chose the blobby figure because of its greater degree of *animacy* (known to be a relevant factor in linguistics), as expressed by Yamamoto (2006, p.30): "only animal beings can be agents in a normal sense […] the agency concept goes hand in hand with that of animacy, and that both notions are highly significant determiners of mind-style or world view".

---

[9] The references in this citation are the following: [26] Köhler (1929); [32] Thompson and Estes (2011); [68, 69] Asano et.al (in review) and Miyazaki et. al (2013); [27] Bremner et. al (2013); [70] Davis (1961).

6. Conclusions

The study of non-arbitrariness in speech, often under the label of "sound symbolism" is not new to the fields of linguistics and cognitive semiotics. However, the linkage between iconicity in speech to iconicity in music has not been widely studied, especially not through empirical research. In order to gather relevant empirical data, an experiment was designed, where participants were to match representamina of either musical or linguistic nature to different objects, represented either by images or words.

In this study there was no significant difference in the way participants perceived iconicity in speech in comparison to how they perceived iconicity in this specific piece of programmatic music. This could mean that the cognitive processes used for the perception of iconicity in speech and in music are domain-general, meaning that these processes involved are not limited to a single sensory modality or semiotic system (i.e. language vs. music), but rather that they cut across these. This could be taken as an indication that not only language, but also other semiotic systems like music rest on the universal human capacity for *bodily mimesis* (Donald, 1991; Zlatev, 2019).

The understanding of music as a semiotic system has not been explored sufficiently in previous research. This does not mean, of course, that there have been no studies that delve into the relationship between music and language (e.g. Coker, 1972; Osmond-Smith, 1972; Lerdahl and Jackendoff, 1983; Monelle, 1991). This article has focused on the referential aspects of music, which could be considered as a peripheral property of music but is sufficient to validate the claim that the similarity between music and language lies not only in comparable syntactic structures. This take on the referential aspects of music may be a small, but still very relevant part out of the many possible ways we can investigate the relations between music and language. The similar results obtained between music and linguistic tasks show that there is indeed a similarity in the way referential iconicity in music works in relation to language. These results will only help to construct a more exhaustive understanding of music as a semiotic system in relation to language and will further help us to steer future research into other possible similarities, or differences, between these two semiotic systems.

Is iconicity in programmatic music and speech, primary, secondary, or a combination of both? Through the present we may conclude, in line with the proposal of Ahlner and Zlatev (2010), that iconicity in programmatic music and speech involves a combination of both primary and secondary iconicity, but with a higher degree of secondary iconicity. This was made clear through the more and less-contrastive conditions employed in the experiment, where the more-contrastive conditions showed higher degrees of secondary iconicity, given that they were only presented with two representamina and two objects. Still, the overall results of the less-contrastive condition showed above chance significance. The positing of a *specific* sign relations between object and representamen on the basis of the perception of the iconic ground between them is a clear indicator of primary iconicity (Ahlner and Zlatev, 2010), and thus, supports the proposal that the two kinds of iconicity are not mutually exclusive but may be combined in an act of interpretation.

Finally, to what extent is referential iconicity in music and speech perceivable by members of different cultures? No cultural differences were found in the perception of iconicity in either music or speech between the Swedish and Chinese speakers, which could point to universal human cognitive semiotic-capacities, as mentioned above. Some differences concerning the linguistic tasks were found in the interviews, with Chinese speakers more often objecting to the "naming" of abstract figures. These, however, did not lead to significant quantitative differences in performance in the experiment. There are, however factors that could have affected these results, such as the fact that all Chinese speakers live in Sweden, they all speak English, and have most likely been exposed to Western music traditions. It would thus be interesting to test monolingual speakers that do not live abroad and have not been exposed to so many different musical cultures.

