SYNESTHETIC ASSOCIATIONS BETWEEN EARLY ACQUIRED CONCEPTS

by

Aida Isteliyeva

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Arts in

Eurasian Studies

at

NAZARBAYEV UNIVERSITY -
SCHOOL OF SCIENCES AND HUMANITIES

2025

THESIS APPROVAL FORM
NAZARBAYEV UNIVERSITY
SCHOOL OF SCIENCES AND HUMANITIES

SYNESTHETIC ASSOCIATIONS BETWEEN EARLY ACQUIRED CONCEPTS

BY

Aida Isteliyeva

NU Student Number: 201929099

APPROVED

BY

Dr. VALENTINA APRESSYAN ASSOCIATE PROFESSOR

ON

The 26th of May, 2025

Signature of Principal Thesis Adviser

In Agreement with Thesis Advisory Committee
Second Adviser: Dr. Mihnea Capraru Assistant Professor

Abstract

The seemingly unexplainable associations between such early acquired semantic categories as colors, days of the week, months, shapes, and digits are an underresearched topic in the research on synesthesia. This work fills this gap in knowledge and delves into the possible reasons for particular pairings to be created. Previous research claims that similar types of synesthesia come from such things as childhood experiences, frequency of word use, and time of acquisition. To research this topic, the work looked into the existing reports of such synesthetic experiences; collected new reports in English, Russian, and Kazakh; and conducted two experiments: to check whether synesthete participants have stable association pairs, and whether the frequency of the words could predict their likelihood to be picked as a pair. The results of the work show that the creation of the association pair could depend on all of the previously mentioned factors. The dependency seems to be hierarchical, meaning that some factors take precedence in determining the pairing. This way, the associations that appear out of constant life experiences seem to not be affected by other factors, however, they only cover a small portion of existing association pairs. The frequency with which the words are used – and arguably their prototypicality – are what determines a large part of the association pairs and to a statistically significant extent. The work also finds that the association pairs are transferable from L1 to L2 and that there are particular trends to what items get paired up the most. Such frequent pairings are constant across participants and languages. While this work does base its findings on bilingual speakers of Kazakh and Russian, a comparison to a English corpus of pre-existing reports of association pairs shows stability of the general trends.

Table of Contents

1. Introduction

1.1 Background information

The Oxford Handbook of Synesthesia defines synesthesia to be a "neuropsychological condition which gives rise to extraordinary sensations" (2013, p.xxi). What this means is better explained by an example. A person listening to music also hears a melody that is not present in reality. A person looking at a text sees the letters in different colors. A person touching a material also smells something associated particularly with this sensation. All these are instances of synesthesia wherein a person receives an additional experience that is unrelated to the sensation that is actually being experienced. A study that created a standardized synesthesia test provides a list of 20 types of synesthesia that are possible to test for online (Eagleman, et al., 2007). Although the prior definitions included only the five main senses, recently, with an addition of the research in the field, there has been an update to include such aspects as word meanings and personification. This work focuses on the kind of synesthesia that fits under the wider definition. The synesthetic associations under investigation will contain items from at least two of the following semantic groups: between colors, days of the week, months, letters, digits, mathematic equations and shapes.

For the lack of an established term for this particular kind of synesthesia and for further convenience, this work will reference this system of synesthetic associations as Network Synesthesia (NS). While most of the existing work looks into associations between two items or senses, this work will focus on connections between more than that – an association network of three and more items – hence the name of the phenomenon under consideration. This work's main objective will be to explore the possible causes behind pairings and groups of items that get created by NS. This work also investigates the interaction between NS and bilingualism. The main methodological part of the work is happening in Kazakhstan which is a predominantly

multilingual population – thus allowing us to look into manifestation of synesthesia in more than one language.

This particular kind of synesthesia, although often reported by people online, is not present in the current discussion on synesthesia and is understudied. Additionally, as per my research, synesthesia is rarely – if at all – studied in the context of Kazakhstan and its multilingual situation. This two aspects make this research novel and relevant to the development of the research field.

This work will first focus on the existing literature on synesthesia to establish some of the terminology, as well as to explain the methodology of the work and the types of analyses that will be applied on the data. Next, it will look into the three methodologies that will be utilized in this work and their results. The first will be a corpus of manifestations of NS that were reported online. The second will be a questionnaire conducted among NU students to find participants and to establish their relation to NS: whether they are synesthetes, and what NS associations they have. The third will be an experiment targeted at testing one of the possible motivations behind creation of associations – namely, frequency of use. The work will then continue to a comparison of the results acquired through different methods.

1.2 Research questions

The main research question of this work is on the motivations of the particular pairings. I am interested to see whether some items get grouped together more often than others. And if it so happens, the question becomes why they are grouped together. The literature on synesthesia and associations in general contains several possible factors that could affect association formation. Some of these factors are: similarity in phonological shape of the associated items, collocation, creation of association based on personal experience, effects of an educational setting, and frequency or prototypicality of an item among its semantic group. Similarity in sound can be illustrated with such an example as *черный* ‘black’ and *четверг* ‘Thursday’. The

two words start with the same consonant which might create an association between them. Such a synesthetic connection will lead to people seeing black in their mind's eye while hearing or reading *четверг*. In absence of such an experience, this is a simple association and does not count to be an instance of NS. The cases of collocation – such as 'black Friday' – also should not be considered under NS as they are cases of simple association. Prototypicality and frequency refer to similar things in my methodology. What it means is that in a semantic group – for example colors – there will be items that will be mentioned more or less in our life. This way, black, white, red, and green are the most frequently mentioned colors in corpora of languages. They are also the more prototypical or common colors. As an apple is a more prototypical fruit than a kiwi, red is a more prototypical color than purple (Rosch, 1977).

Therefore the main research question is: "What are the factors that affect the pairings/groupings associated in NS?" This research question calls for a series of additional secondary questions. Besides, as stated above, the work wants to look into more than just the possible factors. Rather there are additional questions that will be asked along the way that indirectly help answer the main question. This way, as this research is going to analyze data of multilingual participants, we might be able to infer whether the associations happen on the level of words or concepts, as we will be able to see whether the participants name the same pairings for different languages that they know. If it happens that the associations are the same across languages for a participant, we will be able to infer that the association happens on the level of concepts, since, otherwise, the difference in the form across languages would have pushed the participants to either not have an association in other languages, or have different association pairings in different languages. The following list presents the secondary questions of the work:

1. What NS pairs/groups do people report having?
2. Are there any pairings/groupings that appear most frequently across participants?
3. Are there any trends in associations across languages?

4. Are associations stable across languages in multilingual speakers?

5. Do non-synesthetes report having associations that are similar to reports of those with synesthesia?

6. Do items in association share any of the following features: phonological form, semantic collocation, frequency of use?

7. Does the frequency of use/prototypicality of items predict their likelihood to be associated?

8. Does the synesthetic association happen on the level of words or concepts?

The first objective of this work is to collect data on associations across different languages in multilingual speakers. This data will include results of people that do and do not claim to have this type of synesthesia. The second is an application of a synesthesia test to determine whether the associations are stable in participants – to show validity of the data. The third is an analysis of the data across several modes of data collection to see whether there are any common trends among speakers of different languages. The fourth is an analysis of the most common pairings/groupings to see whether any of the aforementioned possible factors really did affect the NS associations.

2. Literature review

Existing research on synesthesia outlines several ideas that are important for this work. One is that synesthesia is a stable phenomenon that could manifest at any point of a person's life. There is also an idea that synesthetic experiences could be present in all people to a different extent. These ideas are reinforced by the research focused specifically on grapheme-color and weekday-color synesthesia. Additionally research presents two main reasons for the creation of particular pairings: one says that the associations come from childhood experiences which applies well to this research as the semantic groups under investigation are some of the first ones to be acquired; the other says that the words and concepts are paired based on their

frequency of perception – the more frequent items pair up with other frequent items. The study of associations presents us with findings that largely agree with those present in synesthetic research. Namely, it has found that the items that are named in association pairs are usually words of high frequency or ones that were acquired early on. Another consideration is whether the associations are created based on similarity of sound form. Due to the existence of the kiki/bouba effect (Nielsen & Rendall, 2011), we can assume that some association can be created through a median cognitive representation of how the words sound.

2.1 Synesthesia

Past research on synesthesia looked into many aspects and types of the phenomenon. The number of the types of the possible synesthetic associations is evaluated at most to be 150 types (Cytowic & Eagleman, 2009). One aspect of synesthesia that differentiates it from other types of associations is that it is stable. The pairs that a participant reports are the same across many trials meaning that it is not merely a made up pairing that could be forgotten but rather a real perceptual experience. The efforts to prove so started in the 80's with Baron-Cohen et al. (1987) that tested and retested synesthetes on their association between words and synesthetically induced colors. However, that method has been since substituted with more experimental procedures. One such experiment is an eye-tracking experiment. Examples of such could be found in works by Paulsen & Laeng (2006) and Carriere et al. (2008). Both of them research Grapheme-color synesthesia but the application of the technology was different. In both cases, the participants were first asked for their association pairs and then presented with colored letters that either agreed or disagreed with their reported pairs. Paulsen & Laeng (2006) showed that the pupils of the participants increased in size when seeing a pair that did not agree with the pairs reported by the participants. Carriere et al. (2008) showed that the participants focused more on the stimuli that agreed with their reported pairs when presented two stimuli at the same time - a pair that agreed and the pair that disagreed.

Another aspect of synesthesia is when it is acquired. One study that investigated this topic is Simner & Bain (2013). They investigated the development of synesthesia in children in a longitudinal study on grapheme-color synesthesia – a type of synesthesia that produces an additional sensation of seeing colors when letters and digits are the real stimuli. The children in the study were presented with a simple questionnaire that asked them to assign colors to letters and digits. The participants retook the test again four years later. An important finding of this study was that both times that the questionnaire was performed new participants showed a presence of synesthesia meaning that it could appear both in early childhood and later on in life.

A research by Maurer et al. (2013) reviews the 'neonatal synesthesia hypothesis' which states that all people are synesthetes at birth and then lose these connections as they grow. Kids show hyperconnectivity between senses which is most likely a sign of natural associations between senses as the exhibited connections cannot be explained by learning. Although this might be one of the origins of synesthesia, the analysis of the current research will not include a discussion on it. The reason is that the items associated in NS are mostly learned concepts and are unavailable to infants and toddlers. Thus, the hypothesis on synesthetic associations being from a natural connection between senses cannot be applied to this research.

One of the secondary research questions of this work also touches upon a topic of whether people that claim to not have synesthesia can also have synesthetic experiences. A study by Martino & Marks (Martino & Marks, 2001; Marks, 2013) introduces as to such phenomena as weak and strong synesthesia. This classification is easy to grasp. Strong synesthesia is what we would call synesthesia in general – a vivid internal sensory experience as a response to another unrelated sensory experience. Weak synesthesia, on the other hand, introduces us to correspondence of sensory stimuli that we might know from synesthetic metaphors. Such things as “sharp smell” and “loud color” are not random correspondences of

stimuli. The authors are assured that there are some shared experiences people have that lead to similar cross-sensory sensations that get expressed best through language. An existence of such correspondences and weak synesthesia leads me to believe that non-synesthetes might too experience some association pairs if prompted to. This work also indirectly agrees with the aforementioned study by Maurer et al. (2013) that at least to some degree all or most people are synesthetes.

2.2 Factors affecting synesthesia

Research on the Grapheme-Color synesthesia has been very productive in the investigation of the reason behind created letter-color pairings. Several works (Simner et al., 2005; Association for Psychological Science, 2008) agree that what could explain the pairings created by this type of synesthesia is the frequency of the items or their prototypicality. Both research projects find that the more frequent letters are paired with the more frequent colors and the other way around – the less frequent letters are paired with the less frequent colors. Association for Psychological Science (2008) states that the association is based on prototypicality of the items rather than on frequency of use. That study states that across participants we can find a consistency in pairings. This way, ‘a’ is associated with the red color, and ‘v’ is associated with purple. Simner et al. (2005) provides such observations that ‘a’ is paired with red, ‘b’ is paired with blue, ‘c’ is paired with yellow – all of which are frequent items in their semantic groups. At the same time, letter ‘w’ is paired with the orange color – both of which are less frequent than the aforementioned counterparts. Two points need to be mentioned here. One is that the authors themselves note that they cannot comment on why the pairings are exactly such – i.e. why ‘b’ cannot be associated with the red color instead of ‘a’. The other is that the study by Simner et al. (2005) was conducted on both synesthetes and non-synesthetes which agrees with the Martino & Marks (2001) study.

Another proposed reason for the pairings comes from the study on the same type of synesthesia by Witthoft & Winawer (2013) who analyzed different people for consistencies and found that many of the pairings they reported agree with color-letter pairs used in a childhood alphabet toy they had. This shows that the associations could be sourced from childhood experiences and toys. As the words that seem to be present in NS come from an early acquired vocabulary, it might be reasonable to hypothesize that these associations also come from some early experience like a children's book. If participants report such an information that they are sure that their association comes from a particular childhood experience, this would agree with this finding. A similar set findings are presented in a research by Hancock (2013).

