Tel-Aviv University

Lester and Sally Entin Faculty of Humanities

Department of Linguistics

Three Men Walk into a Bar: Quantifying Phonological Distance Between Languages on a Universal Scale

MA Thesis submitted by

Alona Golubchik

Prepared under the guidance of:

Dr. Evan-Gary Cohen

July 2022

Acknowledgments

I was never sure what I wanted to do in life. Nothing really interested me, except for reading books and writing stories. In the darker times of my life, books were my comfort, stories were my comfort, *words* were my comfort. When it was time for me to decide what I should learn in university, all I knew is that I wanted to learn something with words – literature, translation, language editing – but I still felt that this was not quite what I was looking for. And then I met Dr. Evan Gary-Cohen at the open day in Tel-Aviv University, and I discovered linguistics, and more specifically – phonetics and phonology. And suddenly I knew that *this* is what I wanted to do in life, even though the words we examine in phonology are not *written* but *spoken*. The spoken words have great influence on the written words and their formation. Knowing where speakers "make mistakes" orally can help us determine why they also make mistakes in writing. I had found a new comfort.

This thesis went through quite a bit (COVID, writer's block, and the accidental deletion of the entire almost-finished draft), and I could not have overcome all these hardships without much needed help from the people around me.

First and foremost, I would like to thank my advisor, Evan-Gary Cohen. There are no words that can even begin to express my gratitude for being a great mentor, a considerate colleague and a (mental) shoulder to cry on. For reminding me from time to time that someone is waiting for me to finish my research. For supporting my crazy idea of a study. For giving me so many opportunities to pursue my dream.

Second, I would like to thank Lior Laks for reading and commenting on my thesis proposal, thus providing me insightful remarks, which eventually lead me to research on matters that I had not thought about before.

I would also like to express my gratitude to the other faculty members whose courses I participated in: Galit Adam, Mira Ariel, Outi Bat-El, Irena Botwinik, Aya Meltzer-Asscher, Nirit Kadmon, Roni Katzir, Fred Landman, Ezer Rasin, Einat Shetreet and Tali Siloni. Special thanks are also in order to the teaching assistants in these courses: Daniel Asherov, Hila Davidovich, Alma Frischoff, Roman Himmelreich, Yuval Katz, Nicole Katzir, Radan Nasrallah, Tal Ness, Dori Sharabani and Itamar Shefi. Without all of you, I would not have known that I love linguistics so much. I also thank Ruti Zussman, the department secretary, who was always glad to help when I was confused about the bureaucracy of the institute, or when I just needed an advice.

I would like to thank the participants of my experiments, and the audiences of the TAU colloquium and the Israeli Phonology Circle for providing their insight on my research and my thesis defense.

Of course, I did not make my linguistic journey alone. Special thanks to Nitsan Cohen, who helped me greatly in both the linguistics department and in my personal life, was my teaching assisting colleague, and also read this thesis and edited it before its submission. Additional thanks to dear my linguistics friends: Mandy Cartner, Lital Elkaiam, Eugenia Kosolapov, Vera Rusyanov (also a fellow teaching assistant), Nofar Rimon, Or Shapira, Anat Sageev, Tomer Yahalomi and Hadas Yeverechyahu. I was also deeply privileged for having such great teaching assisting fellows throughout the years: Shachar Bieber, Peter Brauth, Yoni Heilig, Carlo Meloni, Ziv Plotnik-Peleg and Hadas Zaidenberg. To all of you, I would like to express my sincere gratitude for advising me along the way, or for just being there when I needed a push. I wish you all the luck in whatever path you choose.

Last but not least, I would like to thank my family: my parents who supported me

in any way I needed, and still do so (I promise to move out soon!), my little brothers who guided me technically in times of need, my late grandmother who kept telling me how very proud she was of me and told me so many stories about the linguistics of Russian, my cat Sabaka ("a dog" in Russian. I know it is funny since he is a cat) who knew when I needed a cuddle, and to my dearest non-linguistic friends, who are practically a family to me – Giora Glotser (who was lately seduced to the linguist side too), Ilana Miroshnikov and Raphael Zachis Caravaggi (with his dog Wow!) – who did not give up or run away when I expressed my endless frustration throughout the preparation of the experiment and the writing of the thesis.

I promise my journey has only just begun.

Abstract

Many researches have studied the similarity between languages (e.g. Eden 2018; Crowley and Bowern, 2010; Longobardi and Guardiano, 2009, 2017), but there is no research which quantifies the similarity between languages. The final goal of this study is to examine whether similarity can be measured and quantified using the scales of the acoustical prominence of several phonetic and phonological properties, while merging them into one universal scale of prominence. However, since there is no research in which similarity is measured by phonetic and phonological features alone, the goal of my thesis was to examine which features should be placed in this scale in the first place.

This study contains two experiments, a preliminary one and a main one. In the preliminary experiment, 132 Hebrew speakers rated their familiarity level with each of the 35 languages that appeared in the main experiment. In the main experiment, 362 Hebrew speakers listened to 20 sets of three recordings, a base language and two additional languages, and were asked which of the two additional languages was more similar to the base language. The similarity was determined by the number of the shared features between the base language and the other language, and the features (a total of 41) were taken mostly from the World Atlas of Language Structures Online (WALS) and from Bradlow et al. (2010). One of the additional languages shared more features with the base language (the similar language) and the other language shared fewer features with it (the dissimilar language). The results showed a significant inclination to choose the more similar language over the dissimilar one.

These findings suggest that the similarity can be measured by phonetic and phonological features. However, we know that not all features were created equal; thus, this model can be upgraded by weighting the features, so that more prominent features

will have more weight in similarity quantification. I leave the weighting of the features for future research.

Table of contents

1. Introduction

Three men walk into a bar. This bar specializes in keeping their clients' privacy by separating the tables with a curtain preventing clients from seeing the other tables. The three men sit down and enjoy a dinner when they hear a faint chatter at the adjacent table. One man wonders what a Romanian speaker is doing in their country; another man says that the speaker is Korean; and the third one argues that he is Portuguese. A few hours later, they get up from the table and they see, on their way out, a glimpse of the client who had been sitting next to them the whole time. They were surprised to find out that, without a doubt, the speaker was Japanese.

You must be wondering why each man identified the spoken language differently from the others (and still none of them identified it correctly). Now I can tell you that the first man's L1 was language A, the second man's L1 was language B and the third man's L1 was language C. None of the men knew Japanese, and none of them heard the speaker clearly, but only some linguistic properties of the language he spoke. Each man might have perceived different properties, or perhaps perceived the same properties differently, depending on their L1 (otherwise, they would have agreed on which language had been spoken).

Many researches have studied the similarity between languages: counting phonological features, using cognates, applying various computational methods, or acoustic measures (and many more methods – see §2). While all of these methods seem to work to some extent, they address different kinds of similarity, and it seems that language comparisons are more complex than using only one parameter to compare them. In addition, almost none of them tries to *quantify* the overall similarity between languages. The final goal of the current study will be to measure similarity between

languages using scales of the acoustical prominence of several phonetic and phonological properties and merging these scales into one, universal scale of prominence, which we will be able to use to predict how speakers quantify similarity between languages. However, since there is no research in which similarity is measured by phonetic and phonological features alone, the goal of my thesis was to examine which features should be placed on this scale in the first place.

The outline of this thesis will be as follows: in §2, I will present some previous research on similarity; in §3, I will present my research question and the hypothesis, in §4, I will elaborate on the experiment I based my own experiment on – The Great Language Game; in §5, I will explain my own experiment in detail (participants, material, procedure and results) and will discuss the results; in §6, I will offer a few non-phonological properties that could affect similarity; and in §7, I will conclude the study.

2. Theoretical Background

Let us start with the most fundamental of questions – what is similarity? It's not that we do not know what similarity is. We *do* know, but we do not know *what* we know. For example, is green more similar to yellow or red? Now think about the answer that automatically popped into your mind. Why did you choose that answer? When I ran a small quiz around, everyone answered 'yellow', yet no one could really tell me why they had chosen that answer.

Then I asked them if green is more similar to yellow or blue, and the answers started to vary, yet most still answered 'yellow'. Interestingly, one of the participants tried to use 'precise' measures and argued that green is 'closer' to yellow on the color scales, but when they tried to prove that to me, they found out that green was actually right in the middle of blue and yellow. Of course, it depends on what you have in mind when I say 'blue', 'yellow' and 'green', as each color has a scale of its own. A question we can ask is – given the exact same input, will every person around the world give the same answer you did?

Sometimes, you *do* know why you chose a certain answer. For example, when I asked if a motorcycle is more similar to a bicycle or a car, some answered it was more similar to a bicycle because "they both have two wheels", and some answered it was more similar to a car because "they both have an engine and drive fast". In other words, some people compared the appearance of the objects, and some compared the function of the objects. Some people were even so sure of their answer that they said, "well, obviously X". Then, I asked whether a knife is more similar to scissors or a fork. This is a more complicated comparison because the answer might not be intuitive, and indeed I got both answers again, but I also got the answer 'neither'. People who chose 'scissors' struggled

to explain why they had chosen it, while the ones who answered ‘a fork’ said it was ‘because they are both tools used for food’. Now let me tell you that I asked the first question in Hebrew and the second question in Russian. This is important information because in Hebrew ‘bicycle’ is pronounced [o.fa.ˈna.im], ‘motorcycle’ is pronounced [o.fa.ˈno.a] and ‘car’ is pronounced [me.xo.ˈnit], while in Russian ‘knife’ is pronounced [ˈnoʐ], ‘scissors’ are pronounced [ˈnoʐ.nʲi.t͡sɨ] and ‘fork’ is pronounced [ˈvʲiɬ.ka]. All Hebrew speakers said that a motorcycle is more similar to a bicycle and all Russian speakers said that a knife is more similar to scissors. Other people chose both options. In other words, people can use their language as a relevant feature when comparing the similarity of objects. If this is the case, can it be that speakers of different languages compare the similarity of languages differently, based on their knowledge of their own language?

2.1 Similarity Between Languages

It seems similarity depends on the observer’s subjective perspective (Ringbom, 2007:7), i.e. a speaker of one language will perceive some properties of two languages as being the most similar out of three (or more) given languages, and a speaker of another language will perceive other properties of the same languages, and can determine that two other languages are the most similar. We must keep in mind that when a listener observes some unknown language, he uses his prior knowledge on languages, namely his L1 and other languages he might know, to develop a strategy of discrimination between these languages (Vasilescu et al., 2000, 2005 & Barkat and Vasilescu, 2001). For example, Hyman (1970) used English loanwords to examine whether [ð] is closer to [z] or to [d] and found out that French speakers adapt English’s [ð] as [z] and Serbo-Croatian speakers adapt it as [d], even though both [z] and [d] appear in both languages’ inventories. Tversky (1977) says that

there are some stimuli (e.g. faces and countries) which are represented in terms of many qualitative features – I would like to think that languages are represented in our mind the same way. Bradlow et al. (2010) suggest that there might be some sound structure features that have general salience and these features are important to quantify similarity (or rather, the difference) between languages, regardless of the listeners' language background. Bradlow et al. (2010) provide a list of what these features might be, and I will test some of them in my thesis.

Eden (2018) describes several computational comparison methods (e.g. Cognate-base similarity – Crowley and Bowern, 2010; McMahon and McMahon, 2005; mathematical approaches – Longobardi and Guardiano, 2009, 2017; Longobardi et al., 2013; and more) and concludes that the Parametric Comparison (which relies on binary features) and the Cross-Entropy (which relies on the probability of occurrence of some element in a given message) methods are the most reliable when comparing two languages. However, we can see on the surface that speakers do not compare languages using binary features alone. In addition, we cannot use these methods to compare more than two languages: for example, when using one of the two methods listed above, we can establish that Spanish is similar to Portuguese (relatively to all languages) and that Portuguese is similar to Russian (relatively to all languages). Does that necessarily mean that Spanish is similar to Russian when considering all world languages? In other words, is similarity transitive? We cannot be sure of that. This transitivity (triangle inequality, Tversky 1977) is a fundamental problem for many similarity models, and will be discussed thoroughly in my thesis.

When comparing two languages *relatively* to other languages to observe the similarity between these languages, we will not, most likely, observe them only by their

segmental properties, but also by some other properties, as prosodic properties, metrical and intonational structures, phonotactic properties and syllable shapes (Bradlow et al., 2010). However, in order to quantify the overall similarity of languages, we cannot observe different properties separately, but rather we must find a way to normalize all of these properties on a single scale. One possible way to normalize the properties is through observing the acoustics and confusability (i.e. the more confusable two sounds are, the more perceptually similar they are to one another) of phonological features, instead of observing mere features (Steriade 2001, 2001/2008 and Cohen 2009), and this was the main focus of my thesis. For example, Zwicky (1976) notes that nasals are more similar to one another than stops are to one another; i.e. the confusability rate of nasals is higher than the confusability rate of stops. Can we determine which phonological properties form confusability? Rather, can we normalize the confusability rates of all phonological properties onto one scale of confusability rates?

Following Shinohara (2006), who holds that perceptibility scales are universal, I would like to suggest that the similarity (confusability) scale is universal as well, and that the language's quantification of similarity depends on its acoustic, phonetic and phonological properties. However, before constructing some universal similarity scale, it is important to note previous research on similarity within each phonetic and phonological factor (i.e. segments, phonotactics and prosodic rhythms – stress pattern, pitch accent and intonational phrases).

2.2 Phonetic and Phonological Properties of Languages

The examples I provided in the introduction for the different languages the three men chose when they heard Japanese were not random. The man who speaks language A thought it was Romanian, which is segmentally similar to Japanese, yet is different from it in

phonotactics and stress pattern; the man who speaks language B thought it was Korean, which is similar to Japanese by prominence pattern (or rather, the lack of stress), yet is significantly different from Japanese both segmentally and phototactically; and the man who speaks language C thought it was Portuguese, which is similar to Japanese in its phonotactics, yet is significantly different from it both in the segmental and stress pattern aspect. Each speaker relied on a different property of the language they heard and compared this property to the properties of other languages they had heard before. Or so I would like to think.

Following this assumption and the background given above, the speaker of language A must have identified the language giving more weight to its segmental features; the speaker of language B must have identified the language giving more weight to its prominence patterns; and the speaker of language C must have identified the language giving more weight to its phonotactics. i.e. each man chose a different property, which must have been based on their knowledge of languages, and each speaker chose the property he mostly based his comparison on using the properties of his L1 and other languages he knows.

Cole (1973) showed that if we change one feature of any segment in syllables, speakers will not notice the change (though changing two or more features will already be noticeable). It means that as long as the phonotactics are intact, and as long as the segmental change is minimal (the question here is – what is 'minimal'?), speakers will not notice the difference. In other words, Cole (1973) implies that phonotactics are more noticeable to speakers than segments. However, the experiment done by Cole (1973) was done on English speakers with real English words, which is different from identifying a new, unknown language. In addition, not all features are necessarily identical, as we are used to

perceiving some values of some features more (i.e., less marked) than the opposite values of these features (i.e., more marked) (e.g. speakers of almost all languages will be able to tell the difference between [+sonorant] and [-sonorant], since sonority is universally a distinctive feature in languages. However, not all speakers will be able to tell the difference between [+constricted glottis] and [-constricted glottis], since glottalization is only distinctive in a relatively small number of languages).

In addition, Leena et al. (2005) show that the automatic language identification (= LID), a computational program used to identify languages, uses both phonotactics and prosody to identify languages, and Zissman (1996) argues that phonotactics are the most powerful features that LID uses. In both papers, they note that syllables of languages differ in the frequency of occurrence of certain syllables, in possible co-occurrence of syllables, in unique syllables and in pronunciation variations, even in the same syllable. But, contrary to these studies, Leena et al. (2004) show that segmental features also have an impact on language identification. Therefore, all three factors can reportedly influence the perception of language similarity. The question is how much influence does every factor have on language identification?

Before answering this question, we must obtain some background on each of these factors.

2.2.1 Segmental Similarity

Segments are traditionally divided into two groups: consonants and vowels. There is also some variation in the similarity difference within these two groups, as vowels are seemingly more similar to each other than consonants are to each other (Turnbull and Peperkamp, 2017). In my thesis, we examined these groups differently as well, comparing languages with similar consonants inventories but different vowels

inventories, and vice versa.

As was mentioned above, Zwicky (1976) showed that the nasals' confusability rate is higher than the plosives' confusability rate. In addition, some researchers (e.g. Garnes and Bond, 1980; Hung, 2000) show that the confusability rate between liquids is high as well, and some researchers (e.g. Meng et al. 2007) even show that the confusability rate between liquids and nasals is high (i.e. many speakers confuse between nasals and liquids), and that the confusability rate between liquids and glides is high (i.e. many speakers confuse between liquids and glides). Note that if we compare the findings above with the sonority scale of consonants (Clements 1990; see the following (1)), we can see that the more sonorant segments are, the higher their confusability rate is. Therefore, it seems as if the confusability rates of consonants might be determined (to some extent) by the sonority scale (or the other way around), which in turn implies that the sonority scale might help us build the universal similarity scale. However, sonority is probably not the only property which determine the consonants' confusability, and some properties have more effect on confusability than others.

(1) The sonority scale of consonants (Clements 1990)

(Vowels) >> Glides >> Liquids >> Nasals >> Voiced Obstruents >> Voiceless Obstruents

Following this line of thought, we can also try to use the sonority scale for vowels, to determine their position on the similarity scale:

(2) The sonority scale of vowels (Parker 2008)

Low vowels >> Peripheral vowels >> Interior vowels

2.2.2 Phonotactic Similarity

Phonotactics are a little harder to quantify than segments, since the confusability rates of each syllabic position (i.e. onset, nucleus and coda) is not absolute, but contrast dependent (Steriade 2001, 2001/2008). In other words, the prominence of a syllabic position depends on which segment is mapped into that position. As was mentioned above, the more prominent a syllabic position is, the less the segments in this position will be confused with other segments. For example, we can observe (separately) the prominence scales of onsets, nuclei and codas (Prince and Smolensky, 1993:67-82):

(3) a. The prominence scale of onsets

Obstruent >> Nasal >> Liquid >> (Vowel)

b. The prominence scale of nuclei

Vowel >> Liquid >> Nasal >> Obstruent

c. The prominence scale of codas

(Vowel) >> Liquid >> Nasal >> Obstruent

In other words, the onset will be most prominent when the segment that is mapped into the onset position is an obstruent (3a), and the coda will be the most prominent when the segment in the coda position is a liquid (3c). Note the scales here are a mirror image of the sonority scale.The challenge in this factor will be merging all three of these scales into a single quantifiable scale, if such a merger is at all possible.

In the same manner, we can derive prominence scales of clusters based on the sonority distance between the segments that form the clusters, using the Sonority Dispersion Principle (= SDP; Clements 1990), which states that the greater the sonority distance between two segments is, the better the sequence is, and the less marked it is. However, the directionality of this distance also matters. In onset clusters we prefer the

first consonant to be less sonorant than the second, a principle called the Sonority Sequencing Generalization (= Sonority Sequencing Generalization, SSG; Selkirk, 1980). At the other edge of the syllable, we also prefer a coda to be more sonorant than the following onset (= Syllable Contact Law, SCL; Muraay and Vennemann, 1983). See (4) for an illustration.

(4) a. The prominence scale of onset clusters (O=Obstruent, N=Nasal, L=Liquid)

O/L >> O/N, N/L >> L/L, N/N, O/O >> N/O, L/N >> L/O

Sonority rise Sonority plateau Sonority fall

b. The prominence scale of C.C sequences

L.O >> N.O, L.N >> L.L, N.N, O.O >> O.N, N.L >> O.L

SCL preservation SCL violation

2.2.3 Prosodic Similarity

I have not yet found confusability rates of prosodic rhythms or a scale of stress positions' prominence. However, there are some separate scales we know of that could be merged together into one scale (see 5).

(5) The prominence scales of stress and position (Gordon and Roettger 2017; Cooper 1983)

Unstressed syllable >> Secondary stress >> Primary stress

Final syllable >> Final stressed syllable

2.3 Pitch Accent, Tonal and Intonational Languages

Besides stress, other prosodic prominence systems exist, e.g. pitch accent (Ito and Kenstowicz 2017 on Japanese), tone (Hyman 1977 and de Lacy 2002 on Mandarin) and intonation (Jun 2005 on Korean). These prosodic patterns must be placed on the universal similarity scale. There are languages, such as Japanese and Romanian, which are

distinguished primarily by this factor (out of the three factors mentioned above). Japanese is a pitch-accent language while Romanian has a stress pattern.

Regarding intonation, stress and intonation rely on similar acoustic cues: both are characterized by higher pitch (F0) and intensity rates, as well as a longer duration, relatively to unstressed syllables and non-intonational words (Fry, 1955 and Jun, 2005); the difference between them is that stress refers to syllable prominence in a word, while intonation refers to word prominence within an utterance. Many researchers (e.g. Beckman 1986, Jassem 1959 and Fry 1958) claim that when observing the prominence of pitch, duration and intensity in languages, the most acoustically prominent factor is pitch, and the least prominent factor is intensity, though this scale might change depending on the observed language.

In comparison to stress and intonation, pitch accent and tones are characterized *only* by a pitch (F0) change. However, the pitch change may affect other features. e.g. contour tones may lengthen vowels (Remijsen, 2003). As an example, we can look at the Mandarin Chinese words /mā/ *'mother'*, /má/ *'hemp'*, /mǎ/ *'horse'* and /mà/ *'scold'*. The meanings of these words change according to their tones: in the first word F0 is high and steady; in the second word, F0 rises, in the third word, F0 falls, then rises (this tone is also pronounced longer than others); and in the last word, F0 falls.

3. Hypothesis and Research Question

3.1 Research Question

The goal of this study is to be able to predict how a speaker of some language with certain properties will quantify the similarity of two other language unknown to him, in relation to all other languages. However, as mentioned above, I found no study that provides one, unified and universal scale of similarity for all of these properties. In fact, we do not even know yet which phonological and phonetic features we need to consider to quantify similarity.

In this thesis, I focused on finding these features and examining whether they can help distinguish between languages. In this study, all features examined are assigned the same weight, even though it might not be the case, as some features may be more salient than others, i.e. have a larger effect on similarity quantification.

