Down the Rabbit Hole:
Visualizing Linguistic Distance
And Relationships
With Alice in Wonderland

YI-CHIA CHENG
Thesis presented by
Yi-Chia Cheng

To The Department of Art + Design
In Partial Fulfillment of the Requirements for the Degree of Master of Fine Arts in Information Design + Visualization

Advisor: Dietmar Offenhuber
Chair: Pedro Cruz
Respondent: Steven Braun

Northeastern University
Boston, Massachusetts
November, 2019
Abstract

During the cross-language learning or experience process, human brains are aware of psychological "distance" between the languages they already know and the ones that they are currently learning or experiencing. Some seem to be closer, while some are more distant. This thesis will explore the approaches to visualize the phonetic and syntactical distance and relationships across ten of the world languages using notation design systems, scatterplots and human articulator diagrams with sample texts from Alice's Adventures in Wonderland by Lewis Carroll.

The three methods mentioned above in this document are used to visualize different aspects of languages. Through simplifying the design of notation systems and showing the quantity of phonetic feature with scatterplots it allows readers to find out the distinct linguistic features which indicate the 'distance' among languages more effectively.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>02 Research Background</td>
<td>4</td>
</tr>
<tr>
<td>03 Related Work</td>
<td>10</td>
</tr>
<tr>
<td>04 Case Study</td>
<td>23</td>
</tr>
<tr>
<td>– Visualizing Sounds</td>
<td>24</td>
</tr>
<tr>
<td>– Visualizing Sentences</td>
<td>51</td>
</tr>
<tr>
<td>05 Conclusion</td>
<td>58</td>
</tr>
<tr>
<td>06 Discussion &amp; Future Work</td>
<td>61</td>
</tr>
<tr>
<td>– Concern of Data</td>
<td>61</td>
</tr>
<tr>
<td>– Limitations of Visualization</td>
<td>63</td>
</tr>
<tr>
<td>– Future Opportunities</td>
<td>64</td>
</tr>
<tr>
<td>Bibliography</td>
<td>67</td>
</tr>
</tbody>
</table>
01 Introduction

With transportation and telecommunication technology today, cultures from different corners in the world have gained higher mobility and are able to interconnect actively with one another through language translation and learning. To share and exchange information globally, the acquisition and perception of second or even third language play as a fundamental bridge to connect one culture to another.

The pedagogy of learning foreign languages has a long history. Systematic language teaching—learning methodology research could be traced back to the 16th—century Europe (Smith & McEllevant 2014, 1). Back in the middle ages in Europe, foreign language learners mostly were mostly limited in the upper social class including noblemen and —women, merchants, and diplomats, and church educators like priests (Law 2015, 94—95). The languages learned were also highly restricted to Europe region (Glück 2014, 46). As time goes by, learning one or multiple languages is no longer a privilege to wealthy people or confined to languages of certain specific regions. As one of the common activities globally, more people are exposed to the experience of learning different languages aside from their mother tongue.

However, to master a foreign language is never an easy task. During the learning process, human brains are fully aware of some "distance" between the languages they already know and the ones that they are currently learning. Some are closer, while some are more distant (Holmer 1977, 185). What exactly is the distance? If it is not an absolute physical distance that can be measured by rulers, then how to describe it? And how do humans perceive that psychological distance?

If the psychological distance and relationships among languages are still hard to imagine, I would like to give an interesting discovery of Google translate as an example to help illustrate the phonetic relationships of languages. In Google translate, the left column is the text that is desired to be translated and the right one is the translated result. In
this small experiment, let us put some English text and choose any non—English language as the ‘text language’ in the left column. Ignoring Google's suggestion of changing into the language they detect, press the 'listen' button.