Hancock (2013) also discusses a topic present in research by Beeli et al. (2007) and Smilek et al. (2007). These works show that there is a trend in how particular items get associated with more or less luminescent colors. One thing they discuss is that letters and digits that are learned earlier – letters A and E, and digits 1, 2, 3 – are associated with the more luminescent colors like white. In this way, the items that are acquired later on – letters Q and P, and digits 8 and 9 – are associated with the less luminescent colors like blue and purple. Another finding that the works get into is that there is a trend of how the shape of the letter also affects the color it associates to. This way, the letters with angles like X and Z associate with the darker colors, while I and O – the less angular letters – are paired with lighter colors. Overall, the works conclude that there is a variety of explanations that could stand behind the grapheme-color pairings created by synesthetes and there is no definite answer which one is the more influential one or how they interact among each other.

One study that disagrees with the effect of childhood experiences is by Ramachandan & Hubbard (2001) that also focuses on Grapheme-Color synesthesia but does so through experimental methods rather than by questionnaire that paired letters with colors. Their main findings were that induced colors led to a pop-out of corresponding graphemes; a digit that

seems invisible due to crowding can still lead to an experience of color; and that graphemes did not lead to a perception of color only when they were present in a peripheral zone. These prove that this type of synesthesia is a perceptual phenomenon rather than one based on association from childhood, experiences, or metaphors which disagrees with the aforementioned articles. I do not fully agree with this conclusion as I feel like the phenomenon being a perceptual one does not negate that the neural connection between these two semantic groups or sensations could have been created by the experiences that happened early on – in the moment of acquisition.

The study (Ramachandran, V. S., & Hubbard, E. M., 2001) also differentiated two types of synesthesia: higher and lower synesthesia. The lower synesthesia is established when the association is created due to low-level perceptual features of the inducer – the stimuli that is present in reality. An example of such a feature is how some letters and digits are written in a way that is more curvy – as in 0 or 3 – or more angular – as in 4 or 7. The higher synesthesia is established when the association bases on higher level of perception – the conceptual one. An example of such would be a connection between 4 and Roman numeral IV or ∴ sign. The question of whether the association happens on the level of words or concepts. The similarity in shape of the letters that constitute the words – or perhaps only the first few letters of the word – may be a reason for an association. A consideration of this will be present in the later analysis.

2.3 Weekday-Color synesthesia

General discussions on synesthesia aside, there are some other works that discuss topics close to the synesthesia types under investigation in this work. A study by Rouw et al. (2014) researched color associations for days of the week and letters across three languages. The important findings they had is that there are consistencies in color preferences for both days of the week and letters that go across languages. They also were able to find the correspondences in association pairs in both participants that report having synesthesia and those that do not. As

for their analysis, they focused on the possible reasons for the color preferences. Some of the mechanisms influencing the color decision are the linguistic-specific effect, sequence order, and figurative speech.

Another work that discusses the connection between the days of the week and colors is a book by Cytowic and Eagleman (2011) called *Wednesday is Indigo Blue*. It states that this combination is one of the more common types of synesthesia. The authors say that these connections come from a difference in neural networks across people that get created as people go through their lives and get influenced by a range of factors. This at the same time agrees and disagrees with findings of Ramachandan & Hubbard (2001) in that they agree that the phenomenon is a perceptual one but also reference experiences of people as the reason as to why the connections come to exist. The importance of this work has to do with the creation of the Synesthesia Battery webpage https://synesthete.ircn.jp/home which allows for a standardized test on whether a person is a synesthete with an additional verifying component. An adaptation of this battery test is going to be utilized for this work's methodology.

2.4 Associations

Another aspect that can be discussed for this work is the studies on associations. If the synesthetic connections come from personal experience, it can be anticipated that findings in the research in associations in general can be applied to the synesthetic associations as well. One such work is by De Deyne & Storms (2008) that analyzed associations for possible trends. One important finding is that associations created associative networks which established nodes - items in the networks that connected many chains together. The trend for the nodes is that they were nouns that were highly frequent and acquired at an early age. This finding corresponds well to the expected semantic groups that will be present in associations in NS. If it turns out that the more frequent items get associated in synesthetes more than others, then the findings would agree with this study.

The associations are also used to determine the time of the creation of concepts expressed by words. This way, a study by Lowie et al. (2010) looked into concept creation and through a priming experiment in associations in English and Dutch showed that even once achieving fluency in L2, speakers' associations are ruled by the concepts created in their L1 culture and society. This way, it is expected that the multilingual participants of this study will have the same associations in their other languages as they did in their first one as the semantic groups under investigation name concepts rather than physical objects. If it turns out that the NS associations are different in different languages of the speaker, one reasoning for that could be that the association is created on the level of the wordform, rather than on the level of the concept it names.

Another work on the interaction between associations and order of language acquisition is by Fitzpatrick & Izura (2011). They analyzed response time and the type of cue-response correspondence. One of the findings is that when the cue and the response are associated in at least two ways - out of meaning similarity, form similarity, and collocation frequency - the response time is faster in both L1 and L2. The associations in L2 are produced with slower time, however, the response is faster if the response cue is the same in L1 and L2. This shows at least some L1 mediation in producing L2 associations. Our associations could also be connected through their similarity in sound form or collocation with a third word.

2.5 Similarity of Soundform

Another consideration for the analysis of the resulting association pairs/groups is their similarity in sound. One such aspect is the simple similarity in sounds that constitute the words. This way, an association between 'white' and 'Wednesday' may be based on the fact of them beginning with similar sounds. In Russian, the same can be said about something like 'понедельник' and 'прямоугольник'. The length of the words may also affect the likelihood of words being paired up.

Another possible reasoning is the interaction between how we perceive sounds and shapes. An article by Svantesson (2017) talks about exactly that. They call it sound symbolism - a phenomenon when there is some similarity between the item being named and the word used to name it. Their work compiles past findings to show the validity of sound symbolism with how particular sounds are present in words with particular meanings across languages of different language families. The work outlines such cases of sound symbolism wherein particular sounds correspond to the size, shape, and speed of the referred object, action, or description. This way, one example is from the tonal languages of West Africa – Ewe, Twi, and Nupe. In their case, the words for such meanings such as 'small', 'quick', 'bright', and 'light' contain high and front vowels and pronounced in a high tone. The meanings similar to 'large', 'slow', and 'dark' contain back vowels and are pronounced with a low tone. This same trend was found across many languages. In other languages – such as Bohnar, Kammu, and Thongkum – the association between the sounds and the meanings is also stable, however, it is the exact opposite. There, the back vowels are present in nouns, verbs, and adjectives that contain some 'small, ' quick', 'light' meanings; and the front vowels correspond to the 'dark', 'large', 'slow' meanings. The work also discusses the takete/maluma effect present in experimental studies and which shows that people tend to reliably associate the shape of the object being named with particular phonemes. This work introduces us to an idea that particular phonemes are naturally associated with specific meanings – in our case, colors or shapes. It is possible that the phonemes of the words in NS groups might have some similarity. This also calls for a discussion of the 'kiki'/'bouba' effect (Uznadze, 1923). If there exists a stable association between sounds and shapes that they may represent this discussion would echo that of Ramachandan & Hubbard (2001) as it would reference lower synesthesia. The sound of the word or a letter might invoke an image of a shape; a shape that can be comparable with the shape of the individual graphemes (curvy or angular) or a cumulative shape of letters in a word.

2.6 Frequency, Commonality, and Prototypicality

Throughout this work, I use terms such as prototypicality, commonality, and frequency of items. The frequency is the number of times a word appears in a chosen corpus. For the sake of the statistical analysis, the frequency is substituted for relative frequency, where the absolute frequency of the word was divided by the sum of frequencies of all words in a semantic group. The words for the semantic groups were decided based on the words that are taught to children as 'basic'. This way, for the weekday, months, and digits, the members of the semantic groups are stable and understandable. For the group of shapes and colors, I relied on the data that was present in the corpus created in the first methodology, and my personal experience as to what shapes and colors were taught to me the earliest. In this study, a part of the possible explanation for the association has to do with the early or prolonged exposure to the words. The less common items – for example color 'turquoise' – have a much lower chance of appearing in these associations as the first exposure to them happened much later than to the words naming days of the week and digits. The data from the existing corpora used to determine the relative frequency is not rounded, however, the calculated relative frequency is rounded to the first four decimal digits. Commonality is just a convenient substitute for frequency wherein the words are given a category instead of a numerical value to represent how often they appear in corpus. I use the three words almost interchangably, with the items that have the most frequency also being the most common ones and the most prototypical ones. I base my decision to do so on the definition for prototypicality from Rosch (1975) that is also called a statistical model by Geeraerts (2007). By that definition, a prototypical member of a category is the one that is encountered most frequently. This disagrees with some of the studies that divorce the results of a corpus frequency from prototypicality judgement (Boersma, 2006) and Geeraerts does an overview outlining many other models of prototypicality such as physiological, referential, statistical, psychological, and family resemblance models. However, the statistical model is the

one that is the least subjective one and fits with the methodology adapted from Simner et al. (2005).

Aforementioned research shows that synesthetic associations under investigation could be affected by several factors, namely: previous experiences, frequency of interaction, order similarity, collocation, and similarity in sound form. These factors are going to be considered in the later analysis of the pairings that are going to be produced by the participants of this study.

3. Existing reports of Network Synesthesia

As a part of the research into associations created by NS, I decided to start with collecting data on associations that is readily available online. This data will play two main roles. One is that it is a database that we can analyze for possible trends – as to learn what we could expect from our future data. The corpus is also a baseline for comparison to English speakers (as the material was collected in English). The resulting corpus can also be instrumental in determining what semantic groups could be used on the next steps of data elicitation through questionnaire and during the experiment.

3.1 Methodology

The collection of data for the corpus of NS association groups started with a manual search through Twitter. Based on my personal experience and posts that I have previously seen, I decided on some semantic groups that may appear in posts that would report NS experience. I started by inputting expected keywords into Twitter (or X) search engine. After collecting about 20 posts, I determined the semantic groups that appear in such posts most often. Most common semantic groups turned out to be days of the week, colors, months, digits, and letters. The general items of these groups were added to a list together with typical structures that such posts utilized (such as “X and Y feel the same”, “X and Y have the same vibe”, “X is similar to Y”).

The second step is an automatic collection of data with an adaptation of existing algorithms to parse the posts on Twitter based on the list of keywords mentioned above. Such works as Purohit et al. (2015), Llewellyn et al. (2015) provide us with an algorithm of how data could be extracted from Twitter. The extracted data will be cleaned from noise to only include posts that reference a phenomenon similar or equivalent to NS. As a result, we are left with a list of combinations of items from semantic groups that contained from 2 to 6 items – all from different groups. This data can be said to be "naturally occurring" as its report happened outside of an experiment. The examples of the tweets that were used for this corpus can be seen below in figures 1-3

Figure 1

Example of a post on X.com that shows NS

Figure 2

Example of a post on X.com that shows NS

Figure 3

Example of a post on X.com that shows NS

The resulting combinations will be analyzed based on several factors. We will see what items will appear more frequently and in what combinations. Additionally, we will perform a statistical analysis on the pairs to see whether the prototypicality of the items affects the chance of items being grouped up. Whether the item is prototypical or not will be judged based on the judgement of existing studies on whether a letter, a digit or a color is a prototypical/common one. The prototypicality of the days of the week, months, and shapes will be based on their presence in available online corpora of languages necessary (for English, the data is gathered from British National Corpus and from English Web 2021). The values for the letters and colors are borrowed from the information provided by Simner et al. (2005) – their calculation base on the data from British National Corpus. The amount of appearances of each item in a particular corpus will then be calculated against the values for the other items in one semantic category. This way, we determine relative frequency of each item with a value from 0 to 1. Next, depending on the third in which the value of the relative frequency lays, we determine its relative frequency on a categorical range – common, middle values, and uncommon. The idea is that the possibility of the analysis utilizing categorical values will show a better representation of the frequencies. That is because some of the items have a much higher frequency than other items in the same semantic field. This way a word “one” is much more common than any other word that represents a number. While its relative frequency makes it hard to compare the other

values – as it also makes the relative frequencies of the other items clustered together on the lower end – giving it a “common” category together with words for “two” and “three” makes it easier to operate.

The data was analyzed in several ways. First, I did a qualitative analysis for specific cases – such as reports by multilinguals – and those that contained interesting but rare items – such as mathematical equations. Due to my limited abilities, I am unsure how I could include singular cases into a statistical analysis, however, they are too interesting to overlook. Thus, it calls for a separate qualitative overview. There, I looked into whether any of the paired items have similar sound form or if they are a collocation – as these are the possible reasons for the creation of an NS association.