In sum, the study of referential iconicity in music and speech is an important topic for cognitive semiotics and beyond, because it allows for a more concrete understanding of how members of different cultures understand and interpret music and speech. While there have been a wide range of studies touching upon the subject of iconicity in language, not so many have dealt with iconicity in music, and this article contributes to the further understanding of the semiotic and cultural aspects of this phenomenon, across semiotic systems. More generally, this article has provided a new angle to the understanding of music as a semiotic system. It is evident that it has only covered a narrow area out of all the possible connections that can be made between music and language, but nevertheless, the significant results obtained can hopefully contribute to further research on meaning making in music.

Acknowledgments

References

Ahlner, F. and Zlatev, J. (2010). Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign System Studies*, 38 (1/4): 298-348.

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H. and Thierry, G. In review. *Sound symbolism scaffolds language development in preverbal infants.*

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays.* USA: University of Texas Press.

Bernstein, L. (1976). *The Unanswered Question, Six Talks at Harvard*. Cambridge, Mass: Harvard University Press.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J. and Spence, C. (2013). 'Bouba' and 'Kiki' in Namibia? A remote culture makes similar shape– sound matches, but different shape–taste matches to Westerners. *Cognition*, 126: 165-172.

Chen, J. (2015). Comparative study between Chinese and western music aesthetics and culture. *Atlantis Press:* 117-119.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, Mass: MIT Press.

Coker, W. (1972). *Music and meaning: A theoretical introduction to musical aesthetics*. New York: Collier Macmillan.

Cuskley, C., Simner, J. and Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect. *Psychological research,* 81(1): 119-130.

Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *Br. J. Psychol*, 52: 259-268.

Devylder, S. (2018). Diagrammatic iconicity explains asymmetries in Paamese possessive Constructions. *Cognitive Linguistics,* 29(2): 313-348.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and linguistics compass,* 6(10): 654-672.

Dingemanse, M. (2018). Redrawing the margins of language: Lessons from research on ideophones. *Glossa: a journal of general linguistics*, 3(1): 1-30.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of human culture*. Cambridge, MA: Harvard University Press.

Drake, C. and El Heni, J. B. (2001). The quest for universals in temporal processing in music. In I. Peretz and R. J. Zatorre (Eds.) *The cognitive neuroscience of music.* NY: Oxford University Press.

Drake, C. and El Heni, J. B. (2003). Synchronizing with music: intercultural differences. *Ann. N.Y. Acad. Sci*, 999: 429-437.

Editors. (1998). "Program Music". *Encyclopedia Britannica*. Web. 2 September 2018. www.britannica.com/art/program-music

HaCohen, R. and Wagner, N. (1997). The communicative force of Wagner's leitmotifs: Complementary relationships between their connotations and denotations. *Music Perception,* 14(4): 445-476.

Ibarretxe-Antuñano, I. (2017). Basque ideophones from a typological perspective. *Canadian Journal of Linguistics,* 62(2):196-220.

Imai, M., Kita, S., Nagumo, S. and Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109, 54-65.

Imai, M. and Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical transactions of the royal society B, 369*.

Jakobson, R. (1965). Quest for the essence of language. *Diogenes,* 13: 21-38.

Johansson, N. and Zlatev, J. (2013). Motivations for Sound Symbolism in Spatial Deixis: A typological study of 101 languages. *PJOS,* 5(1): 3-20.

Keiler, A. R. (1978). Bernstein's the unanswered question and the problem of musical competence. *Musical Quarterly,* 64: 195-222

Keiler, A. R. (1981). Two views of musical semiotics. Some properties of the design and syntax of tonal music. *SML*, 151-168

Kivy, P. (2002). *Introduction to a philosophy of music.* Oxford, UK: Oxford University Press.

Kivy, P. (2007). *Music, language and cognition.* New York: Clarendon Press.

Köhler, W. (1929). *Gestalt psychology*. New York, NY: Liveright.

Lerdahl, F. and Jackendoff, R. (1983). *A Generative theory of tonal music.* Cambridge, Mass: MIT Pres

Louhema, K. (2018). From Unisemiotic to Polysemiotic Narratives: Translating across semiotic systems. MA Thesis, Lund University.