3.2 Hypothesis

As was written above, since we still do not know what the prominent properties are, this experiment observed the properties 'tabula rasa', i.e., all properties in this experiment were assumed to have the same prominence. Therefore, the hypothesis of this thesis was that we can define similarity between languages based solely on the acoustical measures of some phonetic and phonological properties.

3.3 Methodology

This thesis consists of two experiments: a preliminary and a main experiment. The preliminary experiment rated the familiarity of Hebrew speakers with the languages that appeared in the main experiment, and the main experiment collected data on language identification of Hebrew speakers. The experiment ran online and was available for every Hebrew speaker via the internet, as the goal was to collect as many subjects as possible.

The subjects were given three different three recordings of different languages in each trial, a base language and two additional languages, and were asked which of the two additional languages is the most similar to the base language. Some of the languages, according to the preliminary experiment, were familiar to Hebrew speakers (e.g., French and Russian) and some were unfamiliar (e.g., Hausa and Fijian). This methodology has been used before (e.g., the Great Language Game; see §4 and Skirgård et al., 2017).

4. The Great Language Game

Before explaining the experiments I conducted within the scope of this thesis, I would like to briefly present the experiment I based my own study on – The Great Language Game (Skirgård et al., 2017), which was firstly published in 2013 and ran for nearly five years, collecting data from a great number of speakers from various countries. Note that I will only present things relevant to my study.

4.1 Research Questions and Predictions

There were a few goals for this game: a) to determine which languages are confused with each other; b) to determine whether there are any asymmetries of confusion between languages (i.e., if you hear language A and choose language B as being more similar to language A than the other options available, will you choose language B as more similar when hearing language A?); c) to provide the factors that can predict whether players confuse two languages for each other (see (6.a)); d) to provide the factors that can predict player's accuracy of the answer they give (see (6.b)); e) to examine whether the accuracy of the answer can be predicted by linguistic or non-linguistic factors; and f) to determine whether the importance of phonological cues surpasses the importance of non-phonological cues in predicting the player's accuracy.

There were also a few predictions in this research: a) players will differentiate languages based on phonological properties (e.g. the appearance of retroflex consonants in a language), while some features might be more salient than others (e.g. the appearance of trill rhotics is more salient than the appearance of labiodental fricatives), thus they might have more influence on the confusion between languages; b) the more shared lexical items between the languages, the more they will be confused foreach other; c) languages with more speakers will be easier to recognize and differentiate from other

languages; and d) the clearer the recordings are, the better the differentiation between languages will be.

The factors Skirgård et al. (2017) examined were divided into two categories: factors that can predict the confusion between languages and factors that can predict the accuracy of the answers. See the factors divided by these categories in (6).

(6) a. Factors that can predict the confusion between languages

- Geographical closeness.
- Genealogy.
- Similarity of phoneme inventory.
- Lexical similarity.

b. Factors that can predict the accuracy of the answers

- Acoustic quality of the speech samples: measures the range of frequencies in a signal.
- Proportion of non-native speakers (L2 speakers): measures the number of L2 speakers divided by the sum of L1 and L2 speakers.
- Total native speaker (L1) population: is taken from Ethnologue (Lewis et al., 2014).
- Linguistic diversity of the main country in which the language is spoken: measured by the Greenberg Diversity Index (GDI) from the Ethnologue (Lewis et al., 2014), and reflect the probability of two people from the same country speaking the same first language.
- Number of countries the language is spoken in: is taken from Ethnologue (Lewis et al., 2014).
- Language name transparency: measured by whether the name of the language has a transparent link to the main country in which it is spoken (e.g. Spanish is spoken in Spain so the name of the language is transparent, but Urdu is spoken in Pakistan so the name of the language in not transparent).
- Economic power of main country: measured by the Gross Domestic Product of the main country in which the language in spoken.
- The frequency of occurrence of the language name in Google Books in English

texts, and the Mandarin name of the language in Chinese texts.

4.2 Methodology

4.2.1 Participants

The game was uploaded to the internet in English, thus providing the option for every English speaker, regardless of their level of fluency in English, to participate in the game. Approximately 15 million responses were gathered from participants from all over the world. Nothing is known about these participants, except that they knew English well enough to participate in the game, that they were computer-literate, and that they had some interest in languages. In addition, the IP addresses of the participants were collected, thus we know which country they participated from.

4.2.2 Stimuli

A total of 78 languages were presented in the game. Thirty-nine of these languages were Indo-European and others were from various other families. Each language was represented by a 20 second audio-clip of natural speech, taken from broadcasts. After deciding which languages would be shown in the game, the phonemic inventories were taken from the Phonetic Information Base and Lexicon-database (= PHOIBLE).

4.2.3 Procedure

The participants were presented with an audio-clip of some given language and their goal was to determine which language they had heard. First, they were given four possible answers. After each question they answered, the participants were informed of whether their answer was correct, and if they were wrong, the right answer was presented. Should the participants answer correctly three times, the number of possible answers was increased by one, up to ten possible answers. If the participants were wrong in three questions, the game was over. The participants could participate in the game as many

times as they wanted.

4.3 Results and Discussion

The results showed that there was a 70% probability of guessing a language correctly. Some pairs of languages were confused a lot (e.g., Punjabi and Kannada), while other pairs were rarely confused with one another (e.g., French and Vietnamese). It was found that similarity was not symmetrical: for example, every Slavic language was confused with Russian, but Russian was rarely confused with other Slavic languages. Skirgård et al. (2017) found out that many non-linguistic factors might predict the confusion between languages: historical relations between the languages, geographical relations between languages and cultural knowledge. In addition, languages with very different phonemic inventories (consonants or vowels) are less likely to be confused with one another.

Most of the recognizable languages were from Europe, while the least recognizable languages were from Latin America (and were only spoken in Latin America). There was also one factor that could significantly predict the accuracy of the answers: the "global fame" of the language, i.e., how many times its name appeared in Google search, the economic power of the country in which the language is spoken, and so on. Skirgård et al. (2017) also noted that languages which differed in the presence or absence of some salient phonological properties were less confused with one another (for example, the presence or absence of labial affricates, retroflexes and more).

4.4 The Issues

The Great Language Game provided a vast database on the similarity of languages; it gathered an impressive number of participants from all over the world, and its results can be used in many follow-up researches. However, The Great Language Game was conducted on socio-linguistic grounds, rather than phonological ones: the factors

examined in Skirgård et al. (2017) were factors concerning the history, geography and economy of the countries in which these languages are spoken, and there was minimal reference to phonological properties other than the phonemic inventories of the languages. The participants heard one language and had to choose a name of a language as an answer, without hearing the languages that appeared as answers, thus many unfamiliar languages (mostly not Indo-European languages) could not be chosen answers based on phonology, since no one knows how some of the unfamiliar languages really sound (e.g., does any non-linguist know how Kannada sounds, except for Kannada speakers who livein the southwestern region of India?).

Another issue in this game regards the data gathered from the participants: the researchers only knew the IP address from which the participants played this game. In other words, they did not know the participants' L1 (especially participants from countries with many languages, such as India), they did not know their age (which could affect the participants' level of language knowledge, as well as their phonemic inventory), they did not know which other languages the participants knew (this could affect the answers of the participants, because if they were familiar with some language they could recognize it), they did not know whether the participants lived in the country from which they played, or perhaps they only visited there, and more.

The final issue I would like to mention is the phonemic inventory of the languages which appeared in the game. Since the participants only heard 20 seconds of some language, it is very likely that they did not hear the entire phonemic inventory of the language as it appeared on PHOIBLE, thus the variable of phonemic inventory in this study might be a bit skewed towards the more unmarked segments, and thus there was not enough phonemic contrast to distinguish between languages. For example, if some language has

retroflexes according to its phonemic inventory, it does not necessarily guarantee that retroflexes appeared in the recording, thus they surely could not distinguish between this language and other languages which have no retroflexes.

In conclusion, The Great Language Game was a great experiment which can be used for many sociolinguistic experiments. But I think that since we want to understand how speakers distinguish between languages *phonemically*, we will need to control the experiment further: gather some more data on the participants, choose the recordings wisely so that they will fully represent the phonemic inventory of the languages, and design an experiment in which the participants will not be required to recognize the languages basedon their name alone.

5. The Experiments

The Great Language Game provides a great background for building other experiments. As explained above, the major concern regarding The Great Language Game's experiment is that the recognition of languages was not entirely linguistic (and more specifically, phonological), but it used some other knowledge, e.g., cultural knowledge. In addition, the participants had to choose the *name* of the language they had heard out of a limited number of given options, thus they might not have compared between two languages per se (i.e., discrimination task), but rather they tried to recognize the language they were hearing (i.e., recognition task).

The Hebrew version of the game created by us tries to overcome this issue by asking participants to choose the *recording* they thought was the most similar to the recording presented in the question. This way, by not presenting the name of the languages the participants were hearing, many of the non-phonological factors examined in The Great Language Game, e.g., the language name transparency, were not considered as factors in our version of the game, and the participants only had to use their phonological knowledge to differentiate between languages. In other words, some of the confounds were neutralized in this experiment.

5.1 Preliminary: Language Familiarity Scale

Before conducting the main experiment, and after determining which languages would be presented in it (see §5.2.2), we wanted to determine the level of familiarity of each language for Hebrew speakers. The reason for this is that the familiarity of languages may affect the results of the main experiment: should speakers of some languages hear one very familiar language and one unfamiliar language, they might tend to choose the unfamiliar language to be similar to the language they need to compare them to, because they

"know" the other language, and it is dissimilar to the other one. Therefore, we conducted a preliminary questionnaire to determine this issue.

5.1.1 Participants

The questionnaire (in Hebrew) was created as a Google Form (see the questionnaire here) and was passed on to the participants digitally. A total of 132 participants answered the questionnaire. Most of the participants wrote that they knew English, but since English is not a language participating in the experiment it did not matter here. Eighty-one of the participants (61%) knew other languages (e.g., Russian, Spanish, Ukrainian, German and more). Thirty-five participants (26.5%) had some knowledge in linguistics. The participants had been living in Israel for at least a decade.

5.1.2 Materials

The questionnaire contained a total of 35 languages (see §5.2.2.1 for elaboration of the languages) *in written form* in Hebrew, i.e., the names of the languages appeared in the questionnaire. Hebrew did not appear in the questionnaire, even though it did appear in the main experiment, since Hebrew speakers should know Hebrew.

5.1.3 Procedure

The participants were gathered from Facebook groups and friends who passed the questionnaire on. In the questionnaire, we asked the participants to rate their familiarity with the given languages on a scale of 1-5 (1- unfamiliar, 5-very familiar). The participants could take their time answering it, and it took less than five minutes to fill out.

5.1.4 Results

The final ratings of the participants are presented in table (7).

(7) The familiarity ratings of languages by Hebrew speakers (N= 132)

Russian	3.72
Spanish	3.64
French	3.59
German	3.39
Italian	3.27
Yiddish	3.14
Ukrainian	2.79
Japanese	2.51
Amharic	2.32
Portuguese	2.32
Egyptian Arabic	2.17
Mandarin	2.15
Polish	2.14
Hindi	2.03
Turkish	1.95
Persian	1.91
Korean	1.79
Bulgarian	1.74
Swedish	1.69
Czech	1.64
Hungarian	1.64
Thai	1.58
Finnish	1.48
Norwegian	1.47
Slovak	1.40
Vietnamese	1.40
Croatian	1.38
Xhosa	1.16
Telugu	1.08
Pashto	1.05
Somali	1.05
Yoruba	1.05
Hausa	1.04
Fijian	1.03
Oriya	1.02

Unsurprisingly, most of the languages rated as most familiar were Indo-European languages (e.g., Russian, Spanish, French and German) and the languages rated as the least familiar were "exotic" languages (e.g., Xhosa, Telugu and Somali). Interestingly, Indo-Iranian languages (e.g., Pashto and Oriya; except for Hindi), which are languages in a

sub-family of the Indo-European languages, were rated as unfamiliar to almost all participants. In addition, the "Asiatic" languages (e.g. Japanese and Mandarin) were rated as relatively familiar.

Keeping these results in mind, let us proceed to the main experiment. Further discussion on the familiarity with languages will be discussed in §6.

5.2 The Main Experiment: The Hebrew Great Language Game

After determining the level of familiarity of Hebrew speakers with the languages in the main experiment, we had enough data to build the experiment and form the questions. The main experiment was similar in its design to the experiment conducted in Skirgård et al. (2017), but the questions in the current experiment were based on hearing both the base language (the language which appeared in the question itself) and the languages that could be possible answers. The names of the languages did not appear. In addition, the languages in the questions were not randomly selected but carefully chosen based on their similarity percentage (using the proposed model) to the base language. Should the model proposed in this study work, there will be an inclination of the participants to choose the more similar language over the dissimilar language. If this is indeed the case, then similarity can be quantified by phonetic and phonological features.

5.2.1 Participants

Our goal was to pass the experiment on to as many participants as possible, in order to overcome known confounds such as type I error and variance between speakers. We gathered a total of 362 participants, most of them speakers of solely Hebrew plus English. A hundred and eighty-nine (53%) participants spoke another language/s, e.g., Russian, Spanish, Portuguese and Arabic (M = 2.81, SD = 1.14). A hundred and twenty-four (34%) of the participants had some knowledge in linguistics, but only a few had some advanced

academic linguistic knowledge. All participants were living in Israel. Eventually, each question was answered by at least 23 and at most 77 participants (M = 39.6, SD = 11.01).

5.2.2 Materials

5.2.2.1 The Languages. Thirty-six languages were examined in this experiment: 19 (52.8%) Indo-European languages, consisting of seven (19.4%) Balto-Slavic, four (11.1%) Germanic, four (11.1%) Indo-Iranian and four (11.1%) Italic languages; five (13.9%) Afro-Asiatic languages; two (5.6%) Niger-Congo and two (5.6%) Uralic languages; and the remaining languages (22.1%) were the sole representative of their linguistic family. The languages were chosen so that each had at least one clear recording (i.e., with no background noises).

A total of 64 audio recordings were shown in the experiment: 26 (72.2%) languages were represented by two recordings: one with a male speaker and one with a female speaker; seven (19.4%) languages were represented only by a recording with a male speaker (because a recording with a female speaker was not found for these languages), two (5.6%) languages were represented only by a recording with a female speaker (because a recording with a male speaker was not found for these languages), and one (2.8%) language, Korean, was represented with one recording with a female speaker and two recordings with a male speaker. See Appendix A for the full list of all languages, the family they come from, and how many recordings each language was represented with. The phonological data of these languages were put into one table, so that each segment and prosodic property appearing in these languages was separated for later use (see §5.2.2.2).

5.2.2.2 The Recordings. As was mentioned above, a total of 64 audio recordings were presented to the participants. The recordings were extracted from radio broadcasts downloaded from SBS radio by PRAAT (Boersma and Weenink, 2009), so that each recording was between 3.3-5.2 seconds (M = 4.26, SD = 0.43). The length of the recordings did not exceed two SD above or below the total average length of the recordings. We paid particular attention to avoid possible loanwords (to Hebrew) in the recordings, so as not to indicate what type of language is spoken in a recording.

In addition, there was mostly a balance between the gender of the speakers in the recordings (with a few exceptions), and even though the age of the speakers could not be precisely determined, they did not sound like children or elderly. Since the recordings were taken from broadcasts, they were (almost) "clean", i.e., with no background or white noises.

Before analyzing the recordings, we gathered phonological and general data regarding the languages we wanted to analyze – the family they came from, their consonant and vowel inventories, their phonotactics and their prosody. The data were taken from the Wikipedia pages of the observed languages. Then, each recording was analyzed by transcribing the consonants, the vowels and the syllables in it. The data were then transferred into one Excel file, so that we could observe all the data of the languages together. Later, the single segments were merged into natural phonological classes, separated by recording and by language (i.e., all recordings of the language, independently and together). See an elaboration on the natural classes and other phonological data observed in §5.2.2.3.

Not surprisingly, we found out that not all the segments which appear in the language according to Wikipedia really appeared in the recordings, and not every

segment appeared in both recordings of the language. Therefore, we decided to treat each recording as a different unit, or a different language, for the purpose of this experiment. This is because each recording contained a group of segments which was different, however slightly, from the other groups in the other recordings.

5.2.2.3 The Phonological Properties Examined. After gathering the data from the recordings, we could try to generalize the data and compare it cross-linguistically. In order to do that, we needed to create natural classes of the segments, and measure the similarity of the languages by the number of shared natural classes in two given recordings. Recall that this experiment will observe the properties 'tabula rasa', i.e., all properties in this experiment will be considered as having the same prominence, regardless of previous research, to establish first that phonology has a role in similarity quantification.

However, not *all* natural classes were examined in here. For example, it is very unlikely that the appearance of the most unmarked stops (e.g., /t/) will differentiate between languages because they appear in almost every language existing, and even if they do not appear in a given recording, it is unlikely that anyone will notice their absence. On the other hand, retroflex stops are more marked than alveolar stops, therefore if they appear in languages, they might influence the differentiation between languages.

Some of the factors considered in this experiment were taken from The World Atlas of Language Structures Online (= WALS; Dryer and Haspelmath, 2013), and others were taken from our own knowledge of phonological properties in languages. This section will be divided into two subsections: one will list the factors taken from WALS and will elaborate on them and the second will list the additional factors we added by ourselves. A total of 41 factors were examined in this experiment, 15 of them were taken from WALS

and the others were factors were added by us.

5.2.2.3.1 Phonological Properties from WALS. WALS is a large database of the structural properties of 565 languages. It was gathered from descriptive materials by 55 authors. This site contains 19 chapters about phonology, and in each chapter, there is a description of some property and a distribution of the property among the described languages.

In (8), I will list the factors from WALS that we observed in our experiment, and I will elaborate on the measurement of some of the relevant factors. You can see the full distribution of these properties in the world languages, as well as in our own recordings, in Appendix B.

(8) Phonological properties taken from WALS

a. Segmental properties

- Consonant inventory.
- Vowel inventory.
- Voicing contrast in obstruent consonants.
- The appearance/absence of uvular consonants.
- The appearance/absence of ejective consonants.
- The appearance/absence of implosive consonants.
- The appearance/absence of glottalized consonants.
- The appearance/absence of lateral consonants.
- The appearance/absence of clicks.
- The appearance/absence of inter-dental fricative consonants.
- The appearance/absence of pharyngeal consonants.
- The appearance/absence of front rounded vowels.

b. Prosodic properties

- *Consonant-Vowel Qualities ratio (=C:VQ ratio):* a ratio set by dividing the size of the consonantal inventory by the size of the vowel inventory.

- *Syllable structure:* set by whether the language contains very complex syllables (i.e., a sequence of five consonants), moderately complex syllables (i.e., a sequence of 3-4 consonants), or no complex syllables at all (i.e., a sequence of 1-2 consonants, where at most one consonant is in the onset position and at most one consonant is in the coda position). This division should include all prosodic cases (i.e., languages with no codas at all, languages with codas and languages with complex onset and/or coda).
- *Tone:* set by whether the language has no tones, simple tones or complex tones.

5.2.2.3.2 Other Phonological Properties Examined. In addition to the properties described in WALS, there were few more factors we wanted to examine. Some of these factors were taken from Bradlow et al. (2010) and others were added by our own intuition on similarity between languages upon hearing the recordings. See (9) for the list of factors we added, and the elaboration on some of these factors.

(9) Additional phonological properties examined in the experiment

a. Segmental properties- consonants

- *The existence of non word-initial glottal stops:* as opposed to word-initial glottal stops, which appear either phonemically or allophonically in many languages, the existence of glottal stops in other positions is not as common. Therefore, it might be that their existence may help differentiate between languages.
- *The appearance/absence of non-strident fricatives:* as strident fricatives appeared in 100% of our recordings, the factor of the appearance of stridents was not considered because it could not differentiate between our recordings. However, not all recordings, and not all languages, have non-strident fricatives.
- The appearance/absence of glottal fricative consonants.
- The appearance/absence of rhotic consonants.
- The appearance/absence of glide consonants.
- The appearance/absence of glide consonants with two places of articulation.
- The appearance/absence of retroflex consonants.
- The appearance/absence of affricate consonants.

- The appearance/absence of palatal consonants.
- The appearance/absence of palatalized consonants.
- The appearance/absence of prenasalized consonants.
- The appearance/absence of aspirated consonants.
- The appearance/absence of breathy-voiced consonants.
- The appearance/absence of unreleased consonants.
- The appearance/absence of labio-dental fricatives.

b. Segmental properties- vowels

- The appearance/absence of back unrounded and central vowels.
- The appearance/absence of long vowels.
- The appearance/absence of nasalized vowels.
- The appearance/absence of high vowels.
- The appearance/absence of back vowels.
- The appearance/absence of round vowels.
- The appearance/absence of [-ATR] vowels.
- The appearance/absence of low vowels, except /a/.

c. Prosodic properties

- *Syllables per second (=SPS):* set by the number of syllables in a recording, divided by the length of the recording.
- *Consonant-vowel ratio (=C:V ratio):* as opposed to the *consonant-vowel qualities ratio* factor described in the previous sub-section, which is set by the size of the consonant inventory divided by the size of the vowel inventory, the *consonant-vowel ratio* factor is set by the total number of consonants in the recording (even if the same consonant appeared several times) divided by the number of vowels in the recording (even if the same vowel appeared several times). This ratio is less diverse than the *consonant-vowel qualities ratio*, whose score can be between 2-6.5.
- The appearance/absence of geminates.

5.2.2.4 The Similarity. As I mentioned above, there were a total of 41 factors considered in this experiment. Some of them had binary values (if there was an appearance or absence of a property in two given recordings, the value 1 was given to both recordings in the relevant factors and if one recording contained this property and the other recording was not, the value 0 was given to both recordings), but some factors (mostly prosodic factors, e.g., SPS) could not be given an immediate set value of 0 or 1 because we divided numbers and created ratios of them. However, if we want to put all factors on one scale, our values of all factors must be the same ones. Therefore, we needed to find a way to set a value for these ratios. In order to do that, we used standard deviations. If, given two languages, the distance between the scores of a given factor is smaller than one SD, then these two languages will get the value 1, otherwise the value will be 0. For example, the SPS of the female recording of Bulgarian is 5.04 and the SPS of the female recording of Czech is 6.86. Since the SD of the SPS factor is 0.77 and the distance between the SPS of both recordings is 6.86-5.04 = 1.82, then both of these recordings will get the value 0 in this factor. Note that the mean score of the factor is not considered in the calculation, and only the scores of the two given recordings are considered here, therefore the value they get is relatively to each other, and not to all recordings. See (10) for the list of the non-binary factors and their statistics, especially the SD.