Next, I did a statistical analysis. For the sake of the analysis, all of the items in an association group will be analyzed in pairs. This way, if a group has 3 members, it will result in three pairings. A group with 4 members will result in 6 pairs, and so on. The items in the pairs will be named into ‘first’ and ‘second’ in the corpus for the statistical analysis. The order of the items – which item was named the first and which the second – is random. This work will utilize three statistical models – linear regression model, multinomial model and conditional tree. For these, either the relative frequency or the categorical values were the analyzed material. All of the models show whether the value of the relative frequency of the first item could be predicted by the values of the second item. For the sake of the linear regression model, the values of the frequencies were also transformed logarithmically to achieve a distribution closer to a normal one – as required by the model itself.

3.2 Results

3.2.1 Frequency table

Here, you can find the table of frequencies of the semantic groups under investigation. As mentioned above, it displays the relative frequency of the words, and their commonality

category. ‘Com’ stands for ‘common’, ‘mid’ stands for ‘middle values’, and ‘not’ stands for ‘not common’. The words of a category are ordered from most frequent to least frequent. The whole table of frequencies with data for Kazakh and Russian can be found in the repository in the Appendix section.

Table 1

Frequency and commonality category of words under investigation in English

Word	Frequency	Category	Word	Frequency	Category	Word	Frequency	Category
Months			Colors			Digits		
June	0.1120	Com	White	0.2180	Com	One	0.4671	Com
April	0.1109	Com	Black	0.2221	Com	Two	0.2306	Com
March	0.1085	Com	Red	0.1485	Com	Three	0.1100	Com
July	0.0897	Com	Green	0.1346	Com	Four	0.0595	Mid
October	0.0821	Mid	Blue	0.1123	Mid	Five	0.0476	Mid
September	0.0804	Mid	Yellow	0.0550	Mid	Six	0.0329	Mid
January	0.0781	Mid	Brown	0.0409	Mid	Seven	0.0194	Not
November	0.0725	Mid	Gray	0.0207	Not	Eight	0.0168	Not
December	0.0722	Mid	Orange	0.0205	Not	Nine	0.0114	Not
February	0.0644	Not	Purple	0.0185	Not	Zero	0.0057	Not
May	0.0673	Not	Cyan	0.0100	Not			
August	0.0619	Not						

Days of the week			Shapes		
Sunday	0.2318	Com	Circle	0.3464	Com
Saturday	0.2078	Com	Star	0.3338	Com
Friday	0.1375	Mid	Square	0.1715	Com
Monday	0.1319	Mid	Triangle	0.0841	Not
Wednesday	0.1094	Not	Rectangle	0.0393	Not
Thursday	1.0933	Not	Oval	0.0234	Not
Tuesday	0.0884	Not	Rhombus	0.0015	Not

3.2.2 Qualitative analysis

Although the main analysis method of this work is quantitative, at this point, I would like to highlight some of the findings that would not be represented in a statistical analysis.

One such finding is that many of the reported pairings could be explained by the items having the same position in an ordered list within their semantic groups. Months, names of the days of the week, digits, and letters all have an order that people recognize. Colors to some extent also have an order as they could be named in the order of the rainbow. This way, from what people have reported, the items being in the same position in their ordered lists affects

their likelihood to be associated with one another. An example of such could be an association group that consists of ‘wednesday’, ‘yellow’, ‘3’, and ‘c’ - all of the items take the third position in their respective groups. The same goes for ‘monday’, ‘red’, and ‘a’. Although it is fairly easy to see why these words would be grouped together, people still report not understanding why they make such connections. Such cases constitute around 10% of reported associations.

Another finding comes from the cases where people reported their association groups in more than one language. In all such cases, the reported groups differed across languages. For example, one person reports that digit ‘8’ is associated with color ‘yellow’ in English, but with ‘purple’ in Portuguese. Looking at the differing items, we can find a little tendency: the names for colors and digits contain similar phonemes. This way, a native speaker of Spanish reports that while in Spanish they associate *roho* ‘red’ with *quatro* ‘four’, in English, ‘four’ is associated with ‘brown’ rather than with ‘red’. Their initial association of 4-red does not transfer to English directly. Instead, ‘red’ gets submstituted to ‘brown’ which is the basic color term that is the closest in sound to ‘four’. Additionally, in Spanish, ‘brown’ is *marron* which is also close to *quatro*. Maybe, in an absence of a term for ‘red’ that sounds similar to ‘four’, the speaker ends up going for the color that is the second association to ‘four’ in Spanish. However, this is purely speculative. Such cases were few and reported pairs in different languages - Spanish, French, Portuguese - so it would be impossible to include them into the statistical analysis.

Another type of cases that cannot be used for statistical analysis due to a lower number of instances is one that associated the main semantic groups under consideration with things like shapes, time, and equations. This way, people reported such association groups as thursday-purple-november-7*8=56 and thursday-october-8PM.

3.2.3 Quantitative analysis

First, I would like to start with the general trends of the results. The most reported associations came from two types. The first type is the one that contained items with the same position in ordered lists. The second type is the associations between items that are met less frequently among the members of their semantic groups. This way, one of the most reported associations is between Thursday - an uncommon day of the week - November - an uncommon month - and either purple or yellow - an uncommon color. Aside from being often reported together, these items are overall the most frequently associated items.

Figure 4

Mosaic plot of the interaction of categorical frequency

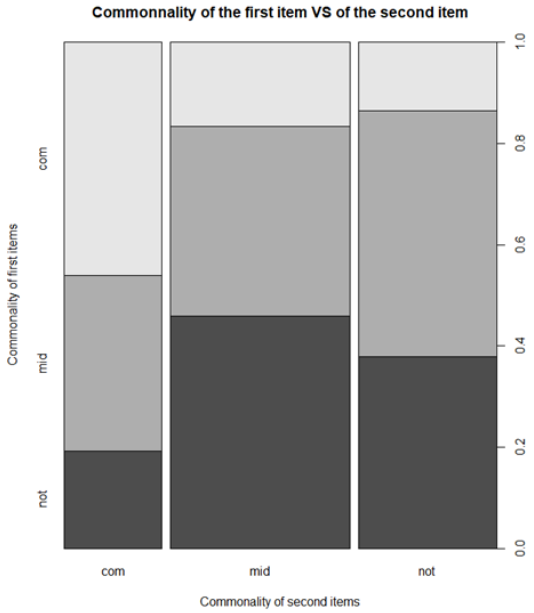


Figure 4 above shows the general trends of how the pairs were made up based on the frequency of the constituent words. We can see that there is a noticeable difference in the amount of interaction of the frequency categories. The common items pair up most often with other frequent items and rarely with items of low frequency. The non-common items and the words of the median category pair equally often with other items of these two categories and rarely pair with items of high frequency.

An additional statistical analysis using multinomial and conditional tree models show that the factor of the frequency of words is a significant predictor of whether words are associated with each other. Figure 5 and 6 show the extent to which the frequency is a good predictor.

Figure 5

Multinomial model coefficients for frequency as predictor of association

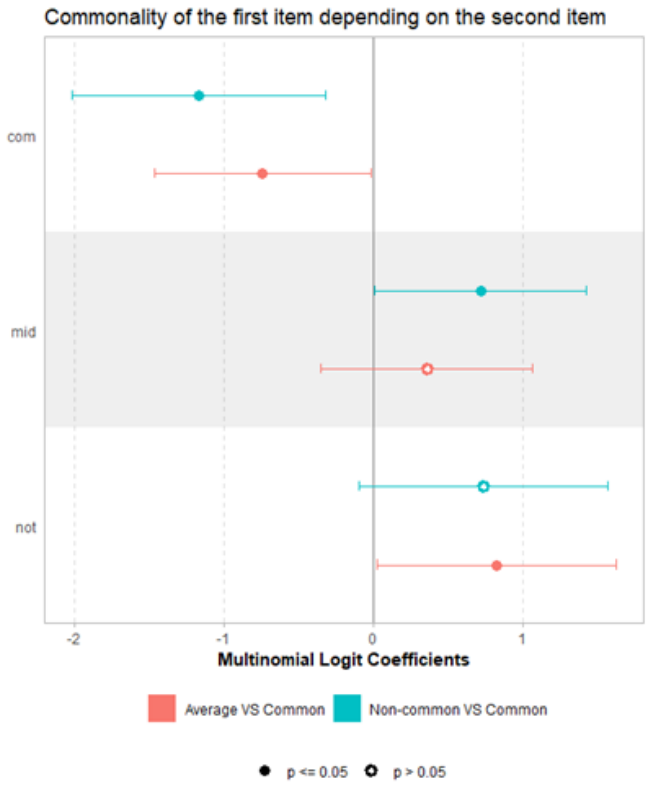
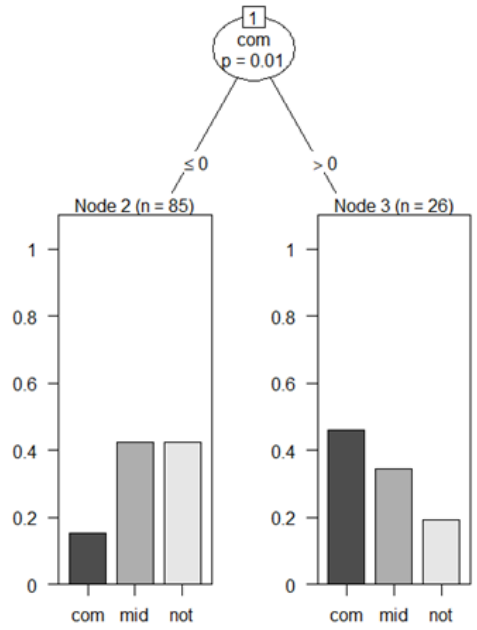


Figure 5 is a graphical representation of how likely it is that the second word will be of a category other than common depending on the frequency category of the first word. We can see that for 4 out of 6 interactions the difference is significant (with a p-value less than 0.05). This would make the frequency a significant predictor, however, as visible from the image, the error bars are quite wide. This means that in general, the model might predict the category of the second word depending on the frequency of the first word, however, its accuracy will not be great. This agrees with the calculated Wald statistics result.

Figure 6

Conditional tree model for frequency as predictor of association



The conditional tree model also shows that the first word being a common one is a significant predictor (with a p-value of 0.01) of the category of the second word. If the word is common it is significantly more likely that the second word is also from the common category.

3.3 Discussion

The main finding of this section is that there is a significant correlation between the frequency of items and their likelihood to be paired into an association. While there are significant errors in the models based on our data, the simple explanation for that is the rather low amount of data. The corpus was built on data that comes from social media. The problem with this is that some posts become viral or come from an account with a big following. Both of these cases lead to many reactions that could also be counted for the corpus. However, these responses are 'primed' to respond with either the same combination of words or with a combination that somewhat agrees with the one in the initial tweet. This results in a situation where some

combinations get overreported. As I did not want to skew the data into some combinations being reported hundreds of times while others - coming from a smaller account and not getting engagement - are only present a couple of times. I think that the data that considers all of the tweets - including those created as a response or a reaction bait - would not be representative of real life distribution of the NS experiences. Because of that, for the statistical analysis, I had to consider only the posts that were created on their own - not as a reaction to another post. This made the number of tokens of data go down significantly. As you can see from the image of the conditional tree, the final analysis only considered the data of 111 tokens. This is a rather small number considering that it distributes over 3 categories of frequency. While I am not sure that a larger amount of data would result in a better model, this is still something that should be considered as a limitation of this methodology stage.

Despite the aforementioned problem, the models still show that the frequency of one word is a significant predictor of the frequency of the second word. Although there are considerable errors, it is still a result that cannot be ignored, as even a mosaic plot shows a preference of words to match for frequency.

Additionally, what we were able to achieve through this step is that we determined the main semantic groups that could be used for the experimental part of the methodology. We now have material that came to be recorded more or less naturally and we will be able to see whether the prompting methodology is an appropriate tool. We also understood some of the possible reasons for association. These include a similarity in sound and similarity in some aspect of meaning. However, as it seems from the data, the similarity has to be clearly apparent for it to result in an association. For example, one similarity type is based on the position of the word in an ordered list with the other members of its semantic group. This way, a lot of people associate ‘monday’ with ‘red’ and ‘a’ as they stand in the first position in their respective lists. At the same time, color ‘yellow’ that takes on the third position is never paired with

‘wednesday’ or ‘c’. It is inferred that for people it is easy to track the first and the last members of the numbered list but not the members in between. The results of this section will be compared to the results of the following sections in the general discussion.

4. Collection of associations in English, Russian and Kazakh

The main methodology of this study targets collecting association groups in English, Kazakh and Russian. The task is completed through a questionnaire and it is necessary to achieve two things: collect as many association pairs/groups as possible in English and to collect such information for multilinguals. By achieving this, I would be able to find out whether synesthetes of this type have consistent associations across languages and whether the trends in association differ in different languages.