In this example above, the computer would play 'Alice's Adventures in Wonderland' in English with a Japanese accent. Play with different languages and it is not difficult to find that some languages have less of an accent when pronouncing English while some have heavier one which somewhat suggests their phonetic distance to English. If a language has light or almost no accent while pronouncing English, it indicates the language might share more similarity of phonetic features with English, and vice versa. Some languages even share similar 'incorrect' ways or intonations while pronouncing English which also implies that they might have more in common in terms of phonetic characteristics. From this easy experiment that most people can do at home, it could give a general concept of languages distance from what they sound like.

However, to analyze language distance as a whole is definitely not such an easy task. Language composition is extremely complex due to its nature, variables involved and historical evolutions. Even today, languages are constantly changing with no stop. New words are coined, and slangs and phrases are created every single day; while at the same time, some words and uses of the languages are also dying. Furthermore, even in the same language, people from different regions might use language differently. Therefore, it is nearly impossible to include all vocabulary, phrases or pronunciation of words into one single research. To make the abstract
linguistic features more understandable for people who do not have professional knowledge of linguistics, I propose using multiple notational visual representations to map out the phonetic and syntactical relationships of languages from in the world through five simple and instinctual design systems. But again, languages have complex characteristics. How to make something so complicated to be easier to perceive would be one of the fundamental questions to answer in the thesis.

Plenty of linguistics studies have been published and conducted every day. However, it is still challenging to bridge the gap between the academic, text—oriented linguistics researches filled with jargons and abstract diagrams, and the public’s understanding of languages? In the book The importance of corpus linguistics to understand the nature of languages, by the linguist Wallace Chafe, it is suggested that in order to understand something, we have to feel "make sense" or be able to connect it with our ordinary experiences (81). General public’s understanding of languages primarily comes from visual and audio channels. We "see" and "hear" multiple languages on a frequent basis today. It seems more natural for human brains to perceive languages in visualized ways focusing on the sounds and looks of them. Therefore, the means of visualizing abstract linguistic concepts to represent language differences would be able to connect scholarly linguistics researches and general understandings and experience of languages. People get to "see" and "perceive" the differences simply by visualizations which need less professional knowledge of linguistics.
02 Research Background

After introducing the concepts of linguistic distance and relationships and what role that visualizations would play to illustrate these vague notions more effectively, I would like to explain the background of the research in the 'Case Study' section. First, I would clarify the materials and data that I use for the case study and explain the rationale of choosing and not choosing for this research. Second, I would discuss the possible issues and concerns that languages collected as data in linguistics analysis. Third, to answer why 'Alice's Adventures in Wonderland' could be a proper data source of text. Later on, I would explain the research scope of this document. And the last, I would give a brief orientation of why the 'backness' of sounds in languages is one of the main aspects to look at in the case study and would discuss more in detail in the next section.

2.1 Data Collection

During the perceiving language experience processes, the spoken form of languages with include sounds and language structure seem to be fairly dynamic and instinctual to a human's perception. The spoken language form also usually comes with a string of text data which can be analyzed with standard forms like phonetic symbols and parts of speech, instead of being individually separated like writing systems which is very challenging to determine the 'distance' from one another when comparing totally different writing system. To see the example below, European languages like English, German and French do share certain writing features since they all use Romanic letters. However, when looking at a broader context: it is not clear enough to pin point the lexical relationships between Arabic, Russian and Korean since they have very distinct writing systems from one another.
Therefore, the sounds and sentence structures which can be analyzed with standardized forms will be the main focus in the thesis; while the written form will not be included in the discussion in this thesis.

2.2 Languages as Data

According to Ethnologue: Languages of the World, over 6,000 languages are spoken today. The number is constantly changing because of the disappearing native speakers of about 2,000 languages (Eberhard et al. 2019). On the other hand, "only 23 languages account for more than half the world's population," cited from the same—named website. It is imaginable how imbalanced of language power could be in terms of its number of speakers, and economic and cultural strength. Before conducting this research, it was aware that the language—power—bias should be minimized by including as more language as possible.