(10) Non-binary factors' average and standard deviation (=SD)

Factor	*Mean*	*SD*
Consonant inventory	13.63	3.10
Vowel inventory	4.39	1.64
Consonant-vowel ratio	1.19	0.196
Consonant-vowel qualities ratio	3.43	1.24
Syllables-per-second (SPS)	5.61	0.77

Now, after all of our factors have a value of 0 or 1, we can calculate the similarity of languages (or, in this experiment, the recordings – recall that since almost every language has two recordings, and since each recording might possess segments that the other recordings do not, we consider all recordings, even recordings of the same language, as a different "language"). We calculated the sum of the values (recall that if a factor is similar in both languages it will get 1, otherwise it will get 0) and divided it by the number of the examined factors – i.e., 41. The outcome will be the percentage of similarity between two given recordings, and both recordings will get the same similarity percentage relatively to each other, as they have the same values relatively to each other. The mean of the total similarity of all recordings was 71%, and the range of similarities was between 44%-93%. The final similarity percentages per recording are presented in Appendix C.

5.2.2.5 The Questions. Now we can proceed to the final stage of the methodology – choosing the recordings that will be presented in the questions in the experiment. Each recording appeared as the question (=target) three times and the question had two possible answers with different similarity percentages, i.e., we needed to find three sets of recordings so that one will be very similar to the target recording and the second one will not be similar to it. However, we needed to control for a few things when choosing the recordings that will be shown along with the target: first, we needed to avoid gender confounds; therefore, the target recordings always differed from the answer recordings in gender.

Second, we needed to control the gap between the similarity percentages of each language, since we could create a confound in which an answer is chosen because the other language is very different from both the target and the other answer; this was done by making sure that no gap was greater than 2 SD from the total mean of all recordings (M = 20.2, SD = 2.5).

Lastly, we needed to ensure that all the recordings appeared more of less the same number of times, since some recordings were very similar to more languages than others. The appearance of a certain recording more times than another recording might affect answers chosen. This was done by counting the number of appearances of each recording and replacing it with another recording if the number of its appearances was higher than 2 SD from the total mean, or add the recording to more questions if the number of its appearances was lower than 2 SD from the total mean (M = 6, SD = 2.4).

A total of 182 questions were prepared for the experiment (see Appendix D for the full list of the questions). Some of the questions were symmetrical, i.e., a target recording appeared as a possible answer for one of its own possible answers (for example,

in one of the questions with Amharic as a target, a possible answer was Japanese, and in one of the questions with Japanese as a target, a possible answer was Amharic), and some of the questions could potentially show transitivity, i.e., when a target in which a possible answer is a target to a different possible answer, it might be that the last possible answer could be a possible answer to the first target (for example, in one of the questions with Hausa as a target, a possible answer was Oriya, and in one ofthe questions with Oriya as a target, a possible answer was Japanese. A transitive question will be a question where the target is Hausa and a possible answer is Japanese).

5.2.2.6 The Game. The experiment was uploaded to a website prepared for the purpose of the experiment's publication. The purpose of the website was to spread the experiment to as many participants as possible, and to gather a large database of possibly thousands of participants. The reason for this was to avoid as many potential confounds as possible, confounds which may have appeared in a smaller experiment.

The experiment consisted of one session of 20 questions, randomly selected by the computer. The possible answers in each question were chosen by us (see previous sub-section), but the order of the answers was also randomized. Each question appeared separately from the others, but the participants could move forward and backward if they wanted to change an answer they had already given, or to skip a question and come back to it later. The text in the question was the same in all the questions – "which of these two languages is the most similar to the first language?", and all they could see after this text was three audio recordings. See (11) for an illustration.

(11) An illustration of a question in the experiment

After finishing the experiment, the participants were informed of their score in the session and their place in the leaderboard. Since there are no right or wrong answers, the answers with the higher similarity percentage were considered as right, and every participant who chose that answer got one point (contrary to participants who chose the answers with the lower similarity percentage, that were considered wrong and provided zero points).

5.2.3 Procedure

The participants were gathered from Facebook groups and via friends on social media.The goal of the experiment was not revealed to them, not even after they finished it. The participants were only aware that the experiment was in fact a game, about comparing languages. In addition, they knew that participants that get the highest number of points (i.e., answer “right” on the highest number of questions) will get into the leaderboard.

Before beginning the game, the participants needed to write down the languages they speak and whether they have linguistic knowledge. Only after answering these questions were they able to proceed to the game itself.

After finishing the session, the participants were informed of the number of right answers they got and their rank in the leaderboard. If they wanted to improve their score,

they could try another session. Each session took approximately 15 minutes. The data of the participants was saved in the database of the site and could be easily extracted and analyzed.

5.2.4 Results

The full results of the main experiment are presented in Appendix E. Each question was first analyzed separately using the binomial distribution test, in order to examine which answer, if at all, was chosen the most in this question alone. Overall, in 102 questions (56%), the participants significantly chose the similar answer more than the dissimilar answer, in 35 questions (19.2%) the participants significantly chose the dissimilar answer more than the similar answer, and in the remaining 45 questions (24.7%) the participants chose both answers equally. There were significantly more similar answers chosen than both dissimilar answers and answers with equal choosing ($\chi^2 = 43.109$, $p < .001$). The significantly similar answers ranged between 61.2%-100%, while the significantly dissimilar answers ranged between 61.4%-90.6%. In three of the questions, the similar answer was always chosen – one question of the male recording of Croatian with Polish and Thai as possible answers (Polish was always chosen), another question of the male recording of Czech with Slovak and Thai as possible answers (Slovak was always chosen), and a question of the male recording of Ukrainian with Slovak and Oriya as possible answers (Slovak was always chosen).

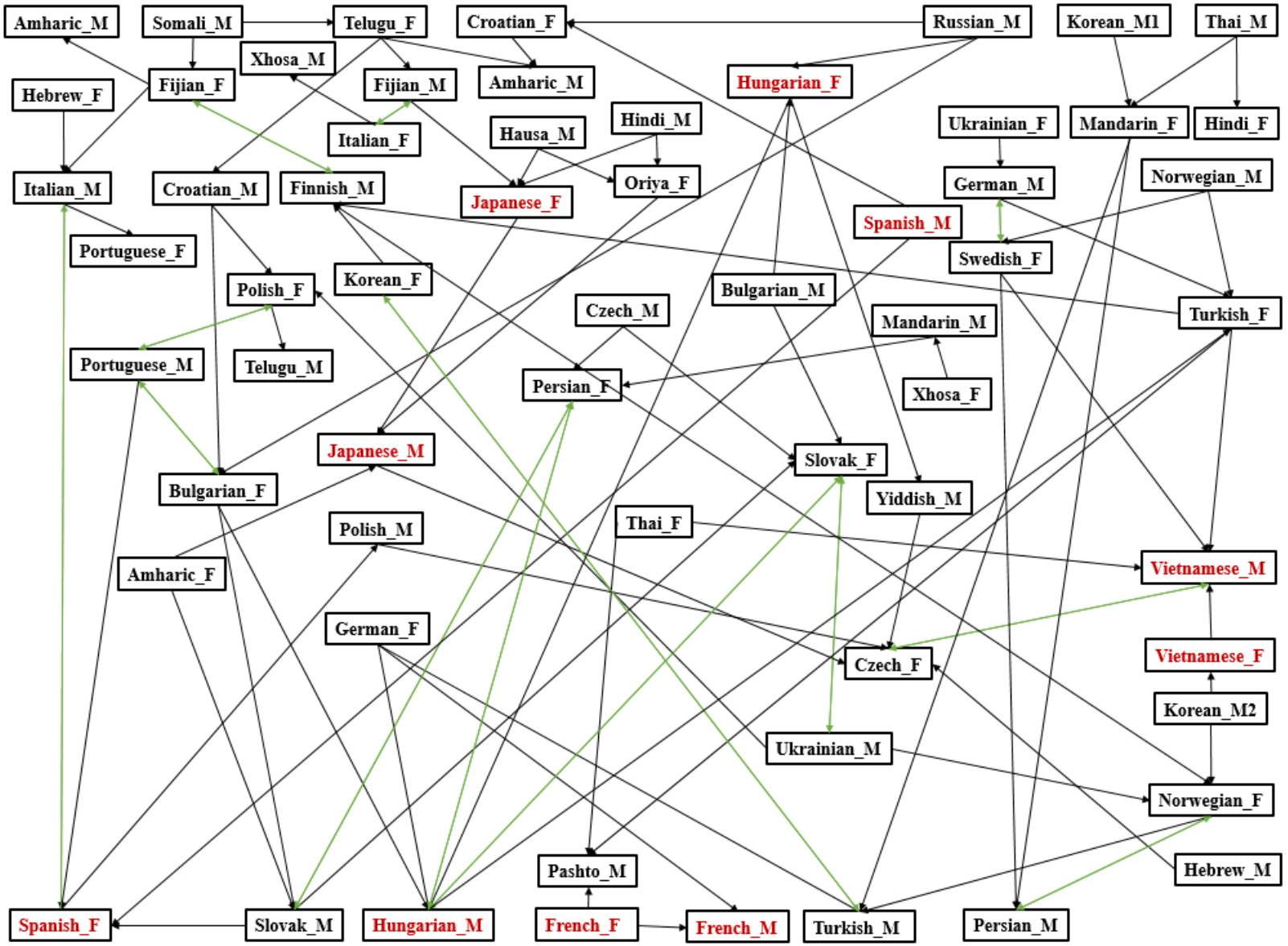
When observing the significant answers per language, we can see that some languages tend to be chosen as answers more times than others (see Appendix F). For example, Slovak was chosen as an answer when it appeared as the similar answer in eight out of eight times (100%), yet Italian was chosen as an answer only in four out of 11 times (34.6%) and Portuguese was chosen only in three out of 12 times (25%). Since

there were not many significant dissimilar answers, languages which were put in the questions as dissimilar answers were not chosen frequently, but some were still sometimes chosen, such as Hungarian in three out of eight times (37.5%), while some were never chosen, such as German in zero out of 11 times.

5.3 General Discussion

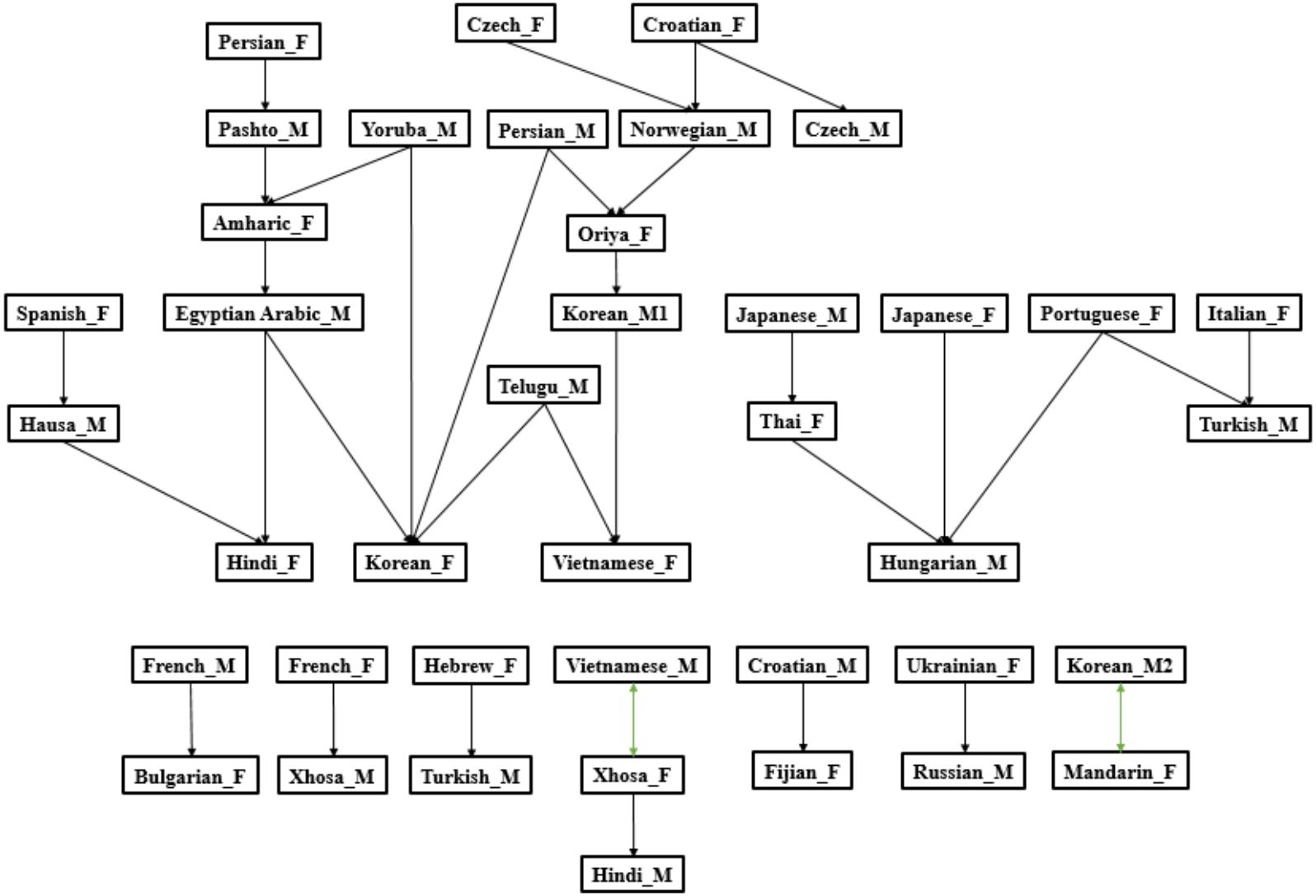
Overall, similar answers were chosen more than dissimilar answers, even if we include questions in which both answers were chosen equally. In addition, we see that some languages are observed by speakers as more similar to other languages than others (e.g., Slovak), while other languages are observed as less similar to other languages than others (e.g., German). Thus, it appears that the model built in this thesis works to some extent, such that the more phonological features two languages share, the more similar speakers will perceive them to be. See (12) for all significant similarities between the languages and (13) for all significant dissimilarities.

(12) A chart of similar significant languages



Language_M – a male recording of the language, *language_F* – a female recording of the language; language in the question → language chosen as the answer; *green line* = symmetrical relation; *red* = languages in which the language in the question and the language in the answer were the same language.

(13) A chart of dissimilar significant languages



Language_M – a male recording of the language, *language_F* – a female recording of the language; language in the question → language chosen as the answer; green line = symmetrical relation.

It is important to note that when some languages appeared as the question, all three answers were significant (whether it was the similar language that was chosen or the dissimilar one) but when other languages appeared as the question, all three answers were insignificant. In other words, some languages were found as similar to three other languages, while other languages were not found as similar to other languages at all. For example, the female recording of French had three significantly similar languages, while the male recording of Amharic had no significantly similar languages at all, which indicates that the former language was similar to other languages more than the latter language. This outcome may suggest that some features of Amharic might feel intuitively different to speakers than other features of that does or does not appear in French. i.e., it might be that some features, when they appear, influence similarity more than other features.

The case of French is specifically interesting to observe, because Skirgård et al. (2017) found that French was the most recognizable language – thus, it would have been logical to assume that no language will be chosen as being similar to French (the same as if Hebrew speakers were asked what language is similar to Hebrew – since Hebrew speakers know Hebrew fluently, they will probably not think that there is a language which is similar to Hebrew). Therefore, the fact that the female recording was found significantly similar to three other recordings (when this recording appeared in the question) may suggest otherwise. However, when we observe the times in which French was chosen as more similar when it appeared as an answer (i.e., when it was needed to be chosen by the participants instead of just passively appearing in the question), French (both recordings) was chosen in two out of 14 times (14.3%). Thus, the current results in this study seem to correspond with the results presented in Skirgård et al. (2017)

regarding French.

5.3.1 Symmetry

The current model shows symmetrical relations in many cases, in both the similar and dissimilar languages. For example, in (12) we see that the male recording of Fijian was found as similar to the female recording of Italian, and vice versa. Another example can be taken from (13), in which the male recording of Vietnamese was found as similar to the female recording of Xhosa and vice versa. These results somewhat contradict Skirgård et al. (2017), who suggested that similarity was not symmetrical. Since the current model is based on the number of shared phonological properties between two languages, it is logical to assume that the similarity percentage will be the same whether the language in the question is language A or language B. However, it is important to note that not in all cases of potential symmetry there was symmetry. For example, the female recording of Telugu was found as similar to the male recording of Amharic, but this similarity was not found the other way around. This asymmetry could have been caused because of statistical reasons, because the third language shared some salient features with the first language, or because there really was not symmetry between these languages; thus this asymmetry might be explained by feature weighing (see §5.6).

5.3.2 Transitivity

In addition to symmetry, the model also shows some transitive relations between languages, though this transitivity could be found only per language and not per recording (since the languages in the questions and the languages in the answers were of different gender. e.g., if language_A_F → language_B_M → language_C_F, we would not find a question in which language_A_F → language_C_F since both recordings are of female speaker). For example, Somali was found as similar to Telugu, and Telugu was

found as similar to Fijian. Then, Somali was found as similar to Fijian, thus showing the transitivity relation. However, unlike the symmetrical relation in which two languages share the same number of phonological features and have same similarity percentage in relation to each other, in transitivity the case is different since the shared features between language A and language B will not necessarily be the shared features between language B and language C, therefore it is not necessarily true that language A will also be similar to language C. Indeed, we see few cases in which transitivity is not shown: for example, even though Croatian is similar to Amharic and Amharic is similar to Japanese, Croatian was not found as similar to Japanese. This result might have been explained by the number of features: it might be that the features that Croatian and Amharic share are not the same features that Amharic and Japanese share, such that the number of features Croatian and Japanese share is low. For example, Croatian and Amharic could have shared features A and B and Amharic and Japanese could have shared features C and D, so Croatian and Japanese do not share any of these four features. However, according to the model, Croatian and Japanese have 78% of similarity, which is considered a high percentage, therefore this explanation cannot be accounted for with this result. Rather, in order to explain this result, we might need to observe the shared features between Croatian and Japanese and consider their weighting on the similarity (see §5.6).

5.3.3 Prototypes

One of the interesting things seen in the results, among other things, is that prototypic languages do not tend to be chosen more times than non-prototypic languages (as opposed to Skirgård et al., 2017), even though they tend to be chosen more when speakers need to recognize languages by their name. In other words, when speakers hear a language and they need to recognize it, they will tend to name the prototypical language

as an answer. But when speakers *hear* a language, and they need to determine what language heard is more similar to it – the prototypical language or some other unrelated language – they will not necessarily choose the prototypical language as more similar.

For example, Russian is the prototypical language of the Balto-Slavic languages, yet speakers thought that the female recording of Slovakian was more similar to the male recording of Ukrainian rather than to the male recording of Russian, thus following the suggested model (as Ukrainian and Slovakian share 88% similarity while Russian and Slovakian share only 68% similarity) rather than the prototypical notion. It is important to note, though, that some Balto-Slavic languages were indeed chosen as more similar to Russian, even when Russian appeared in the question as the dissimilar answer. For example, the female recording of Ukrainian was chosen as more similar to the male recording of Russian (68% similarity) rather than to the male recording of Bulgarian (85% similarity).

This kind of result could be caused by two possible factors: Russian-speaking participants and the quality of the non-shared features. The first option is that participants who know Russian can usually also understand Ukrainian to some extent since both languages have a similar lexicon, and therefore choose them as more similar because of a-prior linguistic knowledge. However, only eight out of the 58 participants (13.8%) who answered this specific question knew Russian, and funnily enough four of them chose Bulgarian while the other four chose Russian as the more similar language. Therefore, this option is not very likely. The second option, the option I tend to believe more, is that some shared features of Russian and Ukrainian, or some unshared features of Bulgarian and Ukrainian, caused Ukrainian to be chosen as more similar to Russian than to Bulgarian.

To emphasize my point, we can observe the shared and unshared features of Ukrainian, Bulgarian and Russian (see 14). It might be that the appearance of low vowels other than /a/ in Bulgarian was salient enough to make participants think that Ukrainian is less similar to Bulgarian than to Russian. On the other hand, it might be that the absence of rhotics in Russian was not salient enough to make participants think that Ukrainian is less similar to Russian than to Bulgarian. In other words, it might be that some features affected similarity more than others.

(14) Dissimilar features – Ukrainian VS. Bulgarian and Russian

Features	**Ukrainian_F**	**Bulgarian_M**	**Russian_M**
Front rounded vowels	✓	X	X
No. of consonants	12	19	17
C:VQ ration	3	4.75	5.67
Aspirated obstruents	X	✓	X
High vowels	✓	X	✓
Low vowels (except for /a/)	X	✓	X
Rhotics	✓	✓	X
Glides	✓	✓	X
Glides with 2 POA	✓	✓	X
Geminates	X	X	✓
Back vowels	✓	✓	X
Round vowels	✓	✓	X
[-ATR] vowels	✓	✓	X
Long vowels	X	X	✓
SPS	5.19	5.32	6.15
No. of consecutive consonants	3	3	2

Green cells in a row – languages with the same value of feature. Red cell in a row – a language with a different value of feature.

In addition, it is very possible that the number of appearances of some features may also affect similarity. For example, if speakers hear a click consonant only once they might think that it was an accidental utterance or some background noise, but if this click

consonant continues to appear they will understand that it is a part of the language they hear. The same goes, of course, for other features, too. Thus, the more a feature appears the more salient it might be for the speakers, therefore the number of appearances matters.

Finally, it is interesting to note that only Balto-Slavic languages were found as 100% similar to each other in the experiment (e.g. Slovak was found as 100% similar to Czech when the other option was Thai). It might be that Balto-Slavic languages have some shared prominent features which differentiate them from all other languages. On the other hand, it might be that Thai has a prominent feature which is not shared with Balto-Slavic languages. However, this problem might be considered as a family confound – i.e., it might be that languages from the same linguistic family are more similar to each other (see §6.1).