4.1 Methodology

The questionnaire for this step of the study consists of three main parts. First part collects general information on the participants such as their gender, their language proficiencies, their level of creativity, neurodiversity status, and, most importantly, whether they are aware that they have associations similar to what can be covered by NS. Their language proficiency and the report on synesthetic experiences decides what is shown to them in the questionnaire. If a person reports that they have NS experiences, after general information section they are provided with an open question where they are allowed to report all of the associations that they have in any of the languages. After that, they - and all of the participants that did not claim to have NS - are shown a prompting phase.

Figure 7

Example of the prompting phase

The prompting phase is necessary to collect the pairings that the participants have. The figure 7 above shows an example of how the words were displayed for the prompting phase. The prompt matrices were shown to the participants at random. The words of two semantic categories are displayed in the first row and column of the matrix. The order of the words of one group is random and does not correpond to their usual positions in an ordered list. This way, in the case of the semantic group of the days of the week, the items were not shown to the participants in their regular order from Monday to Sunday. Their positions were randomized in a way so that all of the items are in a place different from their usual order and consecutive items were not displayed together. None of the answers included cases where participants chose the pairings in some particular pattern on the matrix – for example choosing pairs to create a diagonal line. If such cases were present, they would have been excluded from the analysis. The rest of the questionnaire can be found in the Appendix section through a link to a questionnaire preview on Qualtrics. The participants were asked to mark the intersection of two words that they could feel some association for. The participants were free to mark however many pairings they wanted and could skip a prompt matrix all together if they did not feel any associations in it.

Initially, there were two groups of participants, however, I realized that the transfer of data from the Qualtrics data to a format that was understandable to R was too slow. Because of

that, I decided to utilize only the data of one of the groups – students of NU. They provided me with data in three languages – Kazakh, Russian, and English, with most of the participants providing data in more than one language. This data was more fitting for the goals of my research. The second group consisted of foreign respondents recruited through Twitter (X.com) and Reddit.com. Some of the material that I had time to transfer gave me an insight into the difference in associations between different groups.

The participants were first shown the prompting phase in their first language – either Kazakh or Russian – and then, upon their own will, they could also go through the prompting phase in their second language (Russian or Kazakh) or in English. Overall, each prompting phase consisted of 10 matrices showing all of the possible combinations between the 5 semantic groups under investigation: colours, digits, days of the week, months, and shapes. Previously I also wanted to include a discussion on their interaction with letters, however, such a consideration would only lengthen the questionnaire and did not have a compact prompting method. After the prompting phase the participants were free to report any of the groups and pairings that had members of the semantic groups not present in the matrix.

Through this method, I collected synesthetic association pairs and they were analyzed for the following aspects. First, I looked into whether there is a clear reason as to why a pairing was chosen. These reasons include similarity in sound or some similarity in meaning. An example of the first one is kazakh *qara* ‘black’ and *qarasha* ‘november’ which start with the same two syllables and have a rather opaque semantic connection. Another example comes from the interaction of shapes and digits where there is a clear connection between digits and the number of angles of a shape. In both such cases, the created association is explainable and will not be considered in the following analysis for frequency. I looked into whether most of the associations are created in this way.

Next, I looked into whether the frequency category of the paired words explains their association. For this, I looked only into pairs that do not seem to have an explanation for their association. For this and the future parts of the study I gathered frequency information for each member of the investigated semantic groups. The information regarding the words in Russian was collected from the Russian National Corpus; for words in Kazakh, their frequency was collected from the Almaty Corpus of Kazakh. The collected absolute frequencies were turned into relative frequencies among the members of one semantic group and then each word was assigned a frequency category of either 'common', 'median' or 'not common' depending on their relative frequency. Here, I used linear regression, multinomial and conditional tree models to see whether the factor of frequency is significant in the creation of pairs.

Another aspect that was considered in the analysis of the responses is the association category of the reported pairings. What I mean by association category is a pairing of the semantic groups the words in the association pair come from. This way, some of the association groups are week-color, week-month, color-digit and so on. I will be referring to these aspect of the data as association category for the rest of the thesis.

I also looked into whether multilingual participants report consistent association pairs across their languages. The outcome of this could show us whether the association happens on the level of words or concepts. Relying on findings on associations by Lowie et al. (2010) and Fitzpatrick & Izura (2011), for this work, I assume that words in different languages that name the same thing are tied to a single conceptualization. Therefore if the reported pairs are different across languages, it could mean that the association does not happen between the named concepts.

Another aspect of the analysis considers whether there are trends as to what pairings are reported across participants.

4.2 Results

4.2.1 Self-reported associations

In the beginning of the questionnaire, the participants that marked that they have experienced synesthetic associations had a chance to report their own pairings with no guidence of the prompting matrices. I had 30 such participants and most of them reported only a few pairings. I understand such a small amount of reported pairings as it is hard to come up with examples on a spot. There are several things that are interesting about the pairings that were produced from self-report. First is that most of the groupings included other semantic groups that were not present in the prompting phase. The groups include mathematical equations – of the type 5*2=10 and 7*8=56; time – 17:00, 4PM; seasons; and school subjects. I guess that the report includes such thing as school subjects at such a high rate – 17 out of 30 participants – because the participants are current students and the school subjects are a recent memory. From the material of the foreign participants that, while they also report associations with time and equations, there are only two responses that mention school subjects. The foreign participant pool is much more diverse in age, so maybe there is a recency bias that leads to the university students reporting school subject pairings much more frequently.

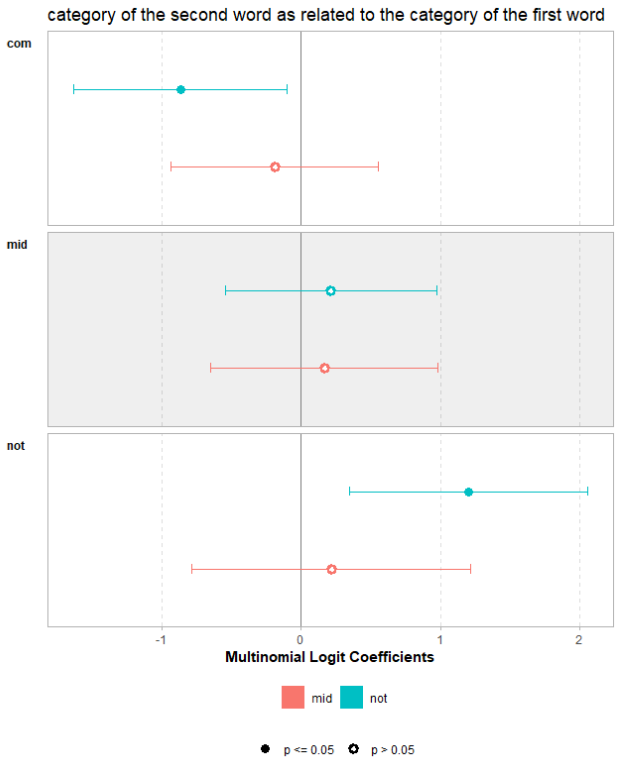
Another interesting thing that we cannot learn from prompting is about the association groups. In the reports, the participants provided groups of up to 7 members, and what we can find there is that often there is one member of the group that does not seem to fit as well. This way, in a group of yellow-triangle-seven-Thursday-July-L, there is a clear trend that all members besides July are on the lower end of prototypicality. What I see here is that ‘seven’ that has a connection to other low frequency words brought July in as the connection between july and seven is quite strong. The same can be said about Thursday-purple-nine-november-april. First, it is strange that there are two months present in the group. Most times, people do not repeat semantic groups in their associations. Second, again, there is a clear trend for lower

frequency items with Thursday-purple-nine-november. My guess is that the strong connection between the fourth items – Thursday and April – forced April into a grouping where it does not belong by frequency.

A multinomial analysis of the data – as well as a conditional tree model – shows that there is some significant correlation in the reported pairs. As we can see from figure 8 below, there is still a statistically significant behaviour in that the common words do not pair with the uncommon ones.

Figure 8

Multinomial analysis of self reported data



4.2.2 General trends in responses

Some participants – 10 to be exact – commented that they do not understand how these unrelated words could be paired together in any way. The only matrices that they filled were the ones associating digits to shapes and digits to months. For them, there was a clear connection

between the number of angles of a shape and a digit, and the ordered position of a month and a digit.

For synesthetes an important point to mention is that in 93% of the situations, the pairings mentioned in the self-reporting section of the questionnaire were also marked to be pairs in the prompting phase of the experiment.

After a general analysis of the results we can also find trends in the most common answers. Figures 9 and 10 below show the most common associations between colors and months, and colors and days of the week. The results can also be seen in table 2, for a better understanding regarding the chosen colors. The trends for these categories are the same across all three languages. Other trend that could be found in the resutls has to do with the association between shapes and digits. While the cases where the associated digit was the same as the number of angles of a shape – for example, rectangle and 4 – were the majority of the cases for this category, there were also cases that did not agree with this logic and still appeared across many participants. This way, there is a common association between a rhombus and 7, as well as a rectangle and 8.

Figure 9

Association trends between colors and months

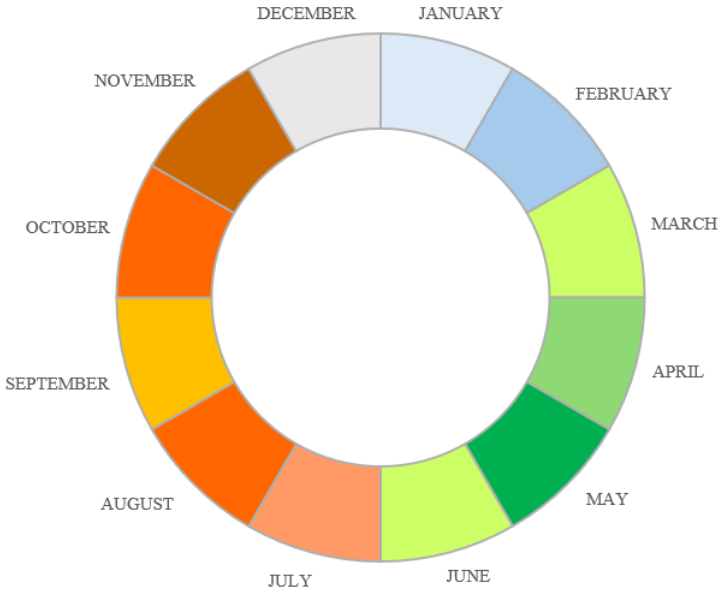


Figure 10

Association trends between colors and days of the week

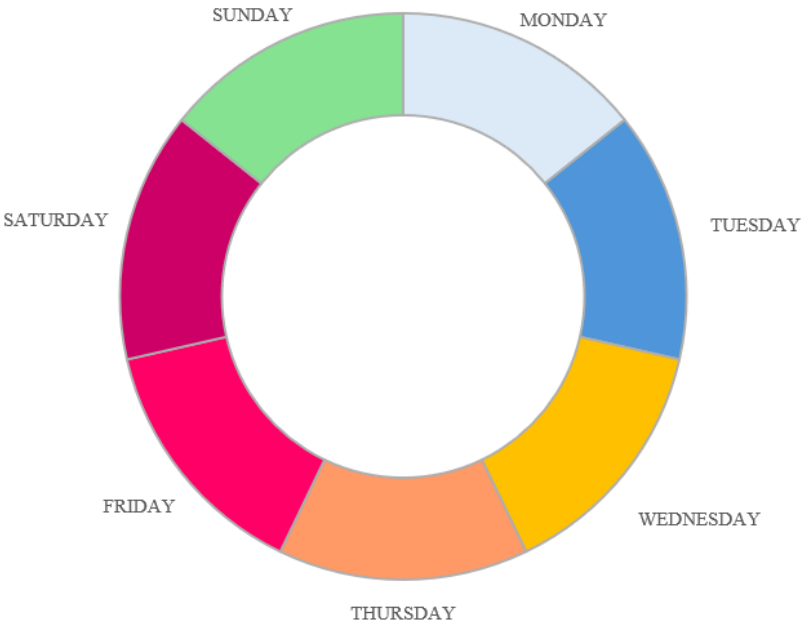


Table 2

Most frequent association pairs between colors and days of the week, colors and months

Months	Colors	Days of the week	Colors
January	White, blue, gray	Monday	White, gray,cyan
February	Blue, white, gray	Tuesday	Blue
March	Green, yellow	Wednesday	Yellow, orange
April	Green, yellow	Thursday	Orange, purple, black
May	Green	Friday	Red, purple
June	Yellow, green	Saturday	Purple, red
July	Yellow, red	Sunday	White, green
August	Red, orange		
September	Orange, yellow		
October	Orange, red, purple		
November	Black, gray, orange		
December	White, blue, gray		

4.2.3 Results in English

The association pairs reported in the questionnaire can be found in the repository linked in the appendix section of this work. Overall, out of the 1400 tokens of English association pairs, 350 pairs had a similarity in some aspect of their meaning, and 25 had a similarity in sound. The linear regression analysis basing on the frequencies of the paired words showed an insignificant p-value.