Lynch, M. P. and Eilers, R. E. (1992). A study of perceptual development for musical tuning. *Percept. Psychophys*, 52: 599-608.

Maurer, D., Pathman, T. and Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science*, 9(3): 316-322.

Miyazaki, M., Hidaka, S., Imai, M., Yeung, H. H., Kantartzis, K., Okada, H. and Kita, S. (2013). The facilitatory role of sound symbolism in infant word learning. In Proc. 35th Annual Conf. of the Cognitive Science Society, Berlin, Germany, 31 July–3 August 2013 (Eds. M. Knauff., M. Pauen., N. Sebanz and I. Wachsmuth), pp.3080-3085. Austin, TX: Cognitive Science Society.

Monelle, R. (1991). *Linguistics and semiotics in music*. London: Routledge.

Nattiez, J. J. (1987). *Musicologie generale et semiologie*. Paris: Bourgois

Osmond-Smith, D. (1972). The Iconic process in musical communication. *VS, Quaderm di Studi Semiotici*, 31-42.

Ohala, J. J. (1994). The frequency code underlines the sound-symbolic use of voice pitch. In: L. Hinton, J. Nichols and J. J, Ohala (Eds.) *Sound Symbolism.* Cambridge: Cambridge University Press.

Ohala, J. J. (1997). Sound Symbolism. Proc. 4th Seoul International Conference on Linguistics [SICOL] 11-15 Aug 1997.

Patel, A. D. (2008). *Music, language and the brain*. New York: Oxford University Press.

Peirce, C. S. (1974 [1931]). The Icon, Index, and Symbol. In C. Hartshorne and P. Weiss (Eds.), *Collected papers of Charles Sanders Peirce*. Cambridge: Harvard University Press.

Peirce, C. S. (2003). Basic concepts of Peircean sign theory. In: M, Gottdiener., K, Boklund-Lagopoulou and A, Lagopoulos (Eds.), *Semiotics. Vol. 1, Part one. Fundamentals: The constitution of the field.* London: Sage publications, 101-135.

Piaget, J. (1971). *Structuralism* (C. Maschler, Trans.). London: Routledge.

Ramachandran, V. S. and Hubbard, E. M. (2001). Synesthesia - A Window into Perception, Thought and Language. *Journal of Consciousness Studies*, 8: 3-34.

Saussure, F. (1959[1916]). *Course in General Linguistics*. New York: The Philosophical Library.

Shen, S. (1991). *Chinese music and orchestration: A primer on principles and practice.* Chicago: Chinese Music Society of North America.

Shen, S. (2000). *China: A journey into its musical art.* Chicago: Chinese Music Society of North America.

Sonesson, G. (1997). The ecological foundations of iconicity. In: I, Rauch and G. F, Carr (Eds.), *Semiotics around the world: Synthesis in diversity* (739-742). Proceedings of the Fifth International Congress of the IASS, Berkeley, June 12-18, 1994. Berlin and New York: Mouton de Gruyter.

Sonesson, G. (2009). Prolegomena to a general theory of iconicity. Considerations on language, gesture, and pictures. In: K, Willems and L, De Cuypere (Eds.), *Naturalness and Iconicity in Language* (47-72). Amsterdam: John Benjamins.

Stampoulidis, G. (2019). Stories of Resistance in Greek Street Art: A Cognitive-Semiotic Approach. *The Public Journal of Semiotics*, 8(2): 29-48.

Thompson, P. D. and Estes Z. (2011). Sound symbolic naming of novel objects is a graded function. *Q. Exp. Psychol*, 64: 2392-2404.

Tien, A. (2015). *The Semantics of Chinese Music: Analysing Selected Chinese Musical Concepts.* Amsterdam: John Benjamins.