5.4 The Implications of the Study

The current thesis provides a basis for similarity research, which until now focused mainly on similarity between segments and less on the quantification of similarity as a sum of a given set of features. Of course, this study is basic, and further research should be made regarding the importance of each feature in the similarity quantification (see §5.6). However, even this basic similarity model, which managed to determine the similarity between languages, can provide guidance in various additional linguistic aspects. For example, knowing how similar two languages are to one another can help determine how easily a speaker will learn an additional language given his L1: given two relatively similar languages, either the speaker will have more difficulty learning the additional language because the differences between the languages are hard to be identified, or the speaker will have less difficulty learning the additional language

because of the minor differences between the languages. Similarity between languages can be used to create an experiment to examine this question.

Another example is taken from the forensic field: given some criminal runaway with a given L1, where would they prefer to hide – in a country with a more similar language to their L1 (so they can better fit into the community) or in a country with a less similar language (so the law enforcement will be less likely to find them)? This question cannot be examined without the ability to quantify the similarity between languages.

5.5 The Limitations of the Study

As in any research, not all possible confounds could be controlled for in the experiment. Most of the linguistic confounds are given and elaborated on in the sixth chapter (e.g., the familiarity of the languages to Hebrew speakers, the family from which the languages are, the number of languages the speaker knows, etc.). Yet, there were some methodological confounds and limitations that could affect the results or the analysis of the experiment.

First, it could be that there were not enough participants in the experiment. Although we made sure that there was a sufficient number of participants to answer each question (between 23-77, as was mentioned above), we still cannot be sure that there were no false results because of statistical reasons. In order to avoid that as much as possible, even more participants should have been found, so that each question will have at least 100 responses (though usually statistical significance should be reached with a minimum of 30 responses).

Second, the statistical test done in this experiment is the binomial distribution test, which is considered the weakest statistical test. The reason this test was done was because there was only one variant with two levels, similar or dissimilar, and this variant

needed to be tested in each question independently. The binomial distributional test's hypothesis is that both answers (similar/dissimilar) were not chosen equally, i.e., that the answers were not chosen by "flipping a coin" (when the probability to get the right answer or the wrong answer are the same – 50%). If the number of participants is sufficiently large, even getting a 60-40 chance can provide statistical significance. Although other tests, such as t-test or even Chi square, could have shown significance more accurately, I believe that the binomial distribution test is still good enough to be used for the purpose of the current study.

Third, the questions we asked the participants prior to starting the experiment (the number of languages they know and whether they have some linguistic knowledge) were too general and divided the participants into four sharply-cut groups: participants who speak only Hebrew and English and do not have linguistic knowledge, participants who speaker additional languages and do not have linguistic knowledge, participants who speak only Hebrew and English and have linguistic knowledge and participants who speak additional languages and have linguistic knowledge. However, there is still great variance among speakers in these groups: is a speaker of three languages the same as a speaker of five languages? Does a speaker of only Hebrew and English not hear Russian, Arabic, or French in their everyday life? Is someone who took one very general class in linguistics considered linguistically knowledgeable? What is the level of proficiency of speakers who know more Hebrew and English? And more. This type of limitation could have been avoided by providing the participants a proper questionnaire prior to the experiment. However, since the experiment was marketed to potential subjects as a game, we tried to avoid asking them too many "annoying" questions, so they would enjoy the experiment. But in a "normal" experiment with "normal" participants we will be able to

ask more questions about their background.

Finally, there was not enough variance in different background aspects of the participants: we did not ask the age of the participants, but we know that most of them are students and their friends, therefore very few non-students (people who are younger or older than the average age of students, about 26-27) participated in the experiment. It is important to observe younger people because they have a lot more interaction with languages other than Hebrew and English via the media, and it is important to observe older people because they have a lot less interaction with other languages since they use the media less and are considered "cleaner" speakers (i.e., speakers who are not influenced by other languages). The students' population is also considered (supposedly) more educated than the population of people with no higher education, therefore students may adapt to changes quicker and learn languages quicker, and therefore recognize them better.

5.6 Future Research

Some of the future research that can be done is already mentioned above, in §5.4 (e.g. in language acquisition). But of course, there is still a lot to examine about the quantification of similarity (over and above running the experiment in additional languages, of course). The most important future research for me is the weighting of the features to allow the model to predict similarity even more accurately.

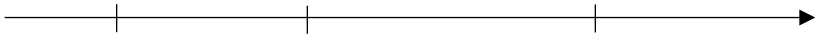
We all know, intuitively, that all features are not created equal. I, as a Hebrew Russian-English speaker, hear palatalized consonants much better than "mere" Hebrew-English speakers, because palatalization is contrastive in Russian, and I am sensitive to this feature as distinctive. For example, say you are a paramedic in battle, and someone shouts for you to take something, you need to understand whether you should take [krovʲ]

'blood' or [krov] *'cover'*. Sometimes perceiving a phonetic contrast can even save your life (or at least prevent a good scare): in Portuguese, there are nasalized vowels as well as regular vowels. I have a Brazilian friend who saw on several occasions Israelis (Hebrew speakers) who entered a shop and asked for a [pao]. What they probably did not know, is that *'bread'* is pronounced with nasalization, [pão], and what they had actually asked for from the now-angry salesman instead was male genitalia (though of course the gloss I gave here is way gentler and more censored than the actual gloss).

In any case, some features are hard for us to hear, and other features are very easy for us to hear. The more perceivable (for us) features, therefore, should have more weight when quantifying similarity: if it is easier for us to hear a feature, we can more easily identify its appearance or absence. For instance, how many of speakers (of any language) will miss the appearance of glottals? Or how many Russian speakers will miss palatalization? I suggest that not only the number of shared features quantifies similarity – but also the sum of the weights of each of these features.

First, we could try to speak of properties (= a group of features: segmental features, prosodic features or stress patterns) instead of mere features: in (15), C is the most salient property, therefore, it has the most weight for similarity; A is the least salient property – therefore, it has the least weight for similarity. Note that the weight difference between A and B might not be the same as the weight difference between B and C, as similarity differences are not necessarily equal; for example, it could be that A's impact on similarity is one point, B's impact is two points, and C's impact is four points.

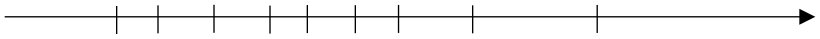
(15) A hypothetical similarity scale given property = {A, B, C}



(less salient) A B C (more salient)

However, it is hard to believe that *all* the components of A are less similar than *all* the components of B, which in turn are less similar than *all* the components of C. Rather than putting just X as a whole on the scale, it is very likely that we put every component of X (e.g. X_1, X_2, X_3, etc.) *separately* on the scale, and each component has its own weight, as can be seen in (16). When the components of each property appear separately on the scale, we do not have to assert that the prominence of one property, *including all components that belong to this property*, is greater than the prominence of another property.

(16) A hypothetical similarity scale given property = {A, B, C} when X = {X_1, X_2, X_3…}



(less salient) A_1 A_2 B_1 C_1 B_2 B_3 A_3 C_2 C_3 (more salient)

The similarity score of languages should be, as written above, the sum of the relevant components the speaker perceives as being in the Base language (the language they compare the two other languages to) and in the languages they compare it to. When a speaker is asked to determine whether language A or language B is closer to the Base language, they first check whether the most salient component exists in the Base language (and how many times it appears if it does exist), in Language A, in Language B and in their own language (=L1). Then, we define the similarity gaps relatively to the component we check between each language (A or B) and the Base language by subtracting the component's score of the Base language and the component's score of the relevant language. If the component also exists in L1, it gets additional points after the subtraction. Note that the greater the similarity gap is, the more different the Base language and the relevant language are, since one has this component, and the other does

not. Finally, after defining the similarity gaps of all components, we sum up the similarity gaps within each language – and the language with the lowest sum score will be reported as the closest to the Base language. See Appendix G for an example of the suggested model.

This is all a suggestion, of course, since this type of model has not been examined before. But should we examine the features and conclude that various features are weighted differently, we can try to explain why some dissimilar answers were chosen in several questions, or why similar answers were not chosen in other questions. Later, we can also consider that the number of appearances of several features can effect similarity quantification (i.e. the language can sound different if features appear more or less than others).

6. Possible Non-Phonological Properties that Might Be Confounds

As simple as the proposed model here is, it is very possible that other non-phonological linguistic properties may also affect similarity judgment. For example, in countries in which many languages are spoken, as in Israel, speakers hear more than one language almost every day; and since different languages have different linguistic properties, speakers hear many linguistic properties throughout their lives. In addition, many people travel abroad nowadays, therefore they may have heard other languages and become familiar with their properties.

In this chapter, I will suggest other non-phonological properties that might have affected the results of the current experiment. Please note that the properties suggested in this chapter are not the only properties that could affect similarity quantification, but only properties that could be derived from my research (see Skirgård, 2017 for more suggestions). Note that there should be a difference between, for example, which family the language in the question is from and which families the languages in the answer are from, and both options will be addressed in each sub-section. See Appendix H for further statistical data on some of the following properties (derived from the current experiment).

6.1 The Families the Languages Come From

Some historical linguists, who study the history of languages, believe that the various languages we have today were derived from one primal language (though this idea is controversial – see Ruhlen, 1994). At some point in time, and due to some causes, the primal language was split into several new languages, and these languages were later split into other new languages. The languages took most of their properties from the languages they were generated from, but some of these properties were created by the new languages or taken from other languages due to language contact. Therefore, it is

believed that the more branches two languages share, the more similar they are. For example, Hebrew, Arabic and Amharic should be more similar to each other than they are to Greek, English and German, because the former languages are Semitic languages, derived from the Proto-Semitic language, and the latter languages are Indo-European languages, derived from the Proto-Indo-European languages. However, Hebrew and Arabic should be more similar to each other than to Amharic, because Hebrew and Arabic are Central Semitic languages, while Amharic is a West Semitic language.

From this we can derive that the families from which the languages were generated may matter when quantifying similarity. Most importantly, it might be harder to choose a similar language as an answer when the languages in the answers are closely related, and it might be easier to choose an answer when one of the languages in the answer is closely related to the language in the question. In addition, it may be easier to choose the similar answer given questions with languages from a more familiar family (e.g., Afro-Asiatic languages versus Niger-Congo languages). However, it is important to note that languages from a given linguistic family are also phonologically similar, since they have a relatively close common ancestor, thus it is possible that the family that the language came from is not a non-phonological factor, but in fact a phonological one.

6.2 The Continents the Languages Are Spoken On

Related languages are often spoken in geographical proximity to one another. However, this is not always the case. For example, Afro-Asiatic languages are spoken, not surprisingly, in Africa and Asia. However, Amharic and Hebrew are examples of Semitic languages, which were generated from the Afro-Asiatic language, yet Amharic is spoken in Africa while Hebrew is spoken in Asia. Being in two different continents can cause exposure to different languages, thus exposure to different linguistic properties.

Following this line of thought, it may be that speakers of a language spoken in Asia will choose a language spoken in Asia as more similar to a language spoken in Asia in contrast to a language spoken in Africa. For example, Hebrew speakers may think that Indonesian is more similar to Assyrian than to Amharic because Hebrew, Indonesian and Assyrian are spoken mainly in Asia, while Amharic is spoken in Africa. In other words, the contact between languages might have an effect on similarity.

It is important to note, though, that a division by continent is not necessarily a good division. For example, both Israel and Russia are located in Asia, but Israel is located right next to Egypt, which is in Africa, and Russian is located right next to Ukraine, which is in Europe. Therefore, assuming we ignore the Russian speakers in Israel and the fact the Israel's population consists of many immigrants who speak in various languages, it is more plausible that Hebrew speakers will hear more Egyptian Arabic than they will hear Russian.

6.3 The Gender of the Speaker In the Recordings

Previous research showed that males and females differ in acoustical properties such as the center of gravity of initial consonants, the VOT of initial plosives, the vowel formant frequencies, the H1-H2 intensity difference in open vowels, the mean F0, the mean duration of dissyllabic words and more (Pépiot, 2015). Therefore, it might be inevitable that the gender of the speaker might have an influence on the quality of the linguistic properties uttered by them.

Since in the current experiment the speaker of the language in the question was of the opposite gender of the speakers of the languages in the answer, it might be that some shared linguistic properties were not noticeable enough. For example, it might be that the difference in vowel formants will cause a different perception of the same vowel, such

that /e/ will be perceived as [e] in one recording (of one gender), and as [ɛ] in the other recording (of the other gender). This type of variance will cause a different number of vowels in the languages' inventory and will cause the appearance of [-ATR] vowels in languages with no ATR distinction.

6.4 The Familiarity of Languages

Not only socio-linguistic properties may affect similarity, especially in Israel, in which speakers of many languages live (e.g., Hebrew, Arabic, French, Russian, Amharic and more). In addition, in the last few decades the technological improvements have allowed people to fly safely abroad and interact with speakers of other languages. Therefore, discussing the family of languages or the continent in which the languages are spoken as single independent properties may be a wrong decision. The languages the speakers are familiar with may be from various families and from various continents.

In addition, we cannot guarantee that all the speakers of a given language know the exact same group of languages, since each speaker is an independent person who can be in contact with whomever they like. As explained above, exposure to other languages, even if the exposure is not vast or consistent, may cause an exposure to various non-native linguistic properties, thus causing them to sound more familiar to speakers. For example, many Hebrew speakers who live in Israel do not speak Russian or French, but they will identify these languages if they hear them on the street. And some languages, which are not spoken frequently in Israel, but which speakers have come into contact with outside the country, will probably not be identified by Hebrew speakers but they will probably say it "sounds familiar".

In any case, it might be that the familiarity of languages to speakers participating in the experiment has affected the results. However, the influence familiarity can have on

similarity is not clear: on the one hand, the familiar languages can prime their linguistic properties so that these properties will be taken into consideration when comparing languages (e.g., if one of the answers is a familiar language and it has click consonants, and if the language in the question has click consonants as well, then the familiar language will be chosen because of this property). On the other side, when we know a language, and especially when we know it well, it stands apart from other languages we do not know, and it may feel unique and special, so that it cannot be compared to other languages (e.g., if one of the answers is a familiar language and the other language is less familiar, then the familiar language is more distinguishable to the speaker than the less familiar language, and it cannot be compared to the language in the question). See Van Engen (2010), Flemming et al. (2014), and Sternin et al. (2021) for more information.

6.5 The Knowledge of the Speakers

Finally, I think we cannot discuss similarity between languages without taking into consideration the knowledge the speaker has on languages. The more the speakers know about languages, the more they notice differences between languages; and we can almost guarantee that speakers who notice changes between languages will analyze languages differently than speakers who have less knowledge about the way languages work. The linguistic knowledge can come mostly from two areas: knowing many languages and learning the linguistic knowledge consciously.

6.5.1 The Linguistic Knowledge of Speakers

It is a safe assumption to make that linguists know more than non-linguists about languages and their properties. Almost every linguist, whether they are phoneticians, phonologists, semanticists, syntacticians or from other linguistic fields, knows the basic aspects that make a language the way it is. Therefore, linguists have more information to

rely on and to use when they quantify the similarity of languages, even if they do so unconsciously. However, one does not have to be a linguist to know about languages. These days, a simple Google search about languages will suffice. In other words, the more the speaker knows about languages, the more tools they will have to recognize languages and distinguish between them.

Knowing the properties of languages should, as was said above, help people recognize languages, or at least recognize the properties of the languages. However, since the linguistic knowledge of speakers differs not only in their level of knowledge but also in their field of knowledge (e.g., it is not guaranteed that academic institutions teach the exact same knowledge, as the teaching is done by researchers which are replaced sometimes), it is hard to hypothesize how linguistic knowledge helps speakers. For example, some speakers might have heard about click consonants before so they might pay more attention to finding click consonants, while some speakers might have learned the family trees of languages and will be able to recognize languages of the same linguistic family.

6.5.2 The Number of Languages the Speaker Knows

Finally, the number of languages the speaker knows may affect their perspective on languages and on their properties. The more languages the speaker knows, especially if these languages are from different linguistic families, the more knowledge they will have on languages (even if this knowledge is not conscious).

Since the linguistic knowledge of speakers is mostly unconscious, the speakers will not calculatedly choose a language based on their knowledge, but instead they will choose a language based on their linguistic intuition: their decision will probably be based on the linguistic properties of the languages they speak. For example, a Hebrew

and Russian speaker will notice both the appearance of glottal stops and the appearance of palatalized consonants (at least to some extent), since glottal stops exist (although not always) in Hebrew, and palatalized consonants exist in Russian.

7. Conclusions

Human languages are complex things: they are composed of many little components that merge into a form of communication that other people can understand and respond to. Different languages have different components, and speakers seem to know them and how to use them. Speakers have linguistic intuitions, which are based on their own knowledge, even if this knowledge is unconscious. Therefore, we can ask ourselves what the differences between languages are, how speakers perceive these differences, and how they use these differences in their day-to-day life.

In this thesis, I built a model that quantifies similarity between languages by calculating the percentage of acoustic and phonological features they share. This model quantifies similarity among languages, but potentially quantifies any type of similarity but taking to account the individual features which are relevant for similarity. i.e. breaking down the complex notion of similarity into the individual components it's made from.

I believe that since we all know more or less what languages we hear around us, and we more often than not agree on the identity of these languages, then similarity is quantifiable. If it is quantifiable, then we can find the quantification using various methods of comparison. This may be a very bumpy road, but I think it is not a dead-end; we just need to fasten our seatbelts and enjoy the ride.

References

Barkat, M., & I. Vasilescu. (2001). From perceptual designs to linguistics typology and automatic language identification: overview and perspectives. In *Proceedings Of the Eurospeech 2001*, 1065-1068.

Beckman, M. (1986). *Stress and Non-stress Accent*. Dordrecht: Foris.

Boersma, P., & D. Weenink. (2009). *Praat: Doing phonetics by computer (computer programme)*. (fon.hum.uva.nl/praat/).

Bradlow, A., Clopper, C., Smiljanic, R., & M. Walter. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication, 52,* 930-942.

Clements, G. (1990). The role of the sonority cycle in core syllabification. In J. Kingston and M. Beckman (eds.) *Papers in laboratory phonology I: Between the grammar and physics of speech*. Cambridge: Cambridge University Press. 282–333.

Cohen, E. (2009). The role of similarity in phonology: evidence from loanword adaptation in Hebrew. Ph.D. dissertation. Tel-Aviv: Tel-Aviv University.

Cole, R. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics, 1*, 153-156.

Cooper, E. (1983). The perception of fluent speech. *Annals of the New York Academy ofScience, 405*, 48-63.

Crowley, T., & C. Bowern. (2010). *An introduction to historical linguistics*. Oxford: OxfordUniversity Press.

Dryer, M. S., & M. Haspelmath (eds.). (2013). The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Eden, E. (2018). Measuring Phonological Distance Between Languages. Ph.D. dissertation. London: London's Global University.

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences, 111*(38), 13795-13798.

Fry, D. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America, 27,* 765–768.

Fry, D. (1958). Experiments in the perception of stress. *Language and Speech,* 1, 126-152.

Garnes, S. & Z. Bond. (1980). A slip of the ear: a snip of the ear? A slip of the year? In V. Fromkin (ed.), *Errors in linguistics performance. Slips of the tongue, ear, pen and hand*. Academic press. 231-239.

Gordon, M. & T. Roettger. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard, 3*(1).

Hung, T. (2000). Towards a phonology of Hong Kong English. *World Englishes, 19*(3), 337-356.

Hyman, L. (1970). The role of borrowing in the justification of phonological grammars. *Studies in African Linguistics, 1,* 1-48.

Hyman, L. (1977). Tone and/or accent. In D. Napoli (ed.) *Elements of Tone, Stress andintonation*. Washington: Georgetown University Press. 1-20.

Ito, C., & M. Kenstowicz. (2017). Pitch accent in Korean. In M. Aronoff *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press.

Jassem, W. (1959). The phonology of Polish stress. *Word, 15*, 252-269.

Jun, S. (2005). Korean intonational phonology and prosodic transcription. In S. Jun (ed.) *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press. 201-229.

de Lacy, P. (2002). The interaction of tone and stress in Optimality Theory. *Phonology, 19*, 1-32.

Leena, M., Srirama Murthy, K., Mahadeva Prasanna, S., & B. Yegnanarayana. (2004). Features for speaker and language identification. In *Proceedings of Odyssey*, 156-159.

Leena, M., Srinivasa Rao, K., & B. Yegnanarayana. (2005). Neural network classifiers for language identification using phonotactics and prosodic features. In *Proceedings of 2005 international conference on intelligent sensing and information processing*, 404- 408.

Lewis M. P., Simons G. F., & Fennig C. D. (2014). *Ethnologue: Languages of the World, 17th ed*. Online version: http://www.ethnologue.com/17/. Dallas, Texas: SIL International.

Longobardi, G., C. Guardiano. (2009). Evidence for syntax as a signal of historical relatedness. *Lingua*, *119*(11), 1679-1706.

Longobardi, G., & C. Guardiano. (2017). Phylogenetic Reconstruction in Syntax: The Parametric Comparison Method. In A. Ledgeway and I. Roberts (eds.) *The Cambridge Handbook of Historical Syntax*. Cambridge: Cambridge University Press. 241-272.

Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A., & A. Ceolin. (2013). Toward a syntactic phylogeny of Modern Indo-European Languages. *Journal of Historical Linguistics, 1*(3), 122-152.

McMahon, A., R. McMahon. (2005). *Language classification by numbers*. Oxford: Oxford University Press.

Meng, H., Yee Lo, Y., Wang, L., & W. Yin Lau. (2007). Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 437-442.

Murrey, R., & T. Vennemann. (1983). Sound Change and Syllable Structure in Germanic Phonology. *Linguistic Society of America, 59*(3), 514-528.

Parker, S. (2008). Sound level protrusions as physical correlates of sonority. *Journal of Phonetics, 36*, 55-90.

Pépiot, E. (2015). Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers. *Corela. Cognition, représentation, langage*, (HS-16).

Prince, A., & P. Smolensky. (1993). *Optimality theory: Constraint Interaction in Generative Grammar*. Boulder: Rutgers University.

Remijsen, B. (2003). New perspectives in word-prosodic typology. *International Institutefor Asian Studies: The Newsletter, 32*, 29.

Ringbom, H. (2007). *Cross-linguistic Similarity in Foreign Language Learning*. England: Clevedom.

Ruhlen, M. (1994). *On the Origin of Languages: Studies in Linguistic Taxonomy*. Redwood City: Stanford University Press.

Selkirk, E. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry, 11*, 563-606.