However, once a language is treated as data, it comes the question that if there are enough materials to be analyzed as data which highly echos back to the language—power—bias that is mentioned above. For this case study, I mainly acquire the phonetic symbol transcriptions, as the data, through a language learning website called EasyPronunciation.com which provides resources for learning pronunciations of different languages online. The phonetic symbol research will therefore mainly focus on the ten languages on its list: English, Spanish, French, German, Italian, Danish, Portuguese, Russian, Chinese and Japanese. Not even mention the languages with a very small number of speakers, even the languages like Hindi and
Indonesian which have a large number of native speakers but with relatively less economic, cultural and global influence have the problem of data insufficiency. Contrast to some European languages like German and Italian which have abundant phonetics data but with a comparatively smaller population of speakers.

To deal with the insufficiency of data, using romanized letters instead of phonetic symbols seem to be helpful with this issue. However, the reasons why I still choose phonetic symbols transcription instead of romanizing languages is to not only avoid the inaccuracy of pronunciation which I would further discuss in the section 06 Discussion and Future Work, but also avoid to reinforce the strength of the languages that are already with relatively more economic, cultural and global power. In other word, the people who use or are native to roman alphabets are traditionally also the speakers of European languages with sufficient linguistic data.

To better address this imbalance of language—power—bias, I decide to approach more languages through the sound spectrograms and sentence structure which are not as limited. Eventually, the research is expected to continue to include more languages in the world into it in the future.

2.3 Text Adventures with Alice in Wonderland

For the thesis research, phonetics and sentence structure from pieces of texts and paragraphs would be the main approach to collect data and help pull out the final visual representations. First, I would like to analyze the same pieces of text in different languages to inspect multiple linguistics variables: through recording audio books for comparing sound shapes and identifying sounds into IPA chart and spectrograms to visualize the phonetic attribute; and through analyzing the order of sentences and grammatical rules to investigate the syntactical feature.

The sample texts from Alice’s Adventures in Wonderland by English writer and mathematician—Lewis Carroll will be my
data source for three major reasons. First, Alice's Adventures in Wonderland as one of the best known children's books, has received worldwide recognition and has been translated into more than 170 languages globally. With plain language illustrating an exciting adventure of a 10-year-old girl, Alice in Wonderland is also a popular book used for language learning. Audiobooks and side-by-side bilingual texts are easily accessible. Therefore, it provides an ideal material for text analysis to compare the same piece of texts in multiple languages with natural translations by human professionals. Second, the book is not only a beloved children's book but also plays with linguistic possibilities which could act as meta—linguistic research that connotes relevant content to the thesis. For example, when mock turtle mispronounces the word "reading" as "reeling", few people might relate it to the phonetics of linguistics research. In other words, the author does not simply pick up a random word but considers the ways of English speakers pronouncing words. /d/ and /l/ sounds are both voiced and produced at the same position of human speech organ which could be explained why the /d/ sound is replaced by /l/ in this joke.

Even though Lewis Carroll is not a linguist, but he coins new words by playing with phonetic features of English throughout the book is another brilliant reinvention of languages.

However, to "translate" the joke would be one of the most challenging tasks for translators. They need to translate into something natural in their languages, but still keep the essence of how Lewis Carroll plays with English phonemes.

In some of the Chinese translations, translators use exactly same pronunciation but wrong characters to transmit the joke; while in some of the Japanese translations, translators choose to keep the same number of the syllables of "reading" and "writing" in Japanese, but with wrong phonemes. These versions of translation do not necessarily translate from English word by word; however, successfully transform the joke to make the texts natural and witty to read which is one of the major reasons that I use translated texts as analysis material.
Below is the research scope of this document. It can be dangerous to determine how close two languages are merely based on one or two aspects of linguistics. Therefore, I would like to clarify my research scope. What I included here is just a small part of language, so do most of the researches on linguistic distance and relationships which will be further illustrated in the 'Related Work' section.
2.5 Where Do Sounds Occur: The Backness of the Sounds