The multinomial analysis of the data present a better interpretation of the dependence of associations on the frequency of the words. Figure 11 and 12 show the graphical representation of the multinomial analysis of the data. The dots on the graph show how likely it is that the first word will be either from the median or lower frequency than of high frequency depending on the frequency category of the second word. For these graphs, the data was divided into two groups: data of the participants that claimed to have synesthesia, and those who did not.

The fully colored dots represent the collection of cases that choose a particular category with a significant accuracy. This way, we can see from figuer 11, that if the second word of the pairing is from a common category, the first word is more likely to be a common word; if the

second word is from the uncommon category, then the first word is also more likely to be an uncommon word. The dots that have a white inclusion in the center represent the collection of cases for which the p-value is more than 0.05. For figure 11, these represent only the cases in which the second word is from the median frequency category. This means that when the second word is from the median frequency, the model cannot accurately predict what will be the category of the first word. Figure 12 represents the results for the self-reported non-synesthetes and we can notice a difference in that only one of the collection of cases has a p-value that is less than 0.05. It predicts that if the second word is from the non-common category then the first word will be significantly more likely to also be from the non-common category.

Figure 11

The prediction of the frequency of the first category depending on the frequency category of the second word among synesthetes

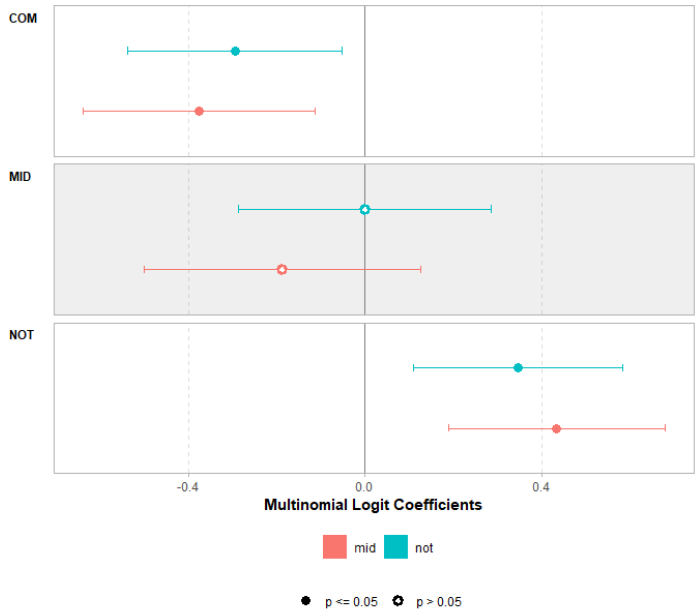
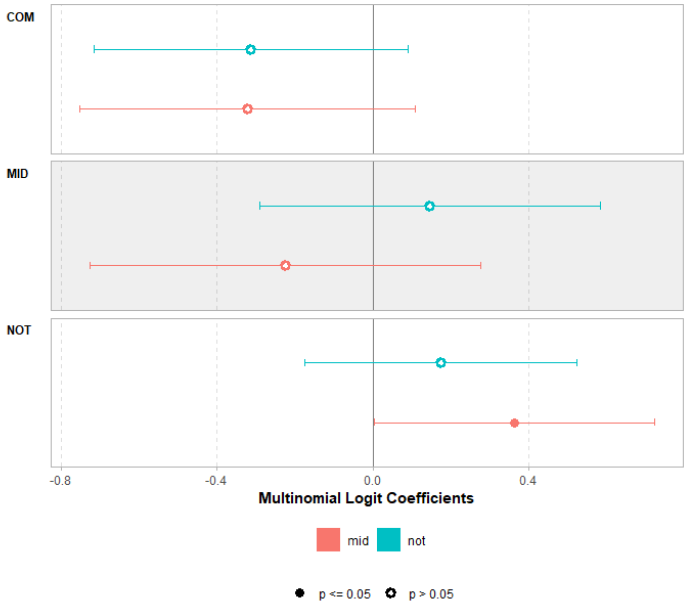


Figure 11

The prediction of the frequency of the first category depending on the frequency category of the second word among non-synesthetes



4.2.4 Results in Kazakh

Only a few participants decided to report their pairings in Kazakh, which resulted in 400 association pairs. Out of 400, 42 were pairs based on a similarity in sound, and 97 were pairs based on a similarity in meaning. For example, we had a pairing between *ақ* 'white' and *ақпан* 'february' where the similarity is both in sound – since they start from the same syllable – but also in meaning because february is one of the winter months which are heavily associated with snow – at least in Kazakhstan's climate. A statistical analysis of Kazakh shows a greater dependency on frequency than in English. A linear regression model that predicted the frequency of the first word depending on the frequency of the second word and the category of the pair shows a p-value of 0.0028 for synesthetes and 0.6937 for non-synesthetes. An analysis of the residuals of the model shows us that the model worked best at predicting association

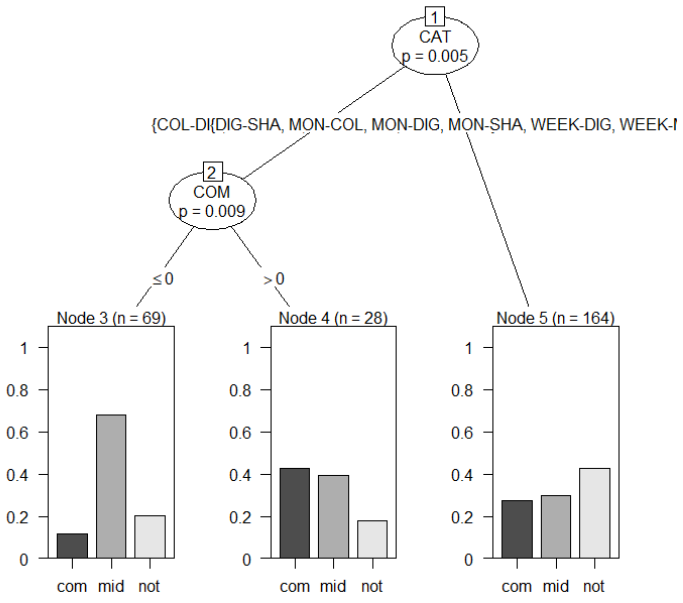
pairings between months and colors, months and shapes, and months and days of the week. The other association categories showed a greater deviation from the predicted values.

4.2.5 Results in Russian

Overall, we had 672 association pairings reported in Russian. Out of them 149 had some similarity in sound or meaning. An example of such can be the Monday-January association pair. A linear regression and a multinomial model analyses of the data did not show any significant results. A conditional tree model – showed in figure 12 below – shows some explanation of the data distribution. We see that the frequency cateogory of the first word can be predicted for only some of the categories. However, even then, we see a general trend that the not common second word results in either median commonality or non-common first word; and that the common word will more likely pair with another common word or a median frequency word than with an uncommon word.

Figure 12

Conditional tree analysis of the data on Russian predicting the commonality category of the first word depending on the commonality of the second word and the association category



4.2.6 Synesthetic associations across languages

From the data of the participants that reported their association pairs in more than one language, we can see several trends. The first one is that there was no participant who reported the exact same pairs in all three languages. Most often, the participant reported different pairs in Russian and Kazakh, however, the pairs reported in English repeated the pairs from the other two languages. If the participant did have a stable pair across all three languages, it was often also present in the self-report section of the questionnaire or was based on some very strong similarity in sound in one of the languages.

For example, a lot of participants reported black-november in a stable manner across all three languages. However, in Kazakh, black is translated to *қара* and november is translated to *қараша*. The two forms are really similar and it seems that this strength overpowers other factors leading to a connection between ‘black’ and ‘november’, and *чёрный* and *ноябрь* even through these two pairs of forms do not have a similarity in sound. This connection also overpowers the proposed frequency factor, because ‘november’ is one of the less common months and ‘black’ is one of the more frequent colors.

4.2.7 Creation of association chains

Another thing that was investigated during this analysis is on the creation of association chains. Looking at the responses of across synesthetes and non-synesthetes, we see a difference in this aspect. The associaiton pairs reported by synesthetes often can be combined into associaiton chains of up to 4 members. For non-synesthetes, we find that if word A is associated with word B from one semantic category and word C from another semantic category, B and C are usually not associated.

What is also important is that rarely all parts of the association chain are in the same frequency category. More often, one of the members will be from another category or all of the

members will be from different categories. In these cases, their pairing was caused by something else – a similarity in sound or meaning. Most of the time, the item that chains all of the words together is either a digit or the color.

4.3 Discussion

This section of the methodology uncovered many of the answers posed in the beginning of the thesis. First, we see that there are general trends present across participants and across languages. By this I mean the ones present in figures 9 and 10, but also some singular cases that are not a part of a trend across the association category. Examples of such association pairings are november and Thursday, cyan and oval, blue and seven, august and Saturday. For the associations between colors and months, we can see a similarity between the colors and the colors of the nature at the time of the month. This way, the winter months are associated with white and blue; spring is associated with green and yellow; summer is associated with green and red; and autumn is associated with brown and yellow. This agrees with findings of other articles on the topic. At the same time, we can also see a trend in the associations between colors and days of the week, however, this time, their connection is a bit harder to judge.

The associaiton chains are also repeated across participants, however, their reported number is too small to make any conclusions on whether some particular one is stable. What is interesting about them – and what was mentioned in 4.2.1 and 4.2.7 – the chains often have one member that stands out from all of the other members. That member often has a strong connection to another word in the group which drags it into association with other words.

Across languages, we see that the participants seem to transfer their association pairs from their native languages to their second ones. This way, we saw that Russian and Kazakh pairs often differed from one another, but English matched pairs with both of them. It seems, that while learning English, the association pairs present in Russian and Kazakh got transferred into English. There is also a clear difference between the soundforms of these words across

languages, so similarity in sound is not one of the factors affecting the association process in these cases. Based on this alone, it seems that the association is tied to the conceptualization of the referent of the word rather than its form. Since we receive a distinct learning that Russian *понедельник* is *Monday* in English, the whole concept of 'Monday' in Russian gets transferred into the English. This results in English pairs being a combination of the pairs in Russian and Kazakh. At the same time, the Kazakh and the Russian pairs are different because of the simultaneous nature of the way these two languages are learned in Kazakhstan. The concepts for them get created separately and do not call for a transfer of meanings between the two languages.

While synesthetes and non-synesthetes were similarly interested in filling out the prompt phase of the experiment, we see a difference in their behaviour in two ways. One is that the participants that do not claim to have synesthesia often did not see any associations between given semantic groups unless there was some direct connection. This way, they felt a connection between digits and days of the week or digits and shapes, but not between colors and any of the other semantic groups. This disagrees with the idea that all people are synesthetes but to a different degree.

Another way is that they did report some association pairings, however, looking across pairings reported by non-synesthetes in general, we see very spread out results where, for example, each participant reports a different color being paired with a word like 'june'. Based on their results alone, there is almost no trend as to which words appear in associations more frequently or which pair comes up more frequently than other pairs of the same association category. Overall the result is close to random. We can see that also from the statistical analysis of Kazakh and English where the dataset of the non-synesthetes does not result in any signifficant predicting power of its model.

As for the effect of frequency, it seems that while the frequency does play a role in the creation of the synesthetic associations, it is not the first deciding factor. I will be returning to this point at a later point, however, from the data of this methodology alone, I can still conclude this. As we saw from the results, there are two points that prove that frequency is not the deciding factor. The first one is that it did not perform equally well across all association categories that we have. The second one is the rather large number of cases that were affected by the similarity of sound or some aspect of meaning of the associated words. Across all three languages, the pairings with words similar in meaning constituted around 25% of all of the cases. While some of them cannot be considered to be cases of NS – like 'two' and 'Tuesday' – some of them cannot be overlooked. This way, while 'zero' and 'white' and 'black' are all related to absence of something, they should still be considered cases of NS because examples like this constitute cases where the person having the synesthetic experience does not understand what caused it. Another factor that is possible due to the semantic categories that we are investigating is a similarity in shape. This way 7 is associated with rhombus and 8 is associated with rectangle. There is a clear similarity between the shapes of the digits in arabic numerals and the geometric shapes.

However, the aforementioned factors can explain the origin of only some of the associations. Once we remove the pairs that could be explained through these factors, we get a set of data where the frequency of the paired words is a good significant predictor of the words being paired together.

Last point that I want to mention has to do with the effect of the order of items on the associations – or rather the saliency of the order. While it does not fully define the association behaviour, it shows an interesting trend noticable from the pairings reported in the prompts. The trend is that the first and the last items of the ordered lists of different semantic groups are regularly associated with each other while the middle members do not immediately associate

with their place on the list. This way, Monday and January get associated as often as Sunday and December. At the same time, there is almost no pairs between July and Sunday even through they are the true seventh items of the list. The same is with June and Saturday. The middle items – those beyond the second position on the list – are hard to keep track of. This is especially visible from the results of participants for whom Kazakh is their second language. In their associations, they still pair up Monday and January, and Sunday and December, however, they do not pair up the items of the same order position beyond that. The order position lose their saliency the farther down the list you go, and this seems to limit the power the order position similarity has over the tendency of the items to be paired in an association.