Shinohara, S. (2006). Perceptual effects in final cluster reduction patterns. *Lingua, 116*, 1046-1078.

Skirgård, H., Roberts, S. G., & L. Yencken. (2017). Why are some languages confused for others? Investigating data from the Great Language Game. *PLoS ONE, 12*(4), e0165934. https://doi.org/10.1371/journal.pone.0165934

Steriade, D. (2001). Directional asymmetries in place assimilation: a perceptual account. In E. Hume and K. Johnson (eds.), *The role of speech perception in phonology*. Academic press. 219-250.

Steriade, D. (2001/2008). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. In K. Hanson and S. Inkelas (eds.), *The nature of the word*. Cambridge: MIT Press. 151-179.

Sternin, A., McGarry, L. M., Owen, A. M., & Grahn, J. A. (2021). The Effect of Familiarity on Neural Representations of Music and Language. *Journal of Cognitive Neuroscience, 33*(8), 1595-1611.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

Turnbull, R., & S. Peperkamp. (2017). The asymmetric contribution of consonants and vowels to phonological similarity. Evidence from lexical priming. *The Mental Lexicon, 12*(3), 404-430.

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first and second language multi-talker babble. *Speech communication, 52*(11-12), 943-953.

Vasilescu, I., Pellegrino, F., & J. Hombert. (2000). Perceptual features for the identificationof Romance language. *In Proceedings of International Congress of Spoken Language Processing* 2, 543-546.

Vasilescu, I., Candea, M., & M. Adda-Decker. (2005). Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages. In *Proceeding of Interspeech*, 1773-1776.

Zissman, M. (1996). Comparison of four approaches to automatic language identification of telephone speech. In *IEEE Acoustics, Speech and Signal processing 4*(1), 31-44.

Zwicky, A. (1976). Well, this rock and roll has got to stop, Junior's head is hard as a rock. In S. Mufwene, C. Walker and S. Steever (eds.) *Proceedings of Chicago Linguistic Society 12*, University of Chicago. 676-697.

Appendices

Appendix A- A List of Languages Presented in the Experiment

No.	Family	Language	Female recording	Male recording
1	Afro-Asiatic	Amharic	V	V
2		Egyptian Arabic		V
3		Hausa		V
4		Hebrew	V	V
5		Somali		V
6	Austroasiatic	Vietnamese	V	V
7	Austronesian	Fijian	V	V
8	Dravidian	Telugu	V	V
9	Indo-European – Balto-Slavic	Bulgarian	V	V
10		Croatian	V	V
11		Czech	V	V
12		Polish	V	V
13		Russian	V	V
14		Slovak	V	V
15		Ukrainian	V	V
16	Indo-European – Germanic	German	V	V
17		Norwegian	V	V
18		Swedish	V	
19		Yiddish		V
20	Indo-European – Indo-Iranian	Hindi	V	V
21		Oriya	V	
22		Pashto		V
23		Persian	V	V
24	Indo-European – Italic	French	V	V
25		Italian	V	V
26		Portuguese	V	V
27		Spanish	V	V
28	Japonic	Japanese	V	V
29	Koreanic	Korean	V	Vx2
30	Kra-Dai	Thai	V	V
31	Niger-Congo	Xhosa	V	V
32		Yoruba		V
33	Sino-Tibetan	Mandarin	V	V
34	Turkic	Turkish	V	V
35	Uralic	Finnish		V
36		Hungarian	V	V

Appendix B- A List of the Phonological Properties of Languages – By WALS

consonant inventory	No. of cons.		No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
small	6-14		22	61%	36	56%	89	16%
moderately small	15-18		10	28%	24	38%	122	22%
average	19-25		4	11%	4	6%	201	36%
moderately large	26-33		0	0%	0	0%	94	17%
large	34 or more		0	0%	0	0%	57	10%
			36		64		563	

vowel inventory	No. of vowels		No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
small	2-4		22	61%	42	66%	93	17%
average	5-6		9	25%	17	27%	287	54%
large	7-14		5	14%	5	8%	154	29%
			36		64		534	

C:VQ ratio	Ratio		No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
low	below 2		3	8%	8	13%	58	10%
moderately low	2.0-2.75		9	25%	9	14%	101	18%
average	2.75-4.5		17	47%	36	56%	234	41%
moderately high	4.5-6.5		7	19%	10	16%	102	18%
high	6.5 or higher		0	0%	1	2%	69	12%
			36		64		564	

Voicing contrast in obstruents			No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
No voicing contrast			1	3%	4	6%	182	32%
Voicing contrast only in plosives			6	17%	14	22%	189	33%
Voicing contrast only in fricatives			1	3%	2	3%	38	7%
Voicing contrast only in affricates			0	0%	1	2%		
Voicing contrast in plosives and fricatives			21	58%	36	56%	158	28%
Voicing contrast in plosives and affricates			1	3%	2	3%		
Voicing contrast in fricatives and affricates			0	0%	0	0%		
Voicing contrast in all			6	17%	5	8%		
			36		64		567	

Uvular consonants			No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
No uvulars			30	83%	55	86%	470	83%
Uvular stops only			1	3%	1	2%	38	7%
Uvular continuants only			5	14%	8	13%	11	2%
Uvular stops and continuants			0	0%	0	0%	48	8%
			36		64		567	

Glottalized consonants			No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
No glottalized consonants			30	83%	54	84%	409	72%
Ejectives only			4	11%	7	11%	58	10%
Implosives only			1	3%	2	3%	55	10%
Glottalized resonants only			0	0%	0	0%	4	1%
Glottalized obstruents only			1	3%	1	2%		0%
Ejectives and implosives			0	0%	0	0%	14	2%
Ejectives and glottalized resonants			0	0%	0	0%	20	4%
Implosives and glottalized resonants			0	0%	0	0%	4	1%
Ejectives, implosives and glottalized resonants			0	0%	0	0%	3	1%
			36		64		567	

Lateral consonants			No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
No laterals			3	8%	12	19%	95	17%
Only /l/, no other laterals			22	61%	38	59%	388	68%
More than one lateral (inc. /l/)			11	31%	14	22%	29	5%
/l/ and lateral obstruents			0	0%	0	0%	47	8%
No /l/, but lateral obstruents			0	0%	0	0%	8	1%
			36		64		567	

Front Rounded Vowels			No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
None			27	75%	50	78%	525	93%
High and mid			6	17%	9	14%	23	4%
High only			2	6%	4	6%	8	1%
Mid only			1	3%	1	2%	6	1%
			36		64		562	

Syllable structure	Structures		No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
Simple - 1-2 consonant sequence	V or CV		20	56%	43	67%	61	13%
Moderately complex - 3-4 cons. seq.	CVC, CCV, VCC, etc.		15	42%	20	31%	274	56%
Complex - 5 cons. seq.	CCCV, CVCCC, etc.		1	3%	1	2%	151	31%
			36		64		486	

Tone			No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
None			21	58%	39	61%	307	58%
Simple	Mostly pitch Accent		9	25%	16	25%	132	25%
Complex	Contour		6	17%	9	14%	88	17%
			36		64		527	

Presence of Uncommon consonants			No. of languages	% of languages	No. of recordings	% of rec.	By WALS	% by WALS
None			28	78%	55	86%	449	79%
Clicks			1	3%	2	3%	9	2%
Labial-velars (e.g. /kp͡/)			0	0%	0	0%	45	8%
Pharyngeals			0	0%	0	0%	21	4%
"Th" sounds			5	14%	5	8%	40	7%
Clicks, pharyngeals, and "th"			0	0%	0	0%	1	0.2%
Pharyngeals and "th"			2	6%	2	3%	2	0.4%
			36		64		567	

Appendix C- The Similarity Between Languages by Percentage

Language	Amharic_F	Amharic_M	Egyptian Arabic_M	Hausa_M	Hebrew_F	Hebrew_M	Somali_M	Vietnamese_F	Vietnamese_M	Fijian_F	Fijian_M	Telugu_F	Telugu_M	Bulgarian_F	Bulgarian_M	Croatian_F	Croatian_M	Czech_F	Czech_M	Polish_F	Polish_M
Amharic_F		76%	49%	71%	66%	73%	63%	71%	66%	71%	59%	66%	66%	68%	59%	68%	73%	73%	71%	63%	66%
Amharic_M	76%		63%	80%	78%	80%	78%	76%	76%	80%	80%	83%	73%	76%	71%	80%	76%	66%	71%	73%	73%
Egyptian Arabic_M	49%	63%		61%	73%	71%	68%	59%	63%	68%	66%	66%	73%	63%	61%	66%	68%	63%	61%	63%	71%
Hausa_M	71%	80%	61%		71%	76%	78%	73%	73%	68%	71%	71%	78%	63%	63%	80%	80%	68%	66%	71%	76%
Hebrew_F	66%	78%	73%	71%		93%	76%	76%	85%	76%	80%	73%	80%	78%	71%	78%	78%	80%	71%	76%	78%
Hebrew_M	73%	80%	71%	76%	93%		76%	73%	76%	76%	80%	80%	88%	80%	76%	85%	90%	85%	80%	78%	83%
Somali_M	63%	78%	68%	78%	76%	76%		73%	73%	85%	78%	78%	66%	61%	71%	73%	68%	61%	73%	66%	68%
Vietnamese_F	71%	76%	59%	73%	76%	73%	73%		90%	73%	66%	73%	66%	66%	61%	78%	68%	80%	71%	66%	63%
Vietnamese_M	66%	76%	63%	73%	85%	76%	73%	90%		73%	76%	68%	71%	66%	61%	76%	71%	83%	68%	68%	66%
Fijian_F	71%	80%	68%	68%	76%	76%	85%	73%	73%		83%	78%	66%	68%	66%	71%	66%	61%	73%	66%	68%
Fijian_M	59%	80%	66%	71%	80%	80%	78%	66%	76%	83%		78%	76%	66%	68%	73%	76%	71%	71%	68%	78%
Telugu_F	66%	83%	66%	71%	73%	80%	78%	73%	68%	78%	78%		73%	71%	68%	76%	73%	68%	66%	63%	71%
Telugu_M	66%	73%	73%	78%	80%	88%	66%	66%	71%	66%	76%	73%		78%	73%	88%	93%	80%	76%	80%	93%
Bulgarian_F	68%	76%	63%	63%	78%	80%	61%	66%	66%	68%	66%	71%	78%		88%	78%	78%	78%	78%	80%	80%
Bulgarian_M	59%	71%	61%	63%	71%	76%	71%	61%	61%	66%	68%	68%	73%	88%		78%	78%	76%	78%	83%	83%
Croatian_F	68%	80%	66%	80%	78%	85%	73%	78%	76%	71%	73%	76%	88%	78%	78%		90%	73%	68%	80%	80%
Croatian_M	73%	76%	68%	80%	78%	90%	68%	68%	71%	66%	76%	73%	93%	78%	78%	90%		76%	71%	83%	83%
Czech_F	73%	66%	63%	68%	80%	85%	61%	80%	83%	61%	71%	68%	80%	78%	76%	73%	76%		85%	83%	80%
Czech_M	71%	71%	61%	66%	71%	80%	73%	71%	68%	73%	71%	66%	76%	78%	78%	68%	71%	85%		78%	80%
Polish_F	63%	73%	63%	71%	76%	78%	66%	66%	68%	66%	68%	63%	80%	80%	83%	80%	83%	83%	78%		88%
Polish_M	66%	73%	71%	76%	78%	83%	68%	63%	66%	68%	78%	71%	93%	80%	83%	80%	83%	80%	80%	88%	
Russian_F	66%	71%	63%	61%	80%	83%	71%	71%	80%	68%	76%	68%	76%	78%	80%	71%	78%	85%	80%	80%	73%

Language	Amharic_F	Amharic_M	Egyptian Arabic_M	Hausa_M	Hebrew_F	Hebrew_M	Somali_M	Vietnamese_F	Vietnamese_M	Fijian_F	Fijian_M	Telugu_F	Telugu_M	Bulgarian_F	Bulgarian_M	Croatian_F	Croatian_M	Czech_F	Czech_M	Polish_F	Polish_M
Russian_M	71%	71%	61%	61%	66%	73%	59%	66%	66%	63%	66%	71%	76%	76%	66%	73%	71%	71%	66%	66%	71%
Slovak_F	76%	76%	66%	71%	78%	83%	68%	76%	68%	76%	68%	73%	78%	90%	83%	76%	78%	83%	83%	83%	80%
Slovak_M	73%	73%	66%	71%	85%	83%	76%	71%	73%	76%	73%	68%	80%	88%	80%	80%	80%	83%	85%	83%	83%
Ukrainian_F	59%	73%	61%	63%	78%	76%	68%	66%	71%	68%	71%	66%	73%	80%	85%	68%	73%	80%	80%	85%	80%
Ukrainian_M	66%	66%	59%	66%	76%	78%	66%	71%	68%	66%	63%	63%	76%	85%	85%	71%	76%	83%	83%	85%	76%
German_F	56%	51%	63%	51%	68%	61%	56%	63%	68%	51%	59%	56%	61%	56%	61%	61%	59%	76%	61%	61%	59%
German_M	56%	61%	66%	61%	71%	68%	56%	66%	71%	56%	63%	54%	63%	71%	76%	66%	63%	78%	68%	76%	68%
Norwegian_F	66%	78%	56%	71%	68%	71%	76%	76%	76%	71%	71%	66%	73%	68%	68%	76%	71%	76%	78%	76%	71%
Norwegian_M	61%	61%	56%	61%	71%	68%	68%	59%	63%	66%	63%	59%	68%	63%	66%	68%	73%	68%	71%	76%	71%
Swedish_F	61%	68%	56%	63%	76%	68%	66%	68%	78%	66%	68%	63%	63%	68%	66%	73%	68%	76%	61%	66%	66%
Yiddish_M	61%	68%	59%	59%	80%	85%	61%	68%	76%	61%	71%	66%	76%	80%	76%	71%	71%	85%	76%	78%	78%
Hindi_F	51%	61%	51%	56%	61%	61%	76%	68%	68%	71%	68%	63%	54%	54%	61%	61%	54%	66%	71%	66%	61%
Hindi_M	49%	56%	46%	61%	59%	61%	66%	59%	54%	63%	61%	71%	59%	63%	63%	66%	59%	61%	66%	59%	61%
Oriya_F	63%	73%	54%	80%	76%	78%	76%	76%	71%	73%	68%	73%	76%	68%	68%	80%	73%	66%	68%	76%	73%
Pashto_M	56%	63%	56%	61%	73%	68%	66%	71%	80%	71%	68%	71%	68%	71%	66%	73%	68%	78%	76%	73%	68%
Persian_F	71%	76%	56%	68%	73%	80%	76%	73%	66%	68%	71%	66%	73%	78%	85%	78%	73%	78%	85%	80%	78%
Persian_M	63%	78%	66%	68%	73%	78%	76%	66%	73%	73%	80%	66%	73%	73%	78%	76%	73%	73%	76%	76%	73%
French_F	44%	54%	56%	46%	66%	66%	63%	59%	63%	59%	63%	68%	54%	56%	61%	63%	59%	59%	54%	66%	56%
French_M	51%	66%	56%	61%	71%	68%	66%	61%	73%	66%	73%	59%	59%	59%	61%	63%	59%	71%	66%	71%	66%
Italian_F	68%	85%	68%	66%	85%	80%	71%	71%	76%	80%	85%	76%	76%	80%	76%	73%	73%	78%	78%	78%	83%
Italian_M	76%	85%	68%	73%	90%	90%	80%	78%	83%	85%	80%	83%	80%	83%	76%	80%	80%	76%	73%	76%	78%
Portuguese_F	68%	78%	61%	76%	85%	93%	73%	68%	73%	68%	76%	76%	85%	78%	73%	83%	83%	73%	71%	80%	76%
Portuguese_M	73%	78%	59%	66%	83%	90%	71%	68%	73%	68%	76%	73%	83%	88%	83%	78%	78%	85%	83%	88%	83%
Spanish_F	66%	80%	73%	68%	90%	85%	80%	68%	78%	80%	88%	73%	78%	80%	85%	76%	80%	76%	78%	85%	83%

Language	Amharic_F	Amharic_M	Egyptian Arabic_M	Hausa_M	Hebrew_F	Hebrew_M	Somali_M	Vietnamese_F	Vietnamese_M	Fijian_F	Fijian_M	Telugu_F	Telugu_M	Bulgarian_F	Bulgarian_M	Croatian_F	Croatian_M	Czech_F	Czech_M	Polish_F	Polish_M
Spanish_M	66%	78%	78%	73%	90%	83%	76%	76%	83%	76%	76%	73%	80%	76%	73%	83%	80%	78%	68%	76%	73%
Japanese_F	73%	88%	66%	78%	76%	83%	90%	73%	73%	88%	85%	80%	76%	68%	71%	78%	80%	71%	80%	71%	78%
Japanese_M	76%	85%	59%	85%	80%	88%	83%	73%	71%	80%	78%	73%	80%	71%	73%	83%	88%	76%	80%	76%	80%
Korean_F	68%	71%	49%	61%	63%	66%	68%	71%	66%	61%	66%	59%	59%	61%	71%	66%	66%	66%	66%	71%	63%
Korean_M1	61%	71%	54%	63%	73%	68%	66%	63%	71%	63%	80%	63%	66%	66%	71%	73%	68%	73%	73%	68%	73%
Korean_M2	63%	71%	49%	68%	66%	63%	68%	71%	73%	66%	63%	68%	61%	56%	54%	68%	63%	68%	71%	59%	59%
Thai_F	54%	66%	51%	61%	68%	66%	73%	73%	78%	63%	73%	63%	61%	59%	66%	71%	61%	71%	66%	68%	66%
Thai_M	59%	73%	54%	66%	68%	68%	66%	68%	73%	61%	68%	63%	63%	68%	73%	63%	68%	78%	73%	78%	73%
Xhosa_F	63%	73%	59%	68%	66%	76%	66%	61%	63%	61%	68%	59%	71%	73%	78%	66%	73%	78%	78%	80%	76%
Xhosa_M	71%	88%	56%	73%	78%	78%	73%	68%	73%	68%	78%	71%	73%	73%	73%	73%	71%	76%	78%	76%	80%
Yoruba_M	56%	78%	61%	71%	78%	71%	76%	61%	73%	76%	90%	73%	68%	68%	71%	71%	68%	71%	73%	76%	76%
Mandarin_F	61%	68%	54%	63%	68%	71%	68%	76%	76%	61%	66%	61%	71%	66%	71%	73%	68%	80%	78%	76%	71%
Mandarin_M	59%	78%	54%	71%	76%	76%	76%	68%	68%	66%	73%	73%	68%	71%	78%	76%	71%	76%	76%	76%	78%
Turkish_F	66%	73%	54%	63%	73%	76%	63%	78%	83%	61%	68%	66%	73%	73%	68%	80%	73%	88%	78%	76%	73%
Turkish_M	66%	71%	56%	63%	66%	73%	71%	80%	76%	63%	66%	66%	71%	68%	68%	78%	76%	80%	76%	73%	68%
Finnish_M	63%	78%	63%	68%	80%	76%	76%	80%	80%	83%	78%	73%	71%	71%	61%	73%	68%	73%	68%	68%	68%
Hungarian_F	61%	71%	59%	66%	71%	73%	59%	68%	68%	63%	68%	66%	71%	78%	71%	71%	68%	78%	71%	73%	73%
Hungarian_M	59%	68%	61%	59%	68%	68%	63%	71%	66%	68%	63%	66%	68%	80%	76%	71%	68%	78%	73%	78%	71%

Language	Russian_F	Russian_M	Slovak_F	Slovak_M	Ukrainian_F	Ukrainian_M	German_F	German_M	Norwegian_F	Norwegian_M	Swedish_F	Yiddish_M	Hindi_F	Hindi_M	Oriya_F	Pashto_M	Persian_F	Persian_M	French_F	French_M	Italian_F
Amharic_F	66%	71%	76%	73%	59%	66%	56%	56%	66%	61%	61%	61%	51%	49%	63%	56%	71%	63%	44%	51%	68%
Amharic_M	71%	71%	76%	73%	73%	66%	51%	61%	78%	61%	68%	68%	61%	56%	73%	63%	76%	78%	54%	66%	85%
Egyptian Arabic_M	63%	61%	66%	66%	61%	59%	63%	66%	56%	56%	56%	59%	51%	46%	54%	56%	56%	66%	56%	56%	68%
Hausa_M	61%	61%	71%	71%	63%	66%	51%	61%	71%	61%	63%	59%	56%	61%	80%	61%	68%	68%	46%	61%	66%
Hebrew_F	80%	66%	78%	85%	78%	76%	68%	71%	68%	71%	76%	80%	61%	59%	76%	73%	73%	73%	66%	71%	85%
Hebrew_M	83%	73%	83%	83%	76%	78%	61%	68%	71%	68%	68%	85%	61%	61%	78%	68%	80%	78%	66%	68%	80%
Somali_M	71%	59%	68%	76%	68%	66%	56%	56%	76%	68%	66%	61%	76%	66%	76%	66%	76%	76%	63%	66%	71%
Vietnamese_F	71%	66%	76%	71%	66%	71%	63%	66%	76%	59%	68%	68%	68%	59%	76%	71%	73%	66%	59%	61%	71%
Vietnamese_M	80%	66%	68%	73%	71%	68%	68%	71%	76%	63%	78%	76%	68%	54%	71%	80%	66%	73%	63%	73%	76%
Fijian_F	68%	63%	76%	76%	68%	66%	51%	56%	71%	66%	66%	61%	71%	63%	73%	71%	68%	73%	59%	66%	80%
Fijian_M	76%	66%	68%	73%	71%	63%	59%	63%	71%	63%	68%	71%	68%	61%	68%	68%	71%	80%	63%	73%	85%
Telugu_F	68%	71%	73%	68%	66%	63%	56%	54%	66%	59%	63%	66%	63%	71%	73%	71%	66%	66%	68%	59%	76%
Telugu_M	76%	76%	78%	80%	73%	76%	61%	63%	73%	68%	63%	76%	54%	59%	76%	68%	73%	73%	54%	59%	76%
Bulgarian_F	78%	76%	90%	88%	80%	85%	56%	71%	68%	63%	68%	80%	54%	63%	68%	71%	78%	73%	56%	59%	80%
Bulgarian_M	80%	66%	83%	80%	85%	85%	61%	76%	68%	66%	66%	76%	61%	63%	68%	66%	85%	78%	61%	61%	76%
Croatian_F	71%	73%	76%	80%	68%	71%	61%	66%	76%	68%	73%	71%	61%	66%	80%	73%	78%	76%	63%	63%	73%
Croatian_M	78%	71%	78%	80%	73%	76%	59%	63%	71%	73%	68%	71%	54%	59%	73%	68%	73%	73%	59%	59%	73%
Czech_F	85%	71%	83%	83%	80%	83%	76%	78%	76%	68%	76%	85%	66%	61%	66%	78%	78%	73%	59%	71%	78%
Czech_M	80%	66%	83%	85%	80%	83%	61%	68%	78%	71%	61%	76%	71%	66%	68%	76%	85%	76%	54%	66%	78%
Polish_F	80%	66%	83%	83%	85%	85%	61%	76%	76%	76%	66%	78%	66%	59%	76%	73%	80%	76%	66%	71%	78%
Polish_M	73%	71%	80%	83%	80%	76%	59%	68%	71%	71%	66%	78%	61%	61%	73%	68%	78%	73%	56%	66%	83%
Russian_F		73%	78%	83%	88%	88%	71%	73%	71%	76%	68%	73%	63%	51%	68%	78%	71%	73%	61%	68%	73%
Russian_M	73%		68%	68%	68%	71%	56%	59%	63%	56%	59%	66%	56%	54%	63%	63%	61%	61%	59%	54%	66%
Slovak_F	78%	68%		90%	80%	88%	61%	68%	78%	71%	68%	78%	61%	66%	73%	68%	83%	73%	59%	61%	80%