Vigliocco, G., Perniss, P. and Vinson, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Phil. Trans. R. Soc. B, 369*(1651), 20130292.

Yamamoto, M. (2006). *Agency and impersonality: Their linguistic and cultural manifestations*. Amsterdam: John Benjamins.

Zlatev, J. (2014). Human uniqueness, bodily mimesis and the evolution of language. *Humana. Mente Journal of Philosophical Studies,* 27: 197-219.

Zlatev, J. (2015). Cognitive semiotics. In P. Trifonas (Ed.), *International handbook of semiotics*, 1043–1067. Dordrecht, Netherlands: Springer.

Zlatev, J. (2019). Mimesis, learning and polysemiotic communication. In M. A. Peters (Ed.) *Encyclopedia of Educational Philosophy and Theory*. Berlin: Springer.

Zlatev, J., Sonesson, G. and Konderak, P. (Eds.) (2016). Introduction: Cognitive semiotics comes of age. In J. Zlatev., G. Sonesson and P. Konderak (Eds.), *Meaning, Mind and Communication* (9-28). Frankfurt: Peter Lang.

Non-academic references

Lane, E. W.trans. (1839). *The Thousand and One Nights Commonly Called the Arabian Nights Entertainments.* 32 Pamphlets [3 Vols]. London: Charles Knight.

Rifkin, D. (2018). Visualizing Peter: The first animated adaptations of Prokofiev's Peter and the Wolf. *Society for music theory,* 24(2): 43-56.

Saint-Exupéry, A. (1943 [1987]). *Le petit prince.* France: Gallimard.

Young, R. W. (1939). Terminology for logarithmic frequency units. *The Journal of the acoustic society of America,* 11: 134-139.

Musical References

Djawadi, R. (2015). *Dance of Dragons.*
Prokofiev, S. (1936). *Peter and the Wolf.*
Saint-Saëns, C. (1886). *Le Carnaval des Animaux.*
Rimsky-Korsakov, N. (1899-1900). *Flight of the Bumblebee.*
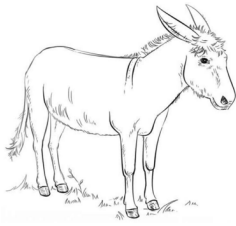Rimsky-Korsakov, N. (1888). *Scheherazade.*
Vivaldi, A. (1721). *Four Seasons.*

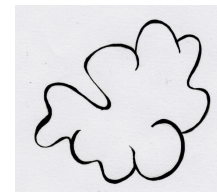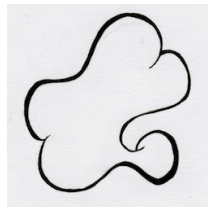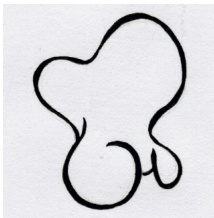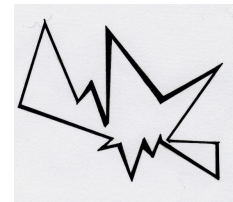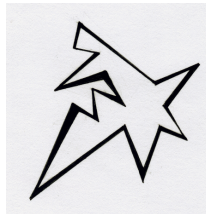**Appendix I. Images.**

1. The six characters in "Peter and the Wolf"
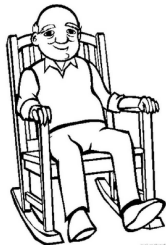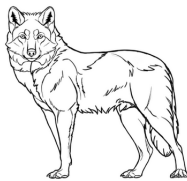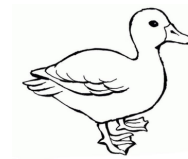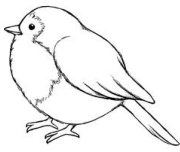
2. Foils for the music tasks
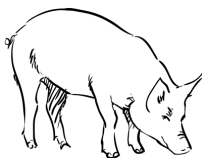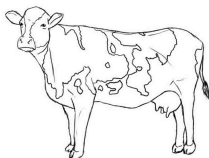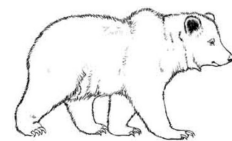
3. Warm-up images for music tasks
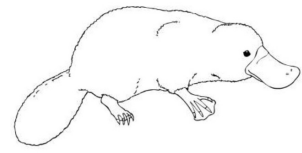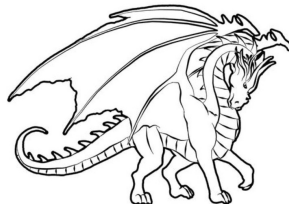
4. Images for the fictive words tasks
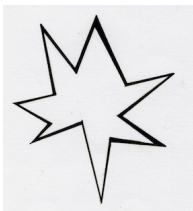
5. Foils for the fictive words tasks
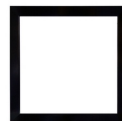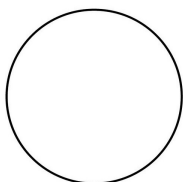
6. Warm-up images for fictive words tasks

Appendix II. Written words.

1. Words for music tasks.

<u>Swedish</u>:
<u>Peter and the Wolf Characters</u>: (Bird, Cat, Duck, Grandfather, Hunter and Wolf respectively)

FÅGEL   KATT   ANKA

MORFAR   JÄGARE   VARG

<u>Chinese</u>:
<u>Peter and the Wolf Characters</u>: (Bird, Cat, Duck, Grandfather, Hunter and Wolf respectively)

鸟   猫   鸭子

外公   猎人   狼

2. Foils for music task (Ballerina, Bear, Cow, Fighter, Pig and Squirrel respectively)

Swedish:

BALLERINA
KO
GRIS

BJÖRN
BROTTARE
EKORRE

Chinese:

芭蕾女演员

牛

猪

熊

格斗士

松鼠

3. Ideophones

Swedish: (Pil-pil, Grik-grak, Draka-draka, Trinkili-trankala, Zirris-zarraz and Zorro-zorro respectively)

LJUDET AV
KOKANDE
VATTEN

ATT SPRAKA

GALOPPERANDE
HÄST

RÖRA SIG
HÖGLJUTT MED
SVÅRIGHETER

LJUDET AV ATT
NÅGON SÅGAR

SNARKANDE

<u>Chinese</u>: (Pil-pil, Grik-grak, Draka-draka, Trinkili-trankala, Zirris-Zarraz and Zorro-Zorro respectively)

烧水声   发出噼啪声

马蹄声   行动困难声

锯木声   呼噜声

4. Foils. (To groan, whispering and to splash respectively)

<u>Swedish:</u>

ATT STÖNA   ATT VISKA

PLASK

<u>Chinese:</u>
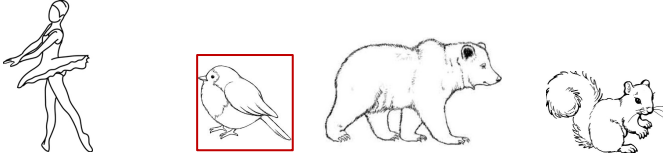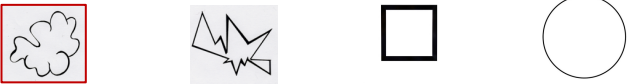
呻吟声   低语声

泼溅声
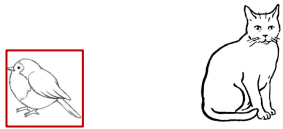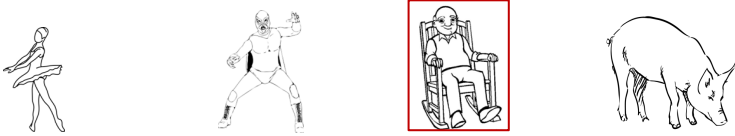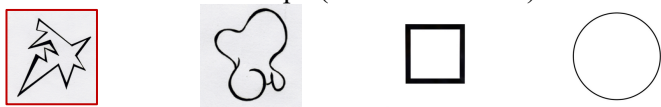
Appendix II. Organization.