5. Consistency of synesthetic experiences

One of the questions that is still debated regarding synesthesia in general is whether it represents a bona fide cognitive experience or not. Some sceptics claim that the experience is fake and people report having synesthesia for the sake of receiving attention. Others say that the experience might be real, however, it depends on the situation and the received experience will be different each time. As a part of the current study, one question that is relevant for the validity of findings is whether the participants that reported to have synesthesia are true synesthetes. In this step of the methodology, it is checked whether the participants have stable synesthetic associations.

5.1 Methodology

During the questionnaire stage of participation, participants were asked to volunteer to participate in a second stage of the methodology - an experiment. One part of this experiment is to check whether the associations reported by the participants in the first stage were stable. In the experiment, the participants were shown a number of screens with two pairs of words displayed to them - one pair that was reported by them in the questionnaire, and another pair where one of the words is from the first pair. An important aspect of the stimuli is that the

substituted words come from one of the two conditions. In one case, if two words of the reported pairing already come from the same prototypicality category, then the substituted word will also be from that category. For example, if the word being substituted is ‘rectangle’ in a reported pairing ‘rectangle-wednesday’, then the new word should be of the same semantic group and of the same frequency category. In this situation the appropriate substitution would be ‘rhombus’ as both of the words are from the semantic group of shapes and both of them are in the less frequent category of the group. In the second case, the words of the reported pair come from different frequency categories. In that case, the subsittuted word will be matching the category of the remaining word to see whether the frequency of the words will sway the participant’s choice. An example of such could be ‘Thursday-red’. ‘Thursday’ is an uncommon word while ‘red’ is a common one. In this case, a substitution could be ‘Saturday-red’ as both ‘Saturday’ and ‘red’ are common.

The words on the screen are always displayed in black and in capital letters. The screens that utilize the pairings that were reported in the questionnaire are not shown consequently. Every second screen is stimuli for the part of the experiment reported in section 6.

The pointer of the participant is always placed in the middle of the screen before they are shown the pairs of words. The participants have to read the displayed text and click on the pair of words that they feel to have an association as soon as possible. The experiment is conducted using a platform called Qualtrics. After the experiment, what is recorded are the pairs that the participants chose and the time it took them to make the decision.

What is going to be analyzed is whether the participants chose the same pairings that they did during the questionnaire stage. There was a time gap between the questionnaire and the experiment so it is expected that the participants forgot the answers that they previously reported. It is assumed that if the participants reported the same pairs across the two stages it is

not because they remember the associations but rather because they actually experience a stable association between the words. Additionally, I also analyzed the amount of time that it took the participants to make their decision. It is usually reported that the synesthetic experience is instantaneous. Thus, a true synesthete should be able to decide whether a pairing does or does not have an association right after reading them. If a participant's reaction time is too long then it is assumed that they start to consciously think about what association pair works best - overruling the natural synesthetic reaction that should occur. In those instances, the answer will not be counted towards the accuracy of the participant.

What is important is that the same experiment was conducted for both: participants that report having synesthesia and those that do not. The expected outcome was that the participants that report having synesthesia will choose the same pairings as reported before with greater accuracy than the non-synesthetes.

5.2 Resutls

Both, synesthetes and non-synesthetes, that participated in this step of the experiment performed on at the same time similar but different levels. Only around 30 participants left their emails as an agreement to participate in the next step of the experiment. 24 of them contacted again to keep an equal amount of synesthetes and non-synesthetes – 12 of each. Statistically, the two groups showed a similar performance in that they got only around a half of answers the same as they were reported before. This way, the synesthete group chose the same pairings as they reported before in 62 cases out of 120, 39 times they chose the incorrect pairing, and 19 times they said that there is no difference between the pairings. The non-synesthete group chose 52 out of 120 cases correctly, 20 incorrectly, and decided to abstain from choosing one option in 48 cases.

While their success rate is similar, what is more important is in which pairings the success happened. Across the 62 correct answers of the synesthetes, 41 was with pairings that

they reported by themselves – outside of the prompting phase. Out of the 39 times they chose the incorrect pairing, only 7 cases were ones that were self-reported. Moreover, if we monitor the effect of the frequency category of the substitute word, we will see that out of the 39 failed answers, 28 were from situations where the substitute word matched the remainging word better in frequency than the inial pair. For the non-synesthetes, there is no category of previous self-report so we cannot know whether some of the pairings that they received in the experiment are more stable for them or not.

5.3 Discussions

I find several things that are of note in these findings. First important aspect found in this experiment is that the preservation rate of the self-reported pairings is much higher than of those that were determined through a prompting phase. This might suggest that this way of elicitation forces synesthetes to overgeneralize their synesthetic pairings.

Second thing is that there seems to be a larger sense of frequency in synesthetes than in non-synesthetes. For non-synesthetes, in 40% of the cases they decided to abstain from answering the question which means that the frequency difference between the subsittuted word and the substitute word was enough to question the pair but not enough to sway the answer. Alternatively, the prompting phase resulted in pairings chosen completely at random which would explain why they are indifferent towards their previous choices. The wrong answer – the one that does not agree with the pairings reported in the questionnaire – was chosen only a few times across the participants. At the same time, we see a that a third of the answers of the synesthete participants do not agree with the previous report. However, 28 of them are the pairings that had substitute word be the better match of the remaining word frequency wise. This, and the fact that there are only a few answers that the participants decided to be neutral on, suggests that the associaiton judgement of synesthetes is ruled by the affect of frequency of the words.

What I also want to mention is that, while measured, the time taken to answer the questions was not significantly affected by any of the variables such as the kind of answer that was given, whether participant was or was not a synesthete, and whether the question was a stimuli question or a filler.

6. Effect of word frequency on synesthetic associations

One of the main questions of this thesis is whether the frequency of words is the factor that decides whether words could be paired together. As mentioned in the literature review section, one of the tendencies that people found is that the more frequent words get paired together with other frequent words and the words with low frequency - or the non-prototypical members of the semantic group - a more likely to be in association with other words with low frequency (Simner et al., 2005; Association for Psychological Science, 2008). This work will investigate this question with the help of an experiment that will manipulate the pairings depending on their frequency.

6.1 Methodology

This experiment is happening at the same time as the experiment described in section 5. The set up of the experiment is the same. The difference between the experiments lies in how the pairings displayed on the screens are constructed. In the case of this experiment, none of the displayed pairings were reported by the responding participant. Both pairings are new and the difference between the two pairings is in the frequency. In one of the displayed pairings, both of the words will be from the same frequency category. In the second displayed pairing, the words will be from different frequency categories. For example, a possible combination of displayed pairings is thursday-purple and thursday-blue. In the example, ‘thursday’ is one of the less frequent members of the semantic group of days of the week. ‘Purple’ is also the less frequent member of the semantic group of colours, while ‘blue’ is one of the most frequent members.

What is going to be analyzed in this section of the experiment is whether the participants choose pairings that match in their frequency category or not. If the frequency hypothesis is correct, then the participants should choose the pairings wherein words match by their frequency. I calculate the percentage of chosen pairings that agree in frequency and see whether their representation is significantly greater than of the unmatched pairings.

6.2 Results

For the analysis of this experiment, I utilized both a qualitative and statistical method. First, I made a linear regression model on the answers of the experiment that checked whether the frequency of the second word and the association category of the pair could predict the frequency of the first word. What I got was a $2.2e^{-16}$ p-value. The model showed that the p-value for the predicting power of the frequency of the second word was 0.033, for the week-color category it was 0.023, and for the week-shape category it was $9.66e^{-7}$.

The conditional tree model showed the results that can be seen in figure 13 below. We can see that the relationship between the frequencies of the first and the second words changes depending on the association category of the pair with a p-value of less than 0.001. Overall, we can see that the week-shape, week-color, week-month, and month-color categories predict the value of the frequency category in a way that agrees with our proposed hypothesis. The other categories are either bad predictors of the frequency category, or their result is the opposite of the hypothesis – in their cases, instead of the pairing that was matched for frequency, the unmatched pair was chosen by most participants. You can see such cases in nodes 4, 5, 8, 9, and 12. In all of these cases, we see that even though the bar charts show the frequency categories of the first word depending on the category of the second word, there is no clear division between the common and the non-common categories like we see in the right side of the graph. These mean that for these association categories the commonality category of the second word does not predict the commonality of the first word well.

Seening these results, I had to do a qualitative analysis to see the possible explanations for the differences across categories. While most of the pairings showed that their responses agreed with the frequency hypothesis, others did not. This way, for the month-shape category, most of the times, participants chose the pair that did not match for frequency. The same goes for the month-week, and week-shape categorires. Across all cases, where the participants did not answer as they were expected to, there seem to be three main reasons for the diviation and I will exaplain them in the discussion section.

6.3 Discussion

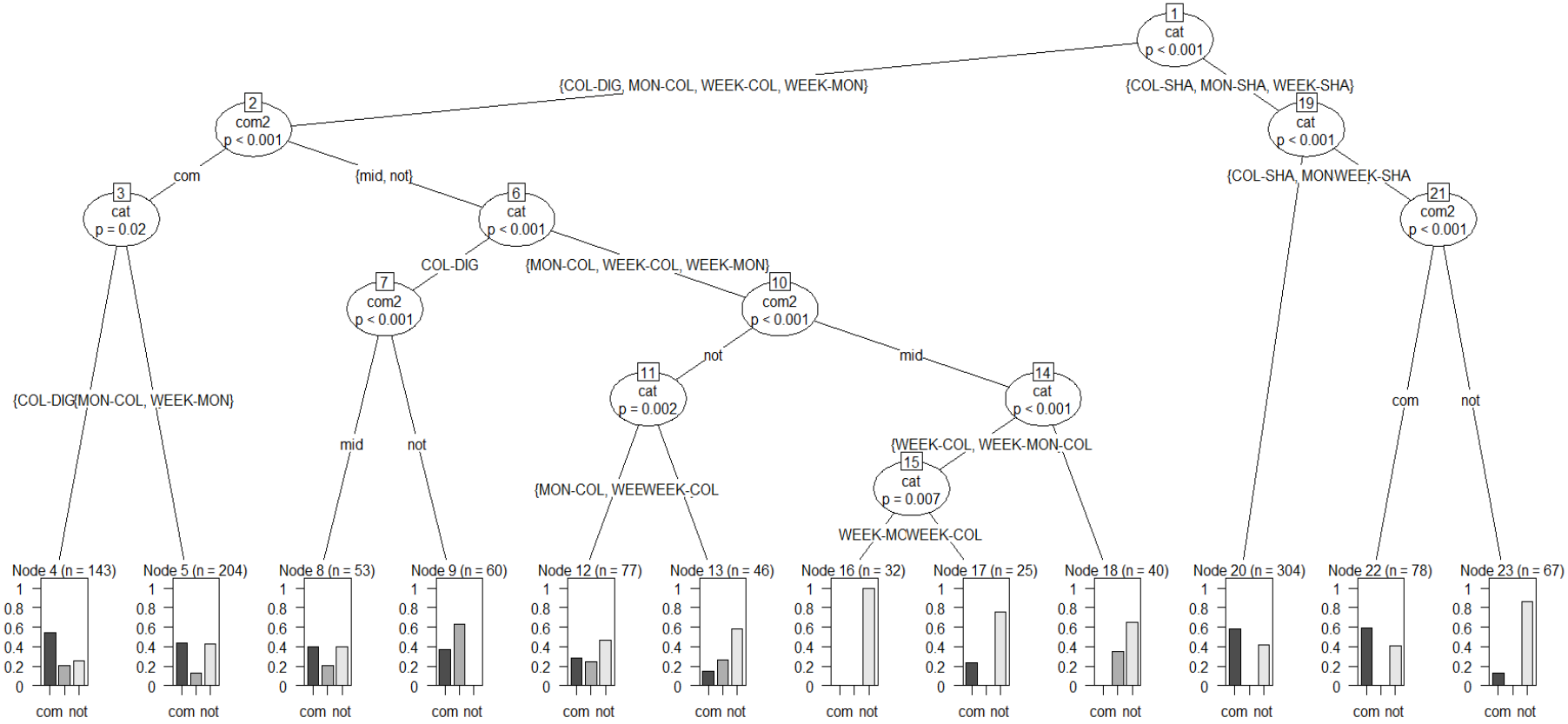
This experiment showed us a couple of things about NS. One is that for certain categories we can predict what pairings a person might have using the frequency hypothesis. While it will not work with a 100% accuracy, we saw that I was able to create pairings that were consistently chosen by other synesthetes. This also tells us that there are certain trends in pairings across synesthetes and we can find an explanation for them.

One unfortunate aspect of this experimental step is that I failed to recruit enough non-synesthetes to make a judgement on whether they have experiences similar to synesthetes. The description of the experiment was attractive to those who have synesthesia and did not interest those that did not.