Language	Russian_F	Russian_M	Slovak_F	Slovak_M	Ukrainian_F	Ukrainian_M	German_F	German_M	Norwegian_F	Norwegian_M	Swedish_F	Yiddish_M	Hindi_F	Hindi_M	Oriya_F	Pashto_M	Persian_F	Persian_M	French_F	French_M	Italian_F
Slovak_M	83%	68%	90%		78%	88%	63%	68%	76%	78%	71%	78%	66%	63%	76%	73%	85%	80%	61%	66%	83%
Ukrainian_F	88%	68%	80%	78%		93%	66%	80%	83%	78%	73%	83%	63%	54%	66%	71%	78%	73%	68%	73%	83%
Ukrainian_M	88%	71%	88%	88%	93%		68%	76%	83%	80%	73%	78%	61%	59%	68%	68%	80%	71%	63%	66%	73%
German_F	71%	56%	61%	63%	66%	68%		78%	71%	61%	68%	71%	63%	51%	54%	66%	61%	68%	71%	71%	59%
German_M	73%	59%	68%	68%	80%	76%	78%		76%	71%	80%	78%	63%	56%	54%	73%	71%	71%	66%	78%	68%
Norwegian_F	71%	63%	78%	76%	83%	83%	71%	76%		76%	78%	73%	71%	63%	71%	73%	83%	88%	61%	76%	76%
Norwegian_M	76%	56%	71%	78%	78%	80%	61%	71%	76%		76%	73%	68%	59%	63%	73%	71%	66%	73%	68%	68%
Swedish_F	68%	59%	68%	71%	73%	73%	68%	80%	78%	76%		78%	68%	66%	59%	76%	71%	80%	73%	76%	71%
Yiddish_M	73%	66%	78%	78%	83%	78%	71%	78%	73%	73%	78%		66%	71%	66%	80%	80%	76%	68%	76%	80%
Hindi_F	63%	56%	61%	66%	63%	61%	63%	63%	71%	68%	68%	66%		76%	68%	76%	66%	66%	78%	78%	66%
Hindi_M	51%	54%	66%	63%	54%	59%	51%	56%	63%	59%	66%	71%	76%		76%	68%	71%	63%	76%	66%	59%
Oriya_F	68%	63%	73%	76%	66%	68%	54%	54%	71%	63%	59%	66%	68%	76%		63%	73%	63%	61%	66%	63%
Pashto_M	78%	63%	68%	73%	71%	68%	66%	73%	73%	73%	76%	80%	76%	68%	63%		68%	78%	78%	76%	76%
Persian_F	71%	61%	83%	85%	78%	80%	61%	71%	83%	71%	71%	80%	66%	71%	73%	68%		83%	54%	61%	76%
Persian_M	73%	61%	73%	80%	73%	71%	68%	71%	88%	66%	80%	76%	66%	63%	63%	78%	83%		61%	73%	78%
French_F	61%	59%	59%	61%	68%	63%	71%	66%	61%	73%	73%	68%	78%	76%	61%	78%	54%	61%		83%	66%
French_M	68%	54%	61%	66%	73%	66%	71%	78%	76%	68%	76%	76%	78%	66%	66%	76%	61%	73%	83%		71%
Italian_F	73%	66%	80%	83%	83%	73%	59%	68%	76%	68%	71%	80%	66%	59%	63%	76%	76%	78%	66%	71%	
Italian_M	85%	73%	85%	85%	78%	76%	61%	61%	73%	71%	76%	76%	63%	59%	76%	73%	71%	80%	66%	71%	88%
Portuguese_F	76%	71%	80%	80%	73%	76%	59%	66%	68%	63%	66%	76%	59%	66%	83%	66%	68%	73%	66%	76%	76%
Portuguese_M	78%	71%	83%	88%	83%	83%	61%	73%	76%	71%	73%	85%	61%	68%	78%	80%	85%	83%	63%	68%	80%
Spanish_F	88%	68%	83%	88%	88%	80%	63%	71%	71%	76%	71%	76%	66%	59%	73%	73%	78%	83%	63%	73%	88%
Spanish_M	83%	66%	80%	83%	76%	78%	71%	73%	71%	71%	76%	71%	66%	59%	78%	71%	73%	83%	63%	63%	78%
Japanese_F	71%	66%	76%	78%	73%	71%	54%	61%	83%	73%	71%	68%	61%	71%	78%	68%	80%	80%	61%	63%	73%
Japanese_M	76%	66%	80%	83%	73%	76%	49%	61%	80%	71%	66%	73%	68%	66%	83%	59%	85%	80%	54%	63%	76%

Language	Russian_F	Russian_M	Slovak_F	Slovak_M	Ukrainian_F	Ukrainian_M	German_F	German_M	Norwegian_F	Norwegian_M	Swedish_F	Yiddish_M	Hindi_F	Hindi_M	Oriya_F	Pashto_M	Persian_F	Persian_M	French_F	French_M	Italian_F
Korean_F	73%	54%	68%	66%	78%	76%	59%	66%	71%	71%	61%	61%	61%	51%	66%	56%	76%	61%	54%	61%	63%
Korean_M1	66%	51%	66%	73%	68%	61%	59%	63%	68%	63%	66%	71%	61%	61%	61%	68%	76%	76%	63%	76%	85%
Korean_M2	66%	68%	63%	63%	63%	63%	51%	59%	71%	68%	66%	66%	66%	68%	68%	71%	66%	61%	63%	63%	59%
Thai_F	66%	61%	61%	68%	68%	61%	68%	68%	73%	63%	73%	73%	68%	63%	59%	78%	66%	73%	78%	80%	76%
Thai_M	80%	56%	73%	71%	83%	76%	59%	73%	73%	68%	63%	76%	83%	51%	63%	71%	73%	71%	56%	73%	80%
Xhosa_F	76%	61%	78%	78%	78%	78%	59%	68%	73%	66%	61%	73%	61%	54%	59%	63%	80%	78%	56%	61%	76%
Xhosa_M	71%	61%	76%	76%	80%	73%	51%	66%	76%	66%	66%	78%	61%	51%	66%	68%	78%	76%	54%	63%	88%
Yoruba_M	68%	61%	71%	76%	76%	66%	59%	66%	76%	66%	73%	73%	56%	73%	76%	68%	73%	80%	68%	83%	83%
Mandarin_F	71%	59%	68%	71%	80%	73%	76%	76%	88%	71%	73%	76%	73%	56%	61%	73%	80%	80%	73%	78%	76%
Mandarin_M	71%	61%	78%	76%	78%	73%	54%	63%	73%	61%	63%	73%	71%	63%	73%	61%	83%	71%	61%	61%	80%
Turkish_F	76%	68%	76%	76%	83%	80%	76%	83%	88%	78%	83%	80%	63%	63%	63%	80%	76%	78%	73%	80%	73%
Turkish_M	80%	66%	76%	78%	76%	83%	76%	73%	83%	76%	68%	68%	68%	63%	76%	68%	83%	71%	61%	66%	63%
Finnish_M	78%	73%	78%	78%	78%	78%	63%	68%	83%	76%	76%	76%	68%	56%	76%	73%	71%	71%	66%	71%	78%
Hungarian_F	68%	73%	73%	73%	78%	78%	66%	83%	80%	73%	78%	80%	63%	66%	66%	73%	73%	76%	68%	76%	73%
Hungarian_M	73%	68%	85%	80%	80%	88%	71%	80%	80%	76%	76%	80%	59%	68%	71%	73%	83%	76%	63%	68%	73%

Language	Italian_M	Portuguese_F	Portuguese_M	Spanish_F	Spanish_M	Japanese_F	Japanese_M	Korean_F	Korean_M1	Korean_M2	Thai_F	Thai_M	Xhosa_F	Xhosa_M	Yoruba_M	Mandarin_F	Mandarin_M	Turkish_F	Turkish_M	Finnish_M	Hungarian_F	Hungarian_M
Amharic_F	76%	68%	73%	66%	66%	73%	76%	68%	61%	63%	54%	59%	63%	71%	56%	61%	59%	66%	66%	63%	61%	59%
Amharic_M	85%	78%	78%	80%	78%	88%	85%	71%	71%	71%	66%	73%	73%	88%	78%	68%	78%	73%	71%	78%	71%	68%
Egyptian Arabic_M	68%	61%	59%	73%	78%	66%	59%	49%	54%	49%	51%	54%	59%	56%	61%	54%	54%	54%	56%	63%	59%	61%
Hausa_M	73%	76%	66%	68%	73%	78%	85%	61%	63%	68%	61%	66%	68%	73%	71%	63%	71%	63%	63%	68%	66%	59%
Hebrew_F	90%	85%	83%	90%	90%	76%	80%	63%	73%	66%	68%	68%	66%	78%	78%	68%	76%	73%	66%	80%	71%	68%
Hebrew_M	90%	93%	90%	85%	83%	83%	88%	66%	68%	63%	66%	68%	76%	78%	71%	71%	76%	76%	73%	76%	73%	68%
Somali_M	80%	73%	71%	80%	76%	90%	83%	68%	66%	68%	73%	66%	66%	73%	76%	68%	76%	63%	71%	76%	59%	63%
Vietnamese_F	78%	68%	68%	68%	76%	73%	73%	71%	63%	71%	73%	68%	61%	68%	61%	76%	68%	78%	80%	80%	68%	71%
Vietnamese_M	83%	73%	73%	78%	83%	73%	71%	66%	71%	73%	78%	73%	63%	73%	73%	76%	68%	83%	76%	80%	68%	66%
Fijian_F	85%	68%	68%	80%	76%	88%	80%	61%	63%	66%	63%	61%	61%	68%	76%	61%	66%	61%	63%	83%	63%	68%
Fijian_M	80%	76%	76%	88%	76%	85%	78%	66%	80%	63%	73%	68%	68%	78%	90%	66%	73%	68%	66%	78%	68%	63%
Telugu_F	83%	76%	73%	73%	73%	80%	73%	59%	63%	68%	63%	63%	59%	71%	73%	61%	73%	66%	66%	73%	66%	66%
Telugu_M	80%	85%	83%	78%	80%	76%	80%	59%	66%	61%	61%	63%	71%	73%	68%	71%	68%	73%	71%	71%	71%	68%
Bulgarian_F	83%	78%	88%	80%	76%	68%	71%	61%	66%	56%	59%	68%	73%	73%	68%	66%	71%	73%	68%	71%	78%	80%
Bulgarian_M	76%	73%	83%	85%	73%	71%	73%	71%	71%	54%	66%	73%	78%	73%	71%	71%	78%	68%	68%	61%	71%	76%
Croatian_F	80%	83%	78%	76%	83%	78%	83%	66%	73%	68%	71%	63%	66%	73%	71%	73%	76%	80%	78%	73%	71%	71%
Croatian_M	80%	83%	78%	80%	80%	80%	88%	66%	68%	63%	61%	68%	73%	71%	68%	68%	71%	73%	76%	68%	68%	68%
Czech_F	76%	73%	85%	76%	78%	71%	76%	66%	73%	68%	71%	78%	78%	76%	71%	80%	76%	88%	80%	73%	78%	78%
Czech_M	73%	71%	83%	78%	68%	80%	80%	66%	73%	71%	66%	73%	78%	78%	73%	78%	76%	78%	76%	68%	71%	73%
Polish_F	76%	80%	88%	85%	76%	71%	76%	71%	68%	59%	68%	78%	80%	76%	76%	76%	76%	76%	73%	68%	73%	78%
Polish_M	78%	76%	83%	83%	73%	78%	80%	63%	73%	59%	66%	73%	76%	80%	76%	71%	78%	73%	68%	68%	73%	71%
Russian_F	85%	76%	78%	88%	83%	71%	76%	73%	66%	66%	66%	80%	76%	71%	68%	71%	71%	76%	80%	78%	68%	73%
Russian_M	73%	71%	71%	68%	66%	66%	66%	54%	51%	68%	61%	56%	61%	61%	61%	59%	61%	68%	66%	73%	73%	68%
Slovak_F	85%	80%	83%	83%	80%	76%	80%	68%	66%	63%	61%	73%	78%	76%	71%	68%	78%	76%	76%	78%	73%	85%

Language	Italian_M	Portuguese_F	Portuguese_M	Spanish_F	Spanish_M	Japanese_F	Japanese_M	Korean_F	Korean_M1	Korean_M2	Thai_F	Thai_M	Xhosa_F	Xhosa_M	Yoruba_M	Mandarin_F	Mandarin_M	Turkish_F	Turkish_M	Finnish_M	Hungarian_F	Hungarian_M
Slovak_M	85%	80%	88%	88%	83%	78%	83%	66%	73%	63%	68%	71%	78%	76%	76%	71%	76%	76%	78%	78%	73%	80%
Ukrainian_F	78%	73%	83%	88%	76%	73%	73%	78%	68%	63%	68%	83%	78%	80%	76%	80%	78%	83%	76%	78%	78%	80%
Ukrainian_M	76%	76%	83%	80%	78%	71%	76%	76%	61%	63%	61%	76%	78%	73%	66%	73%	73%	80%	83%	78%	78%	88%
German_F	61%	59%	61%	63%	71%	54%	49%	59%	59%	51%	68%	59%	59%	51%	59%	76%	54%	76%	76%	63%	66%	71%
German_M	61%	66%	73%	71%	73%	61%	61%	66%	63%	59%	68%	73%	68%	66%	66%	76%	63%	83%	73%	68%	83%	80%
Norwegian_F	73%	68%	76%	71%	71%	83%	80%	71%	68%	71%	73%	73%	73%	76%	76%	88%	73%	88%	83%	83%	80%	80%
Norwegian_M	71%	63%	71%	76%	71%	73%	71%	71%	63%	68%	63%	68%	66%	66%	66%	71%	61%	78%	76%	76%	73%	76%
Swedish_F	76%	66%	73%	71%	76%	71%	66%	61%	66%	66%	73%	63%	61%	66%	73%	73%	63%	83%	68%	76%	78%	76%
Yiddish_M	76%	76%	85%	76%	71%	68%	73%	61%	71%	66%	73%	76%	73%	78%	73%	76%	73%	80%	68%	76%	80%	80%
Hindi_F	63%	59%	61%	66%	66%	61%	68%	61%	61%	66%	68%	83%	61%	61%	56%	73%	71%	63%	68%	68%	63%	59%
Hindi_M	59%	66%	68%	59%	59%	71%	66%	51%	61%	68%	63%	51%	54%	51%	73%	56%	63%	63%	63%	56%	66%	68%
Oriya_F	76%	83%	78%	73%	78%	78%	83%	66%	61%	68%	59%	63%	59%	66%	76%	61%	73%	63%	76%	76%	66%	71%
Pashto_M	73%	66%	80%	73%	71%	68%	59%	56%	68%	71%	78%	71%	63%	68%	68%	73%	61%	80%	68%	73%	73%	73%
Persian_F	71%	68%	85%	78%	73%	80%	85%	76%	76%	66%	66%	73%	80%	78%	73%	80%	83%	76%	83%	71%	73%	83%
Persian_M	80%	73%	83%	83%	83%	80%	80%	61%	76%	61%	73%	71%	78%	76%	80%	80%	71%	78%	71%	71%	76%	76%
French_F	66%	66%	63%	63%	63%	61%	54%	54%	63%	63%	78%	56%	56%	54%	68%	73%	61%	73%	61%	66%	68%	63%
French_M	71%	76%	68%	73%	63%	63%	63%	61%	76%	63%	80%	73%	61%	63%	83%	78%	61%	80%	66%	71%	76%	68%
Italian_F	88%	76%	80%	88%	78%	73%	76%	63%	85%	59%	76%	80%	76%	88%	83%	76%	80%	73%	63%	78%	73%	73%
Italian_M		88%	85%	93%	85%	80%	83%	66%	71%	63%	71%	73%	73%	78%	78%	71%	73%	73%	68%	85%	73%	73%
Portuguese_F	88%		85%	80%	78%	73%	85%	59%	68%	63%	59%	73%	71%	73%	78%	68%	73%	68%	66%	71%	76%	63%
Portuguese_M	85%	85%		85%	80%	73%	80%	61%	68%	59%	63%	78%	83%	83%	76%	73%	78%	76%	73%	73%	80%	78%
Spanish_F	93%	80%	85%		88%	83%	85%	73%	76%	63%	71%	80%	80%	85%	80%	73%	80%	73%	71%	78%	71%	71%
Spanish_M	85%	78%	80%	88%		78%	78%	66%	66%	63%	66%	66%	68%	71%	76%	68%	73%	73%	71%	80%	68%	73%
Japanese_F	80%	73%	73%	83%	78%		93%	73%	66%	78%	66%	68%	71%	76%	78%	63%	73%	71%	78%	76%	63%	71%
Japanese_M	83%	85%	80%	85%	78%	93%		76%	73%	73%	63%	73%	76%	78%	78%	68%	80%	68%	76%	73%	68%	68%

Language	Italian_M	Portuguese_F	Portuguese_M	Spanish_F	Spanish_M	Japanese_F	Japanese_M	Korean_F	Korean_M1	Korean_M2	Thai_F	Thai_M	Xhosa_F	Xhosa_M	Yoruba_M	Mandarin_F	Mandarin_M	Turkish_F	Turkish_M	Finnish_M	Hungarian_F	Hungarian_M
Korean_F	66%	59%	61%	73%	66%	73%	76%		71%	71%	63%	71%	63%	71%	59%	73%	71%	73%	80%	76%	56%	68%
Korean_M1	71%	68%	68%	76%	66%	66%	73%	71%		61%	78%	80%	68%	76%	78%	76%	78%	73%	66%	66%	63%	59%
Korean_M2	63%	63%	59%	63%	63%	78%	73%	71%	61%		66%	68%	66%	63%	66%	59%	66%	71%	78%	71%	63%	66%
Thai_F	71%	59%	63%	71%	66%	66%	63%	63%	78%	66%		78%	71%	68%	71%	83%	73%	78%	68%	66%	66%	63%
Thai_M	73%	73%	78%	80%	66%	68%	73%	71%	80%	68%	78%		76%	85%	78%	80%	85%	80%	68%	63%	71%	66%
Xhosa_F	73%	71%	83%	80%	68%	71%	76%	63%	68%	66%	71%	76%		93%	71%	76%	83%	76%	68%	61%	68%	71%
Xhosa_M	78%	73%	83%	85%	71%	76%	78%	71%	76%	63%	68%	85%	93%		78%	78%	85%	73%	66%	66%	68%	66%
Yoruba_M	78%	78%	76%	80%	76%	78%	78%	59%	78%	66%	71%	78%	71%	78%		63%	80%	66%	66%	73%	71%	68%
Mandarin_F	71%	68%	73%	73%	68%	63%	68%	73%	76%	59%	83%	80%	76%	78%	63%		76%	85%	80%	68%	73%	71%
Mandarin_M	73%	73%	78%	80%	73%	73%	80%	71%	78%	66%	73%	85%	83%	85%	80%	76%		71%	68%	63%	63%	66%
Turkish_F	73%	68%	76%	73%	73%	71%	68%	73%	73%	71%	78%	80%	76%	73%	66%	85%	71%		88%	78%	83%	83%
Turkish_M	68%	66%	73%	71%	71%	78%	76%	80%	66%	78%	68%	68%	68%	66%	66%	80%	68%	88%		80%	76%	83%
Finnish_M	85%	71%	73%	78%	80%	76%	73%	76%	66%	71%	66%	63%	61%	66%	73%	68%	63%	78%	80%		78%	85%
Hungarian_F	73%	76%	80%	71%	68%	63%	68%	56%	63%	63%	66%	71%	68%	68%	71%	73%	63%	83%	76%	78%		90%
Hungarian_M	73%	63%	78%	71%	73%	71%	68%	68%	59%	66%	63%	66%	71%	66%	68%	71%	66%	83%	83%	85%	90%	

Appendix D- The Questions in the Game

Lang.	Type.	Ques. Nom.	Option A	% Similarity	Option B	% Similarity	Gap similiarities
Amharic	F	1	Japanese_M	76	German_M	56	20
		2	Slovak_M	76	French_M	51	25
		3	Croatian_M	73	E.Arabic_M	49	24
	M	4	Japanese_F	88	Thai_F	66	22
		5	Italian_F	85	Czech_F	66	19
		6	Telugu_F	83	Hindi_F	61	22
E.Arabic	M	7	Spanish_F	73	Hindi_F	51	22
		8	Hebrew_F	73	Amharic_F	49	24
		9	Italian_F	68	Korean_F	49	19
Hausa	M	10	Oriya_F	80	Bulgarian_F	63	17
		11	Japanese_F	78	Russian_F	61	17
		12	Portuguese_F	76	Hindi_F	56	20
Hebrew	F	13	Italian_M	90	Korean_M2	66	24
		14	Spanish_M	90	Turkish_M	66	24
		15	Portuguese_M	83	Hindi_M	59	24
	M	16	Portuguese_F	90	Korean_F	66	24
		17	Croatian_F	85	Thai_F	66	19
		18	Czech_F	85	German_F	61	24
Somali	M	19	Japanese_F	90	Xhosa_F	66	24
		20	Fijian_F	85	Turkish_F	63	22
		21	Telugu_F	78	Hungarian_F	59	19
Vietnamese	F	22	Vietnamese_M	90	Slovak_M	71	19
		23	Turkish_M	80	Yoruba_M	61	19
		24	Finnish_M	80	Norwegian_M	59	21
	M	25	Hebrew_F	85	Amharic_F	66	19
		26	Czech_F	83	Bulgarian_F	66	17
		27	Turkish_F	83	Xhosa_F	63	20