What is the importance of discussing the backness of sounds, especially the vowels? I recorded the different sounds in languages and found out that the feature of vowels as more distinctly audible sounds seems to be shared by most languages in the world. When we hear languages, we hear a lot of vowels. For example, when a person shouts out a word, the parts that can be prolonged and emphasized are the vowels. Therefore, vowels seem to be an appropriate element to analyzed. Furthermore, the backness of vowels could also suggests the audio level (dB) among different vowels. Take front vowel: /a/ and back vowel /ɑ/ as example. /ɑ/ not only has higher sound frequency distribution, but also is louder than /ɑ/ on the waveform. More details of vowels and consonants position comparison among languages will be discussed more thoroughly in the ‘Case Study’ section.

Furthermore, phonetics, study of speech sound is another important topic under the category of language form. One the other hand, language meaning includes semantics and semiotics. Last one, language in context is to study language in broader contexts, for example, study languages in history is called historical linguistics; study languages in society is sociolinguistics.

Taking how complicated the linguistics could be in mind, to measure a single linguistic distance between two languages sounds fairly challenging. In this case study, I would focus on phonetic relationships and distance of language along with a brief comparison on their sentence structure and parts of speech.
2.4 Research Scope

Linguistics typically includes the studies on language form, language meaning, and language in context (Martinet 1960, 15). Different linguists might have different ideas on how to categorize sub-topics under linguistics. Here, I would simplify the concepts of linguistics based on André Martinet’s *Elements of General Linguistics*. Language form is to investigate how to form a language which could be taken as grammar rules that construct a language like morphology, phonology and syntax.

**03 Related Work**

From the previous linguistics researches, especially in the field of comparative and historic linguistics, incredibly rich resources are provided to structure the concepts of linguistic relationships. Here are two of the most related researches:

3.1 Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages

In Chiswick and Miller’s Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages, they proposed to address the linguistic distance (LD) with linguistic score (LS) by simply asking: how difficult is it for individuals who know language A to learn language B. It leads to a scalar measure: LD= 1/LS. The higher the learners score, the closer the distance that language A and language B have (2007, 8—9).

Chiswick and Miller's paper has several problems that need to address: first, the ability of learning foreign languages might be not simply determined by how close the learners' mother tongue and the foreign language is, but the length of
learning, the time and efforts dedicated should be all taken into consideration. Therefore, on selecting subjects for this research can be extremely challenging since it has to be learners that have exactly the same experience and length of learning English. Second, the research centers foreign English learners by conducting English language testing. However, if we would like to explore world languages in a more thorough perspective and equal networks, it might need to take more linguistic features into consideration and avoid putting any language as the standard for comparison.

3.2 Linguistic relationship on three levels

In Nils M. Holmer's Linguistic relationship on three levels, he suggested to examine the linguistic relationships with various faces of the entire language structure which a focus on the mechanism of a language. Based on the mere similarity of words, we could investigate the lexical relationships which are strictly controlled by phonetic rules. Other than the sound and look of a word, the "form" of a language acts as a broader context which leads to the examination of morphological—syntactical relationships.

The central idea and case study in this document are largely inspired by Holmer's paper. It sheds light on how to examine linguistic relationship with language as an extremely complex as a whole by 'dissembling' it into parts with different functions: separated but still inter—connected.

Holmer's analysis starts with the historical, morphological and syntactical aspects of the languages. To truly get deep understanding of linguistic relationships across languages, it is important to know the historical evolution and even migration of languages and language families. This is the crucial side of investigating language distance that will not be discussed too much in this thesis, but it is still good to keep in mind.

After reviewing some linguistics studies, it is time to also take a look at what visualizations were made to address language relationships and distance in the past.
3.3 Language Family Tree

Language tree is one of the most common visual representations showing language relationships that most people have already known. It has many versions available which usually focus on different language family since it is unlikely to put every language on one piece of visualization and plus, linguists