Returning to the possible explanations as to why certain pairings were consistently chosen by the participants over the pairings that matched in frequency, the first one is that some aspect of a meaning affects the outcome of the pairing – something that has to do with the lived experience of the participant. This way, ‘august’ is paired with ‘Sunday’, even through the expected pairing was ‘august’ and ‘Wednesday’. In this situation, a possible interpretation of the pairing is that the people are associating August and Sunday for what the month and the day mean for their lifes. This way, both of them are the last month/day of rest before the work starts.

Figure 13

Conditional tree of the choices of pairings depending on the frequenct category and pairing category



The second possible factor is the association of the two words in association with some third word. An example of such is November and square. On their own, I do not see a connection in meaning, however, knowing the results of the questionnaire and the pairs that were chosen the most, I know that these two are the items among their semantic groups that were associated with black. Maybe, upon reading the pairs in this task, the participants had two synesthetic experiences where both of these words called for the same third thing, leading to their association.

The third factor comes from the works of Hancock (2013), Beeli et al. (2007) and Smilek et al. (2007). One of the topics in their work was that the Grapheme-Color synesthetic pairings are created based on the luminocity of the colors. This way, the more frequent items are associtated with items that are of a higher luminocity, and the less frequent items are associated with darker colors. Example of such would be that the participants chose november-black pairing over the novermber-yellow. Black is one of the most frequent colors, while yellow is closer to the middle of the pack. November is one of the less frequent months and was expected to pair with yellow. However, most of the participants paired it with black. While the frequency hypothesis fails to explain the pairing. What we do see is that a dark – less luminous color – is associated with one of the less frequent items. The same tendency – where the frequency of the color seems to matter less than its luminocity – is found for several of the categories. What is also important, for the cases where the frequency does seem to work to predict the chosen association pair, the luminocity factor is also met. For example, in such a pairing as green-Sunday, green is both luminous and frequent and it gets paired up with a frequent word ‘Sunday’ rather than the less frequent ‘Friday’.

Another factor that might be at play here is the similarity in sound, however it is only present in one of the situations which is already covered by the luminocity explanation.

7. General Discussion

Completing the different methodologies, we now have a more thorough understanding of the association pairs and groups created by people with NS and those that do not have it. The next sections will try to cover all of the questions established at the beginning of the thesis.

7.1 Frequent Association Pairs

As a result of the corpus, the questionnaire, and the experiment, we now have a collection of pairings reported and chosen by the participants. The first thing that needs to be mentioned is that, while the data resulting from all of the methodologies is not identical, there are some trends present as to what could be a pairing and some particular pairings that are present throughout the responses. One such trend is between colors and days of the week, and colors and months. It was already discussed in section 4, however, what needs to be added is that the same trends in colors are also present in the corpus and the preferred responses chosen in the experiment.

We can also find stable pairings in other categories as well. This way the associations between the days of the week and months are also quite stable. Some of them have a clear explanation. This way Friday and May are both the fifth member of their category in an ordered list, but also, for students and people in general, they are the last time span before the holidays: Friday is right before the weekends, and May is right before the summer season. Another such pairing is January and Monday both of which are the first items of their categories. However, there are other pairings for which we can find an explanation but it is significantly less direct than the ones I just mentioned. Such pairings are November-Thursday, September-Tuesday and August-Saturday. The four do not have a correspondance in the position in an ordered list or similarity in sound. My guess as to why these pairings appear at such a high rate will be present in the later section.

Next correspondance is between colors and digits. Just like with months and days of the week, there is a clear trend across responses as to what color corresponds to which digit. Below

is an illustration of the correspondence for you to judge. It does not fully agree with the findings that were reported by some of the articles on the topic of Grapheme-Color synesthesia, however, it does not need to. Although there are general trends for these associations, almost no participant reported this exact sequence of association pairs. While I cannot point out a particular similarity to some set of toys or books from childhood like in Witthoft & Winawer (2013), I do see this set of color-digit pairs as familiar.

Figure 14

Most frequent pairs in the color-digit association category

The trend for some pairings to be more frequent than others is also present in the color-shape associaiton category. For this category the most stable pairings are yellow star, black square, black/white/red circle, cyan/orange oval, green/blue rectangle, and purple rhombus. The yellow star and the black square are of course references to the usual depiction of starts and the "Black Square" painting by Kazemir Malevich. These two depictions are referenced quite often so these might or might not be cases of NS. Whether a person really sees them this way or not can only be checked through another experiment. As for the other color-shape pairings, we see a rather expected set – predictable through the frequency hypothesis. These representations were constant across all methodologies and languages, however, I am still unsure, why these exact pairings were chosen. If we look at the hypothesis alone, purple rhombus is as good of a pairing as orange rhombus, however, there is almost no mention of such a combination.

7.2 Association Chains

As was mentioned in the previous sections, the participants do seem to create association chains but not often enough for them to be analyzed statistically. A qualitative analysis of the chains brings us to a couple of inferences. The first one is that the chains are present in the responses of the synesthetes at a much higher rate than in the responses of non-synesthetes. Even in the self-report section, synesthetes mostly produced whole chains rather than pairings. The second one is that the members of a chain usually have something uniting all of them. This can be a position in an ordered list or the color. This way, some of the responses had the following chain: Tuesday-rectangle-blue-seven. Tuesday, rectangle, and seven all associate with blue on their own. There is no apparent reason as to why they should be chained to one another other than the associating color and their rather low relative frequency.

Another point was already mentioned in the previous sections but it seems that the association chains also results in some pairings being unmatched in frequency due to one strong connection chaining a usually unfit member into the chain. This way we have Monday-Red-March-3. Usually, Monday is paired with its mates in the position in an ordered list – January and 1. However, January is unfit to pair with red and 1 is unfit to pair with March. The association between March and red is also strong due to the experinces that we have in our society – in relation to the 8th of March and Nauryz being associated with red and happening in March. Because of the March-3 and March-red connections overpowering the Monday's connections but not denying the Monday-red connection we are left with a chain of associations that disagrees with the usual trends but stays stable across prompts and languages.

7.3 Synesthetes and non-synesthetes

First, I would like to conclude that NS synesthetes seem to be real synesthetes with cognitive experiences rather than simple associations. This comes from the results of the prompting phase and their responses in the experiment. During the prompting phase, it is possible to track that

the participants who claim to have synesthesia choose the same pairings that they mentioned in the self-report stage of the questionnaire. People should not show such a result if they just remembered what associations they have, as it would be too large of a mental load to track which pairs they should choose in a matrix. The prompting phase also resulted in pairs that show a different dependency on frequency across synesthetes and non-synesthetes. The pairs reported by synesthetes created a model that can predict the frequency of the first word of a pair depending on the frequency of the second word with a significant accuracy – the same analysis of the pairs created by non-synesthetes showed a non-significant result.

The stability of the self-reported pairings is reinforced by the results of the experiment sent out to both synesthetes and non-synesthetes to see whether they would choose the pairings that they previously reported. The outcome is that the synesthetes largerly chose the same pairings as the ones in their self-report. They also show a reliance on frequency in their choises – most probably an unintentional behaviour as they are unaware of the frequency hypothesis of this work. The non-synesthetes were also able to report the same pairs as they did during the questionnaire, however considering that they needed to choose out of two options, their success might be a result of chance. There is also a consideration about how most of the correct choises in the non-synesthete group were made by a couple of participants that repeated their responses with a 90% accuracy. In the synesthete group, all of the participants performed on the same level.

7.4 NS in Multilinguals

Not enough multilinguals reported their pairs in more than one language to make a statistical analysis, however, a qualitative analysis is still applicable to the case. The general trends were already described in section 4.2.6. What I could add in this section to further provide evidence for the fact that these synesthetic associations do transfer across languages is from the results of the first experiment. In the self-report phase of the questionnaire, the participants were free

to report their pairings/chains in either Russian, Kazakh or English. However, to standardize the experiment's stimuli, all of the questions were given in English. As the results show, the synesthete participants were successful at choosing the same pairings as in the self-report section independent of the language in which the self-report was first made. This seem to show that the pairings that seem to be the true synesthetic experiences transfer across languages that a multilingual knows.

7.5 Factors affecting the associations

This work considers several possible factors that affect the creation of particular associations. These are the sound form of the word, the similarity of meaning, the frequency of the words, and previous experiences. After conducting all of the steps of the methodology, this list expands with a couple of other – more specific – members that do a good job at explaining some of the association categories.

Most of the methodologies showed that the frequency of the words can be a good predictor of the likelihood of the words being associated. We can see that from the results of the statistical analysis of the corpus, of the material prompted through the questionnaire, and from the results of both of the experiments. The first experiment – investigating whether the participants are true synesthetes – shows that the frequency of the words in the proposed pairings affected their choices in a way that the pairings that matched in frequency were chosen more frequently than those reported in the questionnaire prompts. We also saw that the synesthetes had more reaction to the frequency of the words than those without synesthesia, showing that perhaps they have the unconscious – in the meaning of unintentional – sense of whether the words are frequent or not. The second experiment – targetted at researching whether the frequency is a good predictor of association – showed that frequency can be a significant factor in the choice of a better association. Another point that can be seen from the second experiment, is that the more distinct is the matching the more the number of participants

that agreed on one option. This way, for opposing pairings brown-nine and brown-four, brown and nine are matched quite well by frequency, while four is much different from brown. This resulted in an overwhelming majority of the participants choosing brown-nine. The same trend can be seen in other pairings as well – the more is the difference between the frequencies of the unmatched pairing, the more is the agreement across the participants that it is the worse pairing.

However, what the second experiment also showed is that there seems to be a hierarchy to how the factors affect the associated pairs. As could be seen from figure 13, there are certain cases that were not ruled by the frequency hypothesis. Looking closely into the data from this experiment, I find that all such cases can be explained by separate factors. This now leads me to the idea that the factors take different precedence in deciding the associations. For example, for the month-week association category, results for only 2 out of 6 pairings agreed with they frequency hypothesis. In all of the other cases, it seems that either the similarity in semantic meaning or experiences took precedence. In the cases where the colors were paired with other semantic categories, I saw that what decided the association pair was not the frequency of the color but rather its luminocity – which agrees with Hancock (2013), Beeli et al. (2007) and Smilek et al. (2007). An example of such a situation are pairings november-yellow and november-black. Although black is the more frequent color and does not match november, most of the participants choose their pairing over november-yellow. However, this can be explained by the luminocity affect. Black is not a luminous color which makes it appropriate for a pairing with a non-frequent word such as november.

Another set of cases can be explained by the similarity in meaning – or rather in how we experience the referents of the words. This way, in a choice between frequency-matched August-Wednesday and unmatched August-Sunday, most participants prefered the second option. My guess is that the similarity comes from the similarity in how August is the last month of holidays, the same way Sunday is the last weekend before Monday. Both August and Sunday

in our experiences are filled with rest and recreation rather than work. This idea comes to me from one of the comments left by the participants during the prompting phase. They said that they feel a connection between Thursday and November exactly because there is nothing happening in those days/months. Nothing is associated with these days/months – like the beginning or the end of work – and usually nothing is scheduled for these time spans. I think that this is the kind of experiences that Cytowic & Eagleman (2011), talked about while saying the association between colors and days of the week appear as we live, as we experience life.

Two other factors that should also be considered are the similarity in sound and the similarity in semantics – in the meaning of the same position in an ordered list. In the cases where the similarity of sound was present in the pairing that was unmatched for frequency, the participants choose both of the options to the same extent, seemingly meaning that the two factors have about the same power over the creation of the synesthetic associations. The similarity of sound also plays a special role in the results of the prompting phase of the questionnaire. In several instances, in the association categories that included shapes, the participants left comments that the word associated with a particular shape sounds like the shape. This seems to be similar to the idea of sound symbolism, however, I lack the experimental method to check how true this relation is.

As for the similarity in the position of the ordered list, there is a particular trend that is present mostly in the datasets of the prompting phase rather than in the experiment. What we see there is that the participants are good at associating the last and the few of the first items of the ordered lists together. The closer to the middle we go, the less predicting power this factor has. This probably has to do with availability of information. For the first few items, it is easy to know from the top of your head which position they take. This becomes harder the deeper into the list you go. The availability of information also seems to be important from the results of the prompting phase of the participants that are not proficient in Kazakh. That their responses

show is that they do not have the information on the order of days of the week and months readily available. Because of that, their association pairs rarely include pairings that are based on this similarity, even though such logic is present in their responses in English and Russian. If the order information is unavailable, the respondents mostly choose their pairs based on the similarity of sound, skipped the options, or chose those pairings that agree with the frequency hypothesis.