Lang.	Type.	Ques. Nom.	Option A	% Similarity	Option B	% Similarity	Gap similiarities
Fijian	F	28	Italian_M	85	Mandarin_M	66	19
		29	Finnish_M	83	Korean_M1	63	20
		30	Amharic_M	80	Thai_M	61	19
	M	31	Japanese_F	85	Mandarin_F	66	19
		32	Italian_F	85	French_F	63	22
		33	Hebrew_F	80	Amharic_F	59	21
Telugu	F	34	Amharic_M	85	Thai_M	63	22
		35	Fijian_M	73	French_M	59	14
		36	Croatian_M	73	German_M	54	19
	M	37	Croatian_F	88	Vietnamese_F	66	22
		38	Portuguese_F	85	Thai_F	61	24
		39	Polish_F	80	Korean_F	59	21
Bulgarian	F	40	Portuguese_M	90	Pashto_M	71	19
		41	Slovak_M	80	French_M	61	19
		42	Hungarian_M	80	Korean_M2	56	24
	M	43	Slovak_F	90	Turkish_F	68	22
		44	Portuguese_F	80	Vietnamese_F	61	19
		45	Hungarian_F	80	Amharic_F	59	21
Croatian	F	46	Telugu_M	88	Norwegian_M	68	20
		47	Japanese_M	83	Czech_M	68	15
		48	Amharic_M	80	French_M	63	17
	M	49	Portuguese_F	83	Fijian_F	66	17
		50	Polish_F	83	Thai_F	61	22
		51	Bulgarian_F	78	Hindi_F	54	24
Czech	F	52	Yiddish_M	85	Norwegian_M	68	17
		53	Hebrew_M	85	Korean_M2	68	17
		54	Vietnamese_M	83	E.Arabic_M	63	20
	M	55	Persian_F	85	Korean_F	66	19
		56	Slovak_F	83	Thai_F	66	17
		57	Japanese_F	80	Swedish_F	61	19

Lang.	Type.	Ques. Nom.	Option A	% Similarity	Option B	% Similarity	Gap similiarities
Polish	F	58	Ukrainian_M	85	Russian_M	66	19
		59	Portuguese_M	88	Somali_M	66	22
		60	Telugu_M	80	Korean_M2	59	21
	M	61	Spanish_F	83	Vietnamese_F	63	20
		62	Czech_F	80	Korean_F	63	17
		63	Persian_F	78	German_F	59	19
Russian	F	64	Ukrainian_M	88	Polish_M	73	15
		65	Hebrew_M	83	E.Arabic_M	63	20
		66	Vietnamese_M	80	Hausa_M	61	19
	M	67	Bulgarian_F	76	Swedish_F	59	17
		68	Croatian_F	73	German_F	56	17
		69	Hungarian_F	73	Hindi_F	56	17
Slovak	F	70	Ukrainian_M	88	Russian_M	68	20
		71	Italian_M	85	Fijian_M	68	17
		72	Hungarian_M	85	Korean_M2	63	22
	M	73	Slovak_F	90	Swedish_F	71	19
		74	Spanish_F	88	Mandarin_F	71	17
		75	Persian_F	85	Telugu_F	68	17
Ukrainian	F	76	Bulgarian_M	85	Russian_M	68	17
		77	Yiddish_M	83	Hausa_M	63	20
		78	German_M	80	E.Arabic_M	61	19
	M	79	Slovak_F	88	Oriya_F	68	20
		80	Polish_F	85	Telugu_F	63	22
		81	Norwegian_F	83	Thai_F	61	22
German	F	82	Turkish_M	76	Mandarin_M	54	22
		83	French_M	71	Xhosa_M	51	20
		84	Hungarian_M	71	Japanese_M	49	22
	M	85	Hungarian_F	83	Croatian_F	66	17
		86	Turkish_F	83	Japanese_F	61	22
		87	Swedish_F	80	Fijian_F	56	24

Lang.	Type.	Ques. Nom.	Option A	% Similarity	Option B	% Similarity	Gap similiarities
Norwegian	F	88	Persian_M	88	Bulgarian_M	68	20
		89	Turkish_M	83	Russian_M	63	20
		90	Finnish_M	83	Hindi_M	63	20
	M	91	Turkish_F	78	Portuguese_F	63	15
		92	Ukrainian_F	78	Oriya_F	63	15
		93	Swedish_F	76	Vietnamese_F	59	17
Swedish	F	94	Persian_M	80	Mandarin_M	63	17
		95	German_M	80	Czech_M	61	19
		96	Vietnamese_M	78	Russian_M	59	19
Yiddish	M	97	Czech_F	85	Telugu_F	66	19
		98	Ukrainian_F	83	Fijian_F	61	22
		99	Hungarian_F	80	Korean_F	61	19
Hindi	F	100	Thai_M	83	Hungarian_M	59	24
		101	French_M	78	Russian_M	56	22
		102	Somali_M	76	Croatian_M	54	22
	M	103	French_F	76	Mandarin_F	56	20
		104	Oriya_F	76	Xhosa_F	54	22
		105	Japanese_F	71	German_F	51	20
Oriya	F	106	Japanese_M	83	Russian_M	63	20
		107	Hausa_M	80	Korean_M1	61	19
		108	Spanish_M	78	German_M	54	24
Pashto	M	109	Turkish_F	80	Xhosa_F	63	17
		110	Russian_F	78	Amharic_F	56	22
		111	Czech_F	78	Korean_F	56	22
Persian	F	112	Portuguese_M	85	Pashto_M	68	17
		113	Slovak_M	85	Korean_M2	66	19
		114	Bulgarian_M	85	French_M	61	24
	M	115	Norwegian_F	88	Telugu_F	66	22
		116	Spanish_F	83	Oriya_F	63	20
		117	Swedish_F	80	Korean_F	61	19

Lang.	Type.	Ques. Nom.	Option A	% Similarity	Option B	% Similarity	Gap similiarities
French	F	118	French_M	83	Portuguese_M	63	20
		119	Pashto_M	78	Japanese_M	54	24
		120	Hindi_M	76	Xhosa_M	54	22
	M	121	Turkish_F	80	Slovak_F	61	19
		122	Thai_F	80	Xhosa_F	61	19
		123	Hindi_F	78	Bulgarian_F	59	19
Italian	F	124	Xhosa_M	88	E.Arabic_M	68	20
		125	Fijian_M	85	Hausa_M	66	19
		126	Korean_M1	85	Turkish_M	63	22
	M	127	Spanish_F	93	Norwegian_F	73	20
		128	Hebrew_F	90	French_F	66	24
		129	Portuguese_F	88	Hindi_F	63	25
Portuguese	F	130	Hebrew_M	93	Somali_M	73	20
		131	Italian_M	88	Turkish_M	66	22
		132	Japanese_M	85	Hungarian_M	63	22
	M	133	Bulgarian_F	88	Fijian_F	68	20
		134	Polish_F	88	Thai_F	63	25
		135	Spanish_F	85	French_F	63	22
Spanish	F	136	Italian_M	93	German_M	71	22
		137	Xhosa_M	85	Hausa_M	68	17
		138	Polish_M	83	Korean_M2	63	20
	M	139	Spanish_F	88	Hungarian_F	68	20
		140	Croatian_F	83	Mandarin_F	68	15
		141	Russian_F	83	Hindi_F	66	17
Japanese	F	142	Japanese_M	93	Vietnamese_M	73	20
		143	Somali_M	90	Hungarian_M	71	19
		144	Amharic_M	88	Thai_M	68	20
	M	145	Spanish_F	85	Thai_F	63	22
		146	Portuguese_F	85	Swedish_F	65	20
		147	Czech_F	76	French_F	54	22

Lang.	Type.	Ques. Nom.	Option A	% Similarity	Option B	% Similarity	Gap similiarities
Korean	F	148	Turkish_M	80	Yiddish_M	61	19
		149	Finnish_M	76	Russian_M	54	22
		150	Japanese_M	76	Hindi_M	51	25
	M1	151	Italian_F	85	Vietnamese_F	63	22
		152	Thai_F	78	Oriya_F	61	17
		153	Mandarin_F	76	German_F	59	17
	M2	154	Japanese_F	78	Mandarin_F	59	19
		155	Norwegian_F	71	Bulgarian_F	56	15
		156	Vietnamese_F	71	German_F	51	20
Thai	F	157	French_M	80	Hungarian_M	63	17
		158	Vietnamese_M	78	Croatian_M	61	17
		159	Pashto_M	78	Ukrainian_M	61	17
	M	160	Hindi_F	83	Swedish_F	63	20
		161	Ukrainian_F	83	Croatian_F	63	20
		162	Mandarin_F	80	Fijian_F	61	19
Xhosa	F	163	Portuguese_M	83	Vietnamese_M	63	20
		164	Mandarin_M	83	Russian_M	61	22
		165	Bulgarian_M	78	Hindi_M	54	24
	M	166	Italian_F	88	Hungarian_F	66	22
		167	Spanish_F	85	Swedish_F	66	19
		168	Hebrew_F	78	French_F	54	24
Yoruba	M	169	Italian_F	83	Mandarin_F	63	20
		170	Spanish_F	80	Korean_F	59	21
		171	Japanese_F	78	Amharic_F	56	22
Mandarin	F	172	Persian_M	80	Russian_M	59	21
		173	Thai_M	80	Korean_M2	59	21
		174	Turkish_M	80	Hindi_M	56	24
	M	175	Xhosa_F	83	Hungarian_F	63	20
		176	Persian_F	83	Swedish_F	63	20
		177	Ukrainian_F	78	Amharic_F	59	19

Lang.	Type.	Ques. Nom.	Option A	% Similarity	Option B	% Similarity	Gap similiarities
Turkish	F	178	Vietnamese_M	83	Yoruba_M	66	17
		179	Pashto_M	80	Somali_M	63	17
		180	Finnish_M	78	E.Arabic_M	54	24
	M	181	Norwegian_F	83	Hebrew_F	66	17
		182	Korean_F	80	Italian_F	63	17
		183	Russian_F	80	Fijian_F	63	17
Finnish	M	184	Norwegian_F	83	Amharic_F	63	20
		185	Fijian_F	83	German_F	63	20
		186	Hebrew_F	80	Xhosa_F	61	19
Hungarian	F	187	Hungarian_M	90	Hindi_M	66	24
		188	German_M	83	Somali_M	59	24
		189	Yiddish_M	80	E.Arabic_M	59	21
	M	190	Slovak_F	85	French_F	63	22
		191	Turkish_F	83	Hindi_F	59	24
		192	Persian_F	83	Amharic_F	59	24

Appendix E- The Results of the Main Experiment

No.	Language	Question 1				Question 2				Question 3			
		Similar	**Dissimilar**	**N**	**p**	**Similar**	**Dissimilar**	**N**	**p**	**Similar**	**Dissimilar**	**N**	**p**
1	**Amharic_F**	**Japanese_M**	German_M			**Slovak_M**	French_M			Croatian_M	**Egyptian_Arabic_M**		
		69.4%**	30.6%	36	<0.01	**75.0%****	25.0%	32	<0.01	21.9%	**78.1%*****	32	<0.001
2	**Amharic_M**	Japanese_F	Thai_F			Italian_F	Czech_F			Telugu_F	Hindi_F		
		46.9%	53.1%	49	0.28	49.0%	51.0%	49	0.39	52.3%	47.7%	65	0.31
3	**E.Arabic_M**	Spanish_F	**Hindi_F**							Italian_F	**Korean_F**		
		29.0%	**71.0%****	31	<0.01					22.6%	**77.4%*****	31	<0.001
4	**Hausa_M**	**Oriya_F**	Bulgarian_F			**Japanese_F**	Russian_F			Portuguese_F	**Hindi_F**		
		87.1%**	12.9%	31	<0.001	**68.6%****	31.4%	35	<0.01	25.0%	**75.0%****	28	<0.01
5	**Hebrew_F**	**Italian_M**	Korean_M2			Spanish_M	**Turkish_M**			Portuguese_M	Hindi_M		
		86.2%***	13.8%	29	<0.001	23.4%	**76.6%*****	47	<0.001	51.6%	48.4%	31	0.36
6	**Hebrew_M**	Portuguese_F	Korean_F			Croatian_F	Thai_F			**Czech_F**	German_F		
		59.0%	41.0%	39	0.1	56.0%	44.0%	25	0.21	**76.0%*****	24.0%	50	<0.001
7	**Somali_M**	Japanese_F	Xhosa_F			**Fijian_F**	Turkish_F			**Telugu_F**	Hungarian_F		
		45.3%	54.7%	64	0.19	**61.9%***	38.1%	42	<0.05	**83.0%*****	17.0%	53	<0.001
8	**Vietnamese_F**	**Vietnamese_M**	Slovak_M			Turkish_M	Yoruba_M			Finnish_M	Norwegian_M		
		79.3%***	20.7%	29	<0.001	52.3%	47.7%	65	0.31	42.3%	57.7%	26	0.16
9	**Vietnamese_M**					**Czech_F**	Bulgarian_F			Turkish_F	**Xhosa_F**		
						64.1%*	35.9%	39	<0.05	25.5%	**74.5%*****	51	<0.001
10	**Fijian_F**	**Italian_M**	Mandarin_M			**Finnish_M**	Korean_M			**Amharic_M**	Thai_M		
		69.2%*	30.8%	26	<0.05	**82.2%*****	17.8%	45	<0.001	**72.7%****	27.3%	33	<0.01
11	**Fijian_M**	**Japanese_F**	Mandarin_F			**Italian_F**	French_F						
		81.3%***	18.8%	32	<0.001	**81.8%*****	18.2%	44	<0.001				
12	**Telugu_F**	**Amharic_M**	Thai_M			**Fijian_M**	French_M			**Croatian_M**	German_M		
		65.8%*	34.2%	38	<0.05	**87.5%*****	12.5%	24	<0.001	**78.9%*****	21.1%	38	<0.001
13	**Telugu_M**	Croatian_F	**Vietnamese_F**			Portuguese_F	Thai_F			Polish_F	**Korean_F**		
		14.8%	**85.2%*****	61	<0.001	45.2%	54.8%	42	0.22	27.8%	**88.9%****	36	<0.01

No.	Language	Question 1				Question 2				Question 3			
		Similar	Dissimilar	N	p	Similar	Dissimilar	N	P	Similar	Dissimilar	N	p
14	**Bulgarian_F**	**Portuguese_M**	Pashto_M			**Slovak_M**	French_M			**Hungarian_M**	Korean_M2		
		75.9%**	24.1%	29	<0.01	**83.3%*****	16.7%	36	<0.001	**81.8%*****	18.2%	55	<0.001
15	**Bulgarian_M**	**Slovak_F**	Turkish_F			Portuguese_F	Vietnamese_F			**Hungarian_F**	Amharic_F		
		90.0%***	10.0%	40	<0.001	53.3%	46.7%	45	0.28	**71.9%*****	28.1%	57	<0.001
16	**Croatian_F**	Telugu_M	**Norwegian_M**			Japanese_M	**Czech_M**			**Amharic_M**	French_M		
		19.0%	**81.0%*****	42	<0.001	15.4%	**84.6%*****	39	<0.001	**73.3%****	26.7%	30	<0.01
17	**Croatian_M**	Portuguese_F	**Fijian_F**			**Polish_F**	Thai_F			**Bulgarian_F**	Hindi_F		
		29.6%	**70.4%****	27	<0.01	**100.0%*****	0.0%	40	<0.001	**86.5%*****	13.5%	37	<0.001
18	**Czech_F**	Yiddish_M	**Norwegian_M**							**Vietnamese_M**	Egyptian_Arabic_M		
		23.3%	**76.7%*****	30	<0.001					**74.3%*****	25.7%	35	<0.001
19	**Czech_M**	**Persian_F**	Korean_F			**Slovak_F**	Thai_F			Japanese_F	Swedish_F		
		87.9%***	12.1%	33	<0.001	**100.0%*****	0.0%	30	<0.001	48.4%	51.6%	31	0.36
20	**Polish_F**	Ukrainian_M	Russian_M			**Portuguese_M**	Somali_M			**Telugu_M**	Korean_M2		
		41.4%	58.6%	58	0.07	**88.9%*****	11.1%	36	<0.001	**75.0%****	25.0%	24	<0.01
21	**Polish_M**	Spanish_F	Vietnamese_F			**Czech_F**	Korean_F			Persian_F	German_F		
		51.4%	48.6%	35	0.37	**97.3%*****	2.7%	37	<0.001	60.5%	39.5%	38	0.07
22	**Russian_F**	Ukrainian_M	Polish_M							Vietnamese_M	Hausa_M		
		57.4%	42.6%	54	0.11					40.0%	60.0%	35	0.09
23	**Russian_M**	**Bulgarian_F**	Swedish_F			**Croatian_F**	German_F			**Hungarian_F**	Hindi_F		
		96.7%***	3.3%	30	<0.001	**91.9%*****	8.1%	37	<0.001	**87.2%*****	12.8%	47	<0.001
24	**Slovak_F**	**Ukrainian_M**	Russian_M			Italian_M	Fijian_M			**Hungarian_M**	Korean_M2		
		67.6%*	32.4%	37	<0.05	51.1%	48.9%	45	0.38	**89.7%*****	10.3%	29	<0.001
25	**Slovak_M**	**Slovak_F**	Swedish_F			**Spanish_F**	Mandarin_F			**Persian_F**	Telugu_F		
		77.4%***	22.6%	31	<0.001	**74.4%*****	25.6%	39	<0.001	**89.2%*****	10.8%	37	<0.001
26	**Ukrainian_F**	Bulgarian_M	**Russian_M**			Yiddish_M	Hausa_M			**German_M**	Egyptian_Arabic_M		
		29.3%	**70.7%*****	58	<0.001	57.6%	42.4%	33	0.15	**80.0%*****	20.0%	35	<0.001
27	**Ukrainian_M**	**Slovak_F**	Oriya_F			**Polish_F**	Telugu_F			**Norwegian_F**	Thai_F		
		100.0%***	0.0%	33	<0.001	**91.9%*****	8.1%	37	<0.001	**92.6%*****	7.4%	27	<0.001

No.	Language	Question 1				Question 2				Question 3			
		Similar	Dissimilar	N	p	Similar	Dissimilar	N	P	Similar	Dissimilar	N	p
28	**German_F**	**Turkish_M**	Mandarin_M			**French_M**	Xhosa_M			**Hungarian_M**	Japanese_M		
		82.6%* **	13.8%	29	<0.001	**71.9%*** **	28.1%	64	<0.001	**72.5%** **	27.5%	40	<0.01
29	**German_M**	Hungarian_F	Croatian_F			**Turkish_F**	Japanese_F			**Swedish_F**	Fijian_F		
		52.1%	47.9%	48	0.33	**90.3%*** **	9.7%	31	<0.001	**95.3%*** **	4.7%	64	<0.001
30	**Norwegian_F**	**Persian_M**	Bulgarian_M			**Turkish_M**	Russian_M			Finnish_M	Hindi_M		
		75.5%* **	24.5%	49	<0.001	**83.8%*** **	16.2%	37	<0.001	50.0%	50.0%	64	0.45
31	**Norwegian_M**	**Turkish_F**	Portuguese_F			Ukrainian_F	**Oriya_F**			**Swedish_F**	Vietnamese_F		
		64.5%*	35.5%	31	<0.05	26.1%	**73.9%** **	23	<0.01	**65.5%** **	34.5%	55	<0.01
32	**Swedish_F**	**Persian_M**	Mandarin_M			**German_M**	Czech_M			**Vietnamese_M**	Russian_M		
		82.9%* **	17.1%	35	<0.001	**80.5%*** **	19.5%	41	<0.001	**66.7%***	33.3%	30	<0.05
33	**Yiddish_M**	**Czech_F**	Telugu_F			Ukrainian_F	Fijian_F			Hungarian_F	Korean_F		
		71.9% **	28.1%	32	<0.01	52.4%	47.6%	63	0.31	62.5%	37.5%	24	0.08
34	**Hindi_F**	Thai_M	Hungarian_M			French_M	Russian_M			Somali_M	Croatian_M		
		50.0%	50.0%	50	0.44	46.2%	53.8%	39	0.26	51.6%	48.4%	31	0.36
35	**Hindi_M**	French_F	Mandarin_F			**Oriya_F**	Xhosa_F			**Japanese_F**	German_F		
		58.8%	41.2%	34	0.11	**77.4%*** **	22.6%	31	<0.001	**75.8%*** **	24.2%	33	<0.001
36	**Oriya_F**	**Japanese_M**	Russian_M			Hausa_M	**Korean_M1**			Spanish_M	German_M		
		84.1%* **	15.9%	44	<0.001	38.6%	**61.4%***	44	<0.05	55.3%	44.7%	38	0.21
37	**Pashto_M**	Turkish_F	Xhosa_F			Russian_F	**Amharic_F**			Czech_F	Korean_F		
		67.6%	32.4%	37	<0.05	25.0%	**75.0%*** **	40	<0.001	47.4%	52.6%	38	0.31
38	**Persian_F**	Portuguese_M	**Pashto_M**			**Slovak_M**	Korean_M2			Bulgarian_M	French_M		
		38.1%	**61.9%***	42	<0.05	**61.2%***	38.8%	49	<0.05	51.4%	48.6%	37	0.37
39	**Persian_M**	**Norwegian_F**	Telugu_F			Spanish_F	**Oriya_F**			Swedish_F	**Korean_F**		
		66.7%*	33.3%	36	<0.05	36.7%	**63.3%***	30	<0.05	29.2%	**70.8%***	24	<0.05
40	**French_F**	**French_M**	Portuguese_M			**Pashto_M**	Japanese_M			Hindi_M	**Xhosa_M**		
		97.1%* **	2.9%	35	<0.001	**70.3%** **	29.7%	37	<0.01	35.6%	**64.4%** **	59	<0.01
41	**French_M**	Turkish_F	Slovak_F			Thai_F	Xhosa_F			Hindi_F	**Bulgarian_F**		
		53.8%	46.2%	39	0.26	39.5%	60.5%	38	0.07	9.4%	**90.6%*** **	32	<0.001