For the three versions of the experiment (in Swedish).
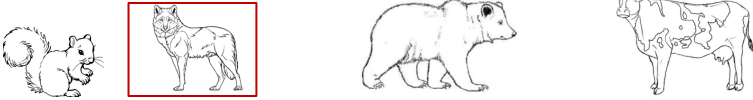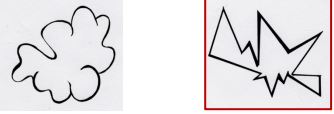
Version 1

T1: Music to Image (More-Contrastive)
T1: Music to Image (Less-Contrastive)
T2: Music to Word ('grandfather' and 'cat' respectively) (More-Contrastive) MORFAR KATT
T2: Music to Word ('pig', 'cow', 'fighter' and 'hunters', respectively) (Less-Contrastive) GRIS KO BROTTARE JÄGARE
T3: Fictive word to Shape (More-Contrastive)
T3: Fictive word to Shape (Less-Contrastive)
T4: Ideophones to Word ('sound of sawing' and 'sound of boiling water') (More-Contrastive) LJUDET AV ATT NÅGON SÅGAR   LJUDET AV KOKANDE VATTEN
T4: Ideophones to Word ('splash', 'to whisper', 'to groan' and 'galloping horse' respectively) (Less-Contrastive) PLASK ATT VISKA ATT STÖNA GALOPPERANDE HÄST

Version 2

T1: Music to Image (More-Contrastive)
T1: Music to Image (Less-Contrastive)
T2: Music to Word ('hunter' and 'wolf' respectively) (More-Contrastive) JÄGARE VARG
T2: Music to Word ('cow', 'duck', 'bear' and 'squirrel' respectively) (Less-Contrastive) KO ANKA BJÖRN EKORRE
T3: Fictive word to Shape (More-Contrastive)
T3: Fictive word to Shape (Less-Contrastive)
T4: Ideophones to Word ('galloping horse' and 'sound of boiling water' respectively) (More-Contrastive) GALOPPERANDE HÄST   RÖRA SIG HÖGLJUTT MED SVÅRIGHETER
T4: Ideophones to Word ('splash', 'to whisper', 'to groan' and 'to crackle' respectively) (Less-Contrastive) PLASK   ATT VISKA   ATT STÖNA   ATT SPRAKA

Version 3

T1: Music to Image (More-Contrastive)
T1: Music to Image (Less-Contrastive)
T2: Music to Word ('duck' and 'bird' respectively) (More-Contrastive) ANKA   FÅGEL
T2: Music to Word ('pig', 'cat', 'ballerina' and 'fighter' respectively) (Less-Contrastive) GRIS   KATT   BALLERINA   BROTTARE
T3: Fictive word to Shape (More-Contrastive)
Speech/ Non-Words/ Less-Contrastive (Target: Lamu)
T4: Ideophones to Word ('to crackle' and 'snoring' respectively) (More-Contrastive) ATT SPRAKA   SNARKANDE
T4: Ideophones to Word ('splash', 'to whisper', 'to groan' and 'sound of sawing' respectively) (Less-Contrastive) PLASK ATT VISKA   LJUDET AV ATT NÅGON SÅGAR   ATT STÖNA

Author Address
Email: veronica.giraldo88@gmail.com

About the author:
Verónica Giraldo's academic background is in music and linguistics. She holds an MA in Language and Linguistics with specialization in Cognitive Semiotics from Lund University. The work presented derives from her master's thesis project. One of her main interest is exploration of the possible correlations between language and music from the perspective of cognitive semiotics. She is currently researching on how to make visual art and museums more accessible to the blind and visually impaired community.