What this discussion leads me to, is to a conclusion that there is a hierarchy as to what factors predict the association pairs. If one of the first factors is anavailable, the next one becomes the deciding factor. The hierarchy is such that the first deciding factor is the experiences that we have with the referents of the words. Next, we have individual cases such as luminocity of color, and sound symbolism between words and shapes. After that we have a similarity in the position of the items in an ordered list and the match between the position and the number of angles of a shape. Lastly, we have the similarity in sound and the similarity in frequency. The experimental results show that they affect the observers/synesthetes to a similar degree, however, what needs to be mentioned is that they apply to a different amount of cases. The words under investigation do not have a lot of cases that match in sound, this way, the sound similarity factor covers only a few cases. At the same time, the frequency factor covers most of the cases left after the other factors – contributing to the prediction of association on most of the cases.

7.6 Concept or word?

The two factors that seem to be important for this consideration are the transfer of associations across languages and the factors that affect association creation. As we saw from section 7.4, the association pairs seem to transfer across languages – from the early acquired languages to those acquired at a later age. If the association was happening on the level of a word, I would imagine that the words that refer to the same referent in different languages would have

different association pairs. We can see that in responses of non-synesthetes that pair ‘february’ with ‘blue’ in English, but with ‘white’ in Kazakh due to the fact that in Kazakh ‘white’ and ‘february’ start from the same syllable.

The second point is present in section 7.5 on the factors affecting association. We find that most of the cases of pairs built upon sound similarity cannot be considered as synesthetic experiences – rather they create simple associations. At the same time, the sound of the words could affect synesthetic pairings across languages as was demonstrated in the example of a Spanish-English speaker in section 3.2.1. The participants also report that the words sound like particular shapes for the prompts that paired shapes with other semantic groups – like days of the week and months. In that moment, the participants do not talk about imagining the concept of the referent taking a particular shape. Rather they say that the combination of sounds that create the word naming the referent sounds like a shape. This way, participants claimed that ‘Tuesday’ sounds rectangular or that ‘June’ sounds like an oval. All these reminds me of the discussion on sound symbolism, however, I do not have an experiment in this thesis that would check for its probable affect on the creation of synesthetic experiments.

Overall, a singular correct answer as to where the association happens – on the level of a word or a concept – is unavailable. It is possible that the associaitons might happen on different levels depending on the situation. It is also possible to argue that the ‘shape’ the word takes could be a part of the conceptualization of a word in a particular language that may or may not transfer to another language. However, at this point, I do not have the data nor the methodology to make a definitive claim.

8. Conclusion

In this thesis, I wanted to answer three main questions: are there association pairs that are more frequent than others and appear across participants; what are the factors affecting the formation

of particular associations – or, more specifically, whether the frequency of the words can affect the likelihood of them being associated; and whether the associations happen on the level of concepts or words.

While I am unable to answer the last question in a way that would satisfy my own standards, I can try to answer the first two. My findings show that some of the association pairs are more frequent than others to a significant extent. Figures 9, 10, and 14 show them in an accessible way and you can be the judge of whether you agree with the general trends or not. While these results are not universal and to not show a response of a particular participant, these are the modes of the responses which also agree with the outlined factors predicting the association of words.

The factors – discussed in detail in section 7.5 above – are the main part of this research work. Looking at the results of all of the methodology stages, we can see that it is impossible to single out one factor that decides the association pairings. All of the deciding factors exist together and interact in a way that seem to build a hierarchy. Agreeing with Cytowic & Eagleman (2011), the main deciding factor seems to be the experiences the people live through. This factor creates both, the associations that are synesthetic and non-synesthetic. A harsh division between the two will be close to impossible to make since only the person themselves knows what experiences they have and their description can easily be false or misunderstood. The second factor seems to be individual for each of the association categories. When the association category includes the semantic group of colors, the luminocity of the color might be the factor affecting the associations. When one of the categories is shape, sound symbolism might be the reason another word is pared with a shape. These factors don’t seem to be generalizable, so they take the second step of the hierarchy together. Then, we have associations created based on a similarity in meaning. These include the associations between objects that take the same position in an ordered list – like how Monday is the first day of the week and

January is the first month. These associations are strong but can still be overpowered by the associations created by experience as seen from the discussion on association chains. The similarity in sound is somewhere near the position of the frequency but since it often creates non-synesthetic associations, it is hard to judge as to how productive it is without having to check each experience on whether it is trully synesthetic or not. And the focus of this thesis – the frequency of words as the predictor of the associations. The results show that it can be a significant predictor, however, mostly for the cases where the other factors cannot be applied. Still, it does the heavy lifting in explaining those associations that are unexplainable on the first glance.

9. Limitations

The main limitation of this study is that there is a possibility that participants responded in a manner that is false of their true experiences. While a single response should not affect the statistics severely enough, it is still unfortunate that the reported information might not represent the reality of the phenomenon. As most of the participants are anonymous and do not get anything for their participation, it is impossible to prevent them from answering in a joking manner. Several of the responses clearly showed that the answers were not serious and were deleted from the data.

Second issue is that the phenomena under investigation covers quite a lot of material. In my methodology, I wanted to pay attention to all of the semantic groups equally which turned into a procedure that takes a lot of time. This led to a number of participants not finishing their questionnaires. While they still provided enough information for it to be helpful for the analysis, it was a lost opportunity to get a full picture of the associations across languages of a multilingual. I think that a possible solution for this program is utilizing a different method - instead of a matrix. I am yet to be able to come up with a method that will not overwhelm the participants while also covering all of the semantic groups.

Another issue is that many of the non-synesthetes participating in the questionnaire spent a lot of time trying to pair all of the items of the matrix. This was not a requirement of the task - as was explicitly stated in the instructions, however, I think that the participants tried to provide as much information as possible in an absence of an authority that would reassure them that it would have been fine if they left a matrix with only 1-2 pairings. I think that this has a simple solution of conducting the questionnaire while a researcher is present for questions and in-person instructions. At the same time, such a set-up would decrease the number of questionnaires left unfinished. However, that creates a situation where the number of answers collected in the same period of time drops drastically. I am unsure as to what set up would be more beneficial for research and I think it depends on the framework of the researcher.

Most of my statistical analysis depends on the relative frequency that I acquire from the corpus. One problem with that is that some words under investigation hold more than one meaning. This way, English 'one' and Kazakh *bir* 'one' could be used to name people as in "That is the one I was looking for". Or English 'march' that is also used in a sense of 'marching'. While I tried to estimate their frequency only for the meaning that is appropriate for this study, there is also a factor that the association to other items can come from the other meanings of the words.

10. Ethics

All of the methodologies explained in sections 4, 5, and 6 received an approval from the SSH IREC committee. All of the collected material was annonymized before being added to the GitHub repository below. The participants would not be identifiable from the information present in the repository. Any identifiable information collected through the methodologies – i.e. e-mails of people that agreed to participate in the second step of the experiment – have been deleted after a randomly generated identification code was assigned to all of their information.

References

Association for Psychological Science. (2008, April 30). Consistencies Found In Synesthesia: Letter 'A' Is Red For Many; 'V' Is Purple. *ScienceDaily*.

Baron-Cohen S, Wyke MA, Binnie C. 1987. Hearing words and seeing colours: an experimental investigation of a case of synaesthesia. *Perception* 16:761–67

Boersma, P. (2006). Prototypicality judgments as inverted perception. *Gradience in grammar: Generative perspectives*, 167-184.

Geeraerts, D. (2007). Where does prototypicality come from. *The cognitive linguistics reader*, 168-185.

Gian, B., Esslen, M., Jancke, L. (2007). Frequency correlates in grapheme-color synaesthesia. *Psychological Science 18* (9), 788-792.

Carriere, J. S., Eaton, D., Reynolds, M. G., Dixon, M. J., & Smilek, D. (2008). Grapheme–color synesthesia influences overt visual attention. *Journal of Cognitive Neuroscience*, *21*(2), 246-258.

Cytowic, R. E., & Eagleman, D. M. (2011). *Wednesday is indigo blue: Discovering the brain of synesthesia*. MIT Press.

Davies, M. (2004) British National Corpus (from Oxford University Press). Available online at https://www.english-corpora.org/bnc/

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior research methods*, *40*(1), 213-231.

Eagleman, D. M., Kagan, A. D., Nelson, S. S., Sagaram, D., & Sarma, A. K. (2007). A standardized test battery for the study of synesthesia. *Journal of neuroscience methods*, *159*(1), 139-145.

Fitzpatrick, T., & Izura, C. (2011). Word association in L1 and L2: An exploratory study of response types, response times, and interlingual mediation. *Studies in Second Language Acquisition*, *33*(3), 373-398.

Hancock, P. (2013). *Synesthesia, alphabet books, and fridge magnets* (pp. 84-99). Oxford, UK: Oxford University Press.

Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D. (2004). The Sketch Engine. Proceedings of the 11th EURALEX International Congress: 105-116. Available online at https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen21_tt31

Llewellyn, C., Grover, C., Alex, B., Oberlander, J., & Tobin, R. (2015). Extracting a topic specific dataset from a Twitter archive. In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19* (pp. 364-367). Springer International Publishing.

Lowie, W., Verspoor, M., & Seton, B. (2010). Conceptual representations in the multilingual mind. *Converging Evidence in Language and Communication Research* (CELCR), 135-148.

Madiyeva, M.E. (2013). Almaty Corpus of Kazakh. Available online at http://web-corpora.net/KazakhCorpus/search/?interface_language=ru

Marks, L. E. (2013). Weak Synesthesia in Perception and Language. In J. Simner & E. Hubbard (Eds.), *Oxford Handbook of Synesthesia* (pp. 761-789).

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and weak. *Current Directions in Psychological Science*, *10*(2), 61-65.

Maurer, D., Gibson, L. C., & Spector, F. (2013). *Synesthesia in infants and very young children* (pp. 46-63). Oxford University Press.

Nielsen, A., & Rendall, D. (2011). The sound of round: evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *65*(2), 115-124.

Paulsen, H. G., & Laeng, B. (2006). Pupillometry of grapheme-color synesthesia. *Cortex*, *42*(2), 290-294.

Purohit, N. S., Angadi, A. B., Bhat, M., & Gull, K. C. (2015, February). Crawling through web to extract the data from Social networking site-Twitter. In *2015 National Conference on Parallel Computing Technologies (PARCOMPTECH)* (pp. 1-6). IEEE.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synesthesia--a window into perception, thought and language. *Journal of consciousness studies*, *8*(12), 3-34.

Rosch, E. (1977). Classification of reald-world objects: Origins and representations in cognition. In Philip Johnson-Laird and Peter C. Watson, eds., *Thinking: Readings in cognitive science*. Cambridge: Cambridge University Press.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

Rouw, R., Case, L., Gosavi, R., & Ramachandran, V. (2014). Color associations for days and letters across different languages. *Frontiers in psychology*, *5*, 64826.

Savchuk, S. O., Arkhangelskiy, T., Bonch-Osmolovskaya, A. A., Donina, O.V., Kuznetsova, Y. N., Lyashevskaya O.N., Orehov, B.V., Podryadchikova, M. V. (2024). Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, (2), 7-34. Available online at https://ruscorpora.ru/en

Simner, J., Ward, J., Lanz, M., Jansari, A., Noonan, K., Glover, L., & Oakley, D. A. (2005). Non-random associations of graphemes to colours in synaesthetic and non-synaesthetic populations. *Cognitive neuropsychology*, *22*(8), 1069-1085.

Simner, J., & Hubbard, E. M. (Eds.). (2013). *The Oxford handbook of synesthesia*. OUP Oxford.

Simner, J., & Bain, A. E. (2013). A longitudinal study of grapheme-color synesthesia in childhood: 6/7 years to 10/11 years. *Frontiers in human neuroscience*, *7*, 603.

Smilek, D., Carriere, J.S.A., Dixon, M.J., Merikle, P.M. (2007). Grapheme frequency and color luminance in grapheme-comor synaesthesia. *Psychological Science 18* (9), 793-705.

Svantesson, J. O. (2017). Sound symbolism: the role of word sound in meaning. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(5), e1441.

Uznadze, D. (1923). Ein experimenteller Beitrag zum Problem der psychologischen Grundlagen der Namengebung [An experimental contribution to the problem of the psychological foundation of naming]. *Psychology Forsch, 5,* 24-43.

Witthoft, N., & Winawer, J. (2013). Learning, memory, and synesthesia. *Psychological science, 24*(3), 258-265.

Appendix

As a part of the thesis work, I collected a corpus and a database on synesthetic and associative pairings. Here is a link to the repository containing all of the collected material in an annonymized form and R code for the statistical part of the analysis: https://github.com/aida-isteliyeva/Synesthesia-of-early-acquired-concepts

My experiments were conducted through an online platform called Qualtrics. The preview for the questionnaire from section 4 can be found here:

https://nukz.qualtrics.com/jfe/preview/previewId/fc73c579-4d1d-41d4-82ed-6182da226d3f/SV_37WAYqKtCKjQaiy?Q_CHL=preview&Q_SurveyVersionID=current

The preview for the experiment in section 6 can be found here:

https://nukz.qualtrics.com/jfe/preview/previewId/3b085936-6da8-4246-ae07-e4e5d74ef093/SV_9XjvnnAOFIAU050?Q_CHL=preview&Q_SurveyVersionID=current