No.	Language	Question 1				Question 2				Question 3			
		Similar	Dissimilar	N	p	Similar	Dissimilar	N	P	Similar	Dissimilar	N	p
42	**Italian_F**	**Xhosa_M**	Egyptian_Arabic_M			**Fijian_M**	Hausa_M			Korean_M	**Turkish_M**		
		80.0%* **	20.0%	40	<0.001	**66.7%***	33.3%	33	<0.05	18.9%	**81.1%*** **	53	<0.001
43	**Italian_M**	**Spanish_F**	Norwegian_F							**Portuguese_F**	Hindi_F		
		97.3%* **	2.7%	37	<0.001					**87.9%*** **	12.1%	33	<0.001
44	**Portuguese_F**					Italian_M	**Turkish_M**			Japanese_M	**Hungarian_M**		
						36.4%	**63.6%***	33	<0.05	38.5%	**61.5%***	52	<0.05
45	**Portuguese_M**	**Bulgarian_F**	Fijian_F			**Polish_F**	Thai_F			**Spanish_F**	French_F		
		63.6%*	36.4%	33	<0.05	**91.3%*** **	8.7%	46	<0.001	**86.2%*** **	13.8%	29	<0.001
46	**Spanish_F**	**Italian_M**	German_M			Xhosa_M	**Hausa_M**			**Polish_M**	Korean_M2		
		86.8%* **	13.2%	38	<0.001	23.7%	**76.3%*** **	38	<0.001	**89.7%*** **	10.3%	29	<0.001
47	**Spanish_M**	**Spanish_F**	Hungarian_F			**Croatian_F**	Mandarin_F			Russian_F	Hindi_F		
		93.5%* **	6.5%	31	<0.001	**97.3%*** **	2.7%	37	<0.001	60.0%	40.0%	40	0.08
48	**Japanese_F**	**Japanese_M**	Vietnamese_M			Somali_M	**Hungarian_M**			Amharic_M	Thai_M		
		67.7%*	32.3%	31	<0.05	9.8%	**90.2%*** **	41	<0.001	41.7%	58.3%	24	0.15
49	**Japanese_M**	Spanish_F	**Thai_F**			Portuguese_F	Swedish_F			**Czech_F**	French_F		
		30.0%	**70.0%** **	30	<0.01	37.5%	62.5%	32	0.06	**82.8%*** **	17.2%	29	<0.001
50	**Korean_F**	**Turkish_M**	Yiddish_M			**Finnish_M**	Russian_M			Japanese_M	Hindi_M		
		97.3%* **	2.7%	37	<0.001	**79.3%** **	20.7%	29	<0.01	58.6%	41.5%	65	0.07
51	**Korean_M1**	Italian_F	**Vietnamese_F**			Thai_F	Oriya_F			**Mandarin_F**	German_F		
		17.5%	**82.5%*** **	40	<0.001	46.8%	53.2%	62	0.26	**78.3%*** **	21.6%	37	<0.001
52	**Korean_M2**	Japanese_F	**Mandarin_F**			**Norwegian_F**	Bulgarian_F			**Vietnamese_F**	German_F		
		23.5%	**76.5%*** **	34	<0.001	**82.0%*** **	18.0%	61	<0.001	**94.9%*** **	5.1%	59	<0.001
53	**Thai_F**	French_M	**Hungarian_M**			**Vietnamese_M**	Croatian_M			**Pashto_M**	Ukrainian_M		
		13.3%	**86.7%*** **	30	<0.001	**92.1%*** **	7.9%	63	<0.001	**91.2%*** **	8.8%	57	<0.001
54	**Thai_M**	**Hindi_F**	Swedish_F			Ukrainian_F	Croatian_F			**Mandarin_F**	Fijian_F		
		66.7%*	33.3%	39	<0.05	45.5%	54.5%	33	0.24	**85.7%*** **	14.3%	63	<0.001
55	**Xhosa_F**	Portuguese_M	**Vietnamese_M**			**Mandarin_M**	Russian_M			Bulgarian_M	**Hindi_M**		
		11.1%	**88.9%*** **	36	<0.001	**75%*** **	25.0%	36	<0.001	34.0%	**66.0%** **	47	<0.01

No.	Language	Question 1				Question 2				Question 3			
		Similar	**Dissimilar**	**N**	**p**	**Similar**	**Dissimilar**	**N**	**P**	**Similar**	**Dissimilar**	**N**	**p**
56	**Xhosa_M**	Italian_F	Hungarian_F			Spanish_F	Swedish_F						
		38.5%	61.5%	26	0.08	51.5%	48.5%	33	0.37				
57	**Yoruba_M**	Italian_F	Mandarin_F			Spanish_F	**Korean_F**			Japanese_F	**Amharic_F**		
		62.1%	37.9%	29	0.07	37.8%	**62.2%***	45	<0.05	28.0%	**72.0%****	25	<0.01
58	**Mandarin_F**	**Persian_M**	Russian_M			Thai_M	**Korean_M2**			**Turkish_M**	Hindi_M		
		82.9%***	17.1%	41	<0.001	20.9%	**79.1%*****	43	<0.001	**66.7%***	33.3%	30	<0.05
59	**Mandarin_M**	Xhosa_F	Hungarian_F			**Persian_F**	Swedish_F			Ukrainian_F	Amharic_F		
		48.1%	51.9%	77	0.32	**82.8%*****	17.2%	29	<0.001	42.9%	57.1%	42	0.14
60	**Turkish_F**	**Vietnamese_M**	Yoruba_M			**Pashto_M**	Somali_M			**Finnish_M**	Egyptian_Arabic_M		
		78.8%***	21.2%	33	<0.001	**76.5%*****	23.5%	51	<0.001	**84.6%*****	15.4%	52	<0.001
61	**Turkish_M**					**Korean_F**	Italian_F			Russian_F	Fijian_F		
						64%*	36.0%	50	<0.05	53.3%	46.7%	60	0.26
62	**Finnish_M**	**Norwegian_F**	Amharic_F			**Fijian_F**	German_F						
		72.5%***	27.5%	51	<0.001	**79.3%*****	20.7%	29	<0.01				
63	**Hungarian_F**	**Hungarian_M**	Hindi_M			German_M	Somali_M			**Yiddish_M**	Egyptian_Arabic_M		
		86.8%***	13.2%	38	<0.001	62.5%	37.5%	24	0.08	**67.9%****	32.1%	53	<0.01
64	**Hungarian_M**	**Slovak_F**	French_F			**Turkish_F**	Hindi_F			**Persian_F**	Amharic_F		
		79.1%***	20.9%	43	<0.001	**67.4%****	32.6%	43	<0.01	**67.7%***	32.3%	31	<0.05

Language_M- a male recording of the language, *language_F*- a female recording of the language; *p<.05, **p<.01, ***p<.001.

Appendix F- Number of Times and Percentages of Languages Chosen as Answers

No.	Language	Gender	Similar			Dissimilar			Mean %
			No. significant	No. of appearance	Percentage	No. significant	No. of appearance	Percentage	
1	Amharic	Female	-	-	-	2	6	33.3%	33.3%
		Male	3	4	75.0%	-	-	-	75.0%
		Total	3	4	75.0%	2	6	33.3%	54.2%
2	E. Arabic	Male	-	-	-	1	3	33.3%	33.3%
3	Hausa	Male	0	1	0.0%	1	4	25.0%	12.5%
4	Somali	Male	0	2	0.0%	0	3	0.0%	0.0%
5	Vietnamese	Female	1	1	100.0%	2	5	40.0%	70.0%
		Male	5	6	83.3%	1	2	50.0%	66.7%
		Total	6	7	85.7%	3	7	42.9%	64.3%
6	Fijian	Female	2	2	100.0%	1	6	16.7%	58.3%
		Male	1	2	50.0%	0	1	0.0%	25.0%
		Total	3	4	75.0%	1	7	14.3%	44.6%
7	Telugu	Female	1	2	50.0%	0	4	0.0%	25.0%
		Male	1	2	50.0%	-	-	-	50.0%
		Total	2	4	50.0%	0	4	0.0%	25.0%
8	Bulgarian	Female	3	3	100.0%	1	4	25.0%	62.5%
		Male	0	3	0.0%	0	1	0.0%	0.0%
		Total	3	6	50.0%	1	5	20.0%	35.0%
9	Croatian	Female	2	4	50.0%	0	2	0.0%	25.0%
		Male	1	2	50.0%	0	2	0.0%	25.0%
		Total	3	6	50.0%	0	4	0.0%	25.0%
10	Czech	Female	5	6	83.3%	0	1	0.0%	41.7%
		Male	-	-	-	1	2	50.0%	50.0%
		Total	5	6	83.3%	1	3	33.3%	58.3%

No.	Language	Gender	Similar			Dissimilar			Mean %
			No. significant	No. of appearance	Percentage	No. significant	No. of appearance	Percentage	
11	Polish	Female	3	4	75.0%	-	-	-	75.0%
		Male	1	1	100.0%	0	1	0.0%	50.0%
		Total	4	5	80.0%	0	1	0.0%	40.0%
12	Russian	Female	0	3	0.0%	0	1	0.0%	0.0%
		Male	-	-	-	1	10	10.0%	10.0%
		Total	0	3	0.0%	1	11	9.1%	4.5%
13	Slovak	Female	5	5	100.0%	0	1	0.0%	50.0%
		Male	3	3	100.0%	0	1	0.0%	50.0%
		Total	8	8	100.0%	0	2	0.0%	50.0%
14	Ukrainian	Female	0	4	0.0%	-	-	-	0.0%
		Male	1	3	33.3%	0	1	0.0%	16.7%
		Total	1	7	14.3%	0	1	0.0%	7.1%
15	German	Female	-	-	-	0	7	0.0%	0.0%
		Male	2	3	66.7%	0	4	0.0%	33.3%
		Total	2	3	66.7%	0	11	0.0%	33.3%
16	Norwegian	Female	4	4	100.0%	0	1	0.0%	50.0%
		Male	-	-	-	2	3	66.7%	66.7%
		Total	4	4	100.0%	2	4	50.0%	75.0%
17	Swedish	Female	2	3	66.7%	0	7	0.0%	33.3%
18	Yiddish	Male	1	3	33.3%	0	1	0.0%	16.7%
19	Hindi	Female	1	2	50.0%	2	8	25.0%	37.5%
		Male	0	1	0.0%	1	6	16.7%	8.3%
		Total	1	3	33.3%	3	14	21.4%	27.4%
20	Oriya	Female	2	2	100.0%	2	4	50.0%	75.0%
21	Pashto	Male	3	3	100.0%	1	2	50.0%	75.0%

No.	Language	Gender	Similar			Dissimilar			Mean %
			No. significant	No. of appearance	Percentage	No. significant	No. of appearance	Percentage	
22	Persian	Female	4	5	80.0%	-	-	-	80.0%
		Male	3	3	100.0%	-	-	-	100.0%
		Total	7	8	87.5%	-	-	-	87.5%
23	French	Female	0	1	0.0%	0	4	0.0%	0.0%
		Male	2	4	50.0%	0	5	0.0%	25.0%
		Total	2	5	40.0%	0	9	0.0%	20.0%
24	Italian	Female	1	6	16.7%	0	1	0.0%	8.3%
		Male	3	5	60.0%	-	-	-	60.0%
		Total	4	11	36.4%	0	1	0.0%	18.2%
25	Portuguese	Female	1	7	14.3%	0	1	0.0%	7.1%
		Male	2	5	40.0%	0	1	0.0%	20.0%
		Total	3	12	25.0%	0	2	0.0%	12.5%
26	Spanish	Female	4	10	40.0%	-	-	-	40.0%
		Male	-	-	-	-	-	-	-
		Total	4	10	40.0%	-	-	-	40.0%
27	Japanese	Female	3	8	37.5%	0	1	0.0%	18.8%
		Male	3	6	50.0%	0	2	0.0%	25.0%
		Total	6	14	42.9%	0	3	0.0%	21.4%
28	Korean	Female	1	1	100.0%	4	9	44.4%	72.2%
		Male 1	0	1	0.0%	1	2	50.0%	25.0%
		Male 2	-	-	-	1	7	14.3%	14.3%
		Total	1	2	50.0%	6	18	33.3%	41.7%
29	Thai	Female	0	2	0.0%	1	8	12.5%	6.3%
		Male	0	2	0.0%	0	3	0.0%	0.0%
		Total	0	4	0.0%	1	11	9.1%	4.5%

No.	Language	Gender	Similar			Dissimilar			Mean %
			No. significant	No. of appearance	Percentage	No. significant	No. of appearance	Percentage	
30	Xhosa	Female	0	1	0.0%	1	5	20.0%	10.0%
		Male	1	2	50.0%	1	2	50.0%	50.0%
		Total	1	3	33.3%	2	7	28.6%	31.0%
31	Yoruba	Male	-	-	-	0	2	0.0%	0.0%
32	Mandarin	Female	2	2	100.0%	1	6	16.7%	58.3%
		Male	1	1	100.0%	0	3	0.0%	50.0%
		Total	3	3	100.0%	1	9	11.1%	55.6%
33	Turkish	Female	4	6	66.7%	0	2	0.0%	33.3%
		Male	4	5	80.0%	3	3	100.0%	90.0%
		Total	8	11	72.7%	3	5	60.0%	66.4%
34	Finnish	Male	4	5	80.0%	-	-	-	80.0%
35	Hungarian	Female	2	4	50.0%	0	4	0.0%	25.0%
		Male	4	4	100.0%	3	4	75.0%	87.5%
		Total	6	8	75.0%	3	8	37.5%	56.3%

Appendix G- A Suggestion of a Similarity Model with Weighted Features

The most salient component in (16) is C_3, and it exists in the Base language, thus we mark 'yes' (=√) in the relevant cell; in language A, C_3 does not exist, so the cell will be marked as 'no' (=X). In both language B and L1, C_3 exists, therefore both are marked as 'yes'. When all cells are marked, we perform the calculation – the Base language is marked as 'yes' and so is language B, thus both get five points, and their similarity gap is zero (Base language minuslanguage B). Since language A is marked as 'no', it gets zero points and the similarity gap between it and the Base language is five (Base language minus language A). If L1 is marked 'no', the languages will get zero additional points; if L2 is marked 'yes', it means that the speaker can better recognize in what languages this component appears, so languages which are also marked 'yes' will get one additional point (as in A_3, for example). After we finish going through all the components, we sum up all the gap similarity points of each language – language A has 12.5 similarity points and language B has 10.8 similarity points, therefore language B should be reported as more similar to the Base language. We should also consider at some point that features might have a conjoined weight in addition to their individual weight and add their conjoined weight to the scale. For example, the features [-back] and [+round] might be common in vowels when they appear separately (i.e., front vowels and round vowels are relatively common), but a vowel with both of these features is much more marked than other vowels (e.g., the front rounded vowel /ø/).

	C_3- 5 points	C_2- 4 points	A_3- 3.5 points	B_3- 3.25 points	B_2- 3 pointes	C_1- 2.8 pointes	…	Similarity sum
Base language	√	√	X	√	X	√		
Language A	X	√	√	√	√	√		
Language B	√	X	X	√	√	X		
L1- 1 point	X	X	√	X	X	√		
Base minus A	*5*	*0*	*4.5*	*0*	*3*	*0*		*12.5*
Base minus B	*0*	*4*	*0*	*0*	*3*	*3.8*		*10.8*

Appendix H- Non-Phonological Properties' Statistical Analysis

(a) The family of the language

Family	Similarity	No. of significant	Percentage	P
Afro-Asiatic	Similar	8	40.0%	= .13
	Dissimilar	5	25.0%	
	None/both	7	35.0%	
	Total	**20**	**100%**	
Austro-Asiatic	Similar	2	40.0%	= .13
	Dissimilar	1	20.0%	
	None/both	2	40.0%	
	Total	**5**	**100%**	
Austro-nesian	Similar ***	5	100.0%	< .001
	Dissimilar	0	0.0%	
	None/both	0	0.0%	
	Total	**5**	**100%**	
Dravidian	Similar	3	50.0%	= .19
	Dissimilar	2	33.3%	
	None/both	1	16.7%	
	Total	**6**	**100%**	
Indo-European - Balto Slavic	Similar ***	26	65.0%	< .001
	Dissimilar	5	12.5%	
	None/both	9	22.5%	
	Total	**40**	**100%**	
Indo-European - Germanic	Similar ***	13	72.2%	< .001
	Dissimilar	1	5.6%	
	None/both	4	22.2%	
	Total	**18**	**100%**	
Indo-European - Indo- Iranian	Similar	5	27.8%	= .38
	Dissimilar	5	27.8%	
	None/both	8	44.4%	
	Total	**18**	**100%**	
Indo-European - Italic	Similar *	13	59.1%	< .05
	Dissimilar	6	27.3%	
	None/both	3	13.6%	
	Total	**22**	**100%**	
Japonic	Similar	2	33.3%	= .31
	Dissimilar	2	33.3%	
	None/both	2	33.3%	
	Total	**6**	**100%**	
Koreanic	Similar	5	55.6%	= .06
	Dissimilar	2	22.2%	
	None/both	2	22.2%	
	Total	**9**	**100%**	
Kra-Dai	Similar *	4	66.7%	< .05
	Dissimilar	1	16.7%	
	None/both	1	16.7%	
	Total	**6**	**100%**	
Niger-Congo	Similar	1	12.5%	< .05
	Dissimilar *	4	50.0%	
	None/both	3	37.5%	
	Total	**8**	**100%**	
Sino-Tibetan	Similar	3	50.0%	= .06
	Dissimilar	1	16.7%	
	None/both	2	33.3%	
	Total	**6**	**100%**	
Turkic	Similar ***	4	80.0%	< .001
	Dissimilar	0	0.0%	
	None/both	1	20.0%	
	Total	**5**	**100%**	
Uralic	Similar ***	7	87.5%	< .001
	Dissimilar	0	0.0%	
	None/both	1	12.5%	
	Total	**8**	**100%**	

(b) The continent the language is spoken in

Continent	Similarity	No. of significant	Percentage	p
Africa	Similar	7	31.8%	= .5
	Dissimilar	8	36.4%	
	None/both	7	31.8%	
	Total	**22**	**100%**	
Asia	Similar **	30	44.8%	<.01
	Dissimilar	15	22.4%	
	None/both	22	32.8%	
	Total	**67**	**100%**	
Europe	Similar ***	59	67.0%	<.001
	Dissimilar	12	13.6%	
	None/both	17	19.3%	
	Total	**88**	**100%**	

(c) The familiarity of the languages

	Similarity	No. of significant	Percentage	p
Familiar	Similar ***	30	55.6%	< .001
	Dissimilar	10	18.5%	
	None/both	14	25.9%	
	Total	**54**	**100%**	
Unfamiliar	Similar *	21	50.0%	> .05
	Dissimilar	12	28.6%	
	None/both	9	21.4%	
	Total	**42**	**100%**	

תקציר

חוקרים רבים חקרו דמיון בין שפות (לדוגמא Eden 2018; Crowley and Bowern, 2010; Longobardi and Guardiano, 2009, 2017), אך טרם פורסם מחקר אשר מכמת את הדמיון בין השפות. המטרה הסופית של המחקר הנוכחי היא לבחון האם ניתן למדוד ולכמת דמיון באמצעות שימוש בסקאלות של בולטות אקוסטית של מספר מאפיינים פונטיים ופונולוגיים, תוך מיזוג הסקאלות הנפרדות לסקאלה אוניברסאלית יחידה של בולטות. עם זאת, מאחר ולא קיים מחקר אשר מודד דמיון אך ורק באמצעות תכוניות פונטיות ופונולוגיות, המטרה של תזה זו הייתה לבחון אילו תכוניות אמורות להתמקם על הסקאלה הזו מלכתחילה.

המחקר הנוכחי מכיל שני ניסויים, ניסוי מקדים וניסוי מרכזי. בניסוי המקדים, 132 דוברי עברית דירגו את רמת ההיכרות שלהם עם כל אחת מ-35 השפות שהופיעו בניסוי המרכזי. בניסוי המרכזי, 362 דוברי עברית הקשיבו ל-20 סטים של שלוש הקלטות, אחת של שפת בסיס ושתיים של שתי שפות נוספות, ונשאלו איזו מבין שתי השפות הנוספות יותר דומה לשפת הבסיס. הדמיון נקבע באמצעות מספר התכוניות המשותפות בין שפת הבסיס לבין כל אחת משתי השפות האחרות, והתכוניות (41 במספרן) נלקחו ברובן מ- World Atlas of Languages Structures Online (WALS) ומ- Bradlow et al. (2010). שפה נוספת אחת חלקה באופן ניכר יותר תכוניות עם שפת הבסיס מאשר השפה הנוספת השנייה (שפה דומה ושפה לא דומה, בהתאמה). התוצאות הראו נטייה מובהקת לבחור בשפה הדומה יותר מאשר בשפה הלא דומה.

הממצאים הללו מציעים כי ניתן למדוד דמיון באמצעות תכוניות פונטיות ופונולוגיות בלבד. עם זאת, אנו יודעים כי לא כל התכוניות חשובות באותה מידה; לכן, המודל הנוכחי יכול לעבור שיפור באמצעות משקול התכוניות, כך שתכוניות הבולטות יותר יקבלו משקל גדול יותר בכימות הדמיון. משקול התכוניות נשאר למחקר עתידי.

אוניברסיטת תל-אביב

הפקולטה למדעי הרוח ע״ש לסטר וסאלי אנטין

החוג לבלשנות

שלושה גברים נכנסים לבר: כימות המרחק הפונולוגי בין שפות על גבי סקאלה אוניברסלית

חיבור זה הוגש כעבודת גמר לקראת התואר

״מוסמך אוניברסיטה״ באוניברסיטת תל-אביב

על ידי

אלונה גולובצ׳יק

העבודה הוכנה בהדרכת:

ד״ר אוון-גרי כהן

יולי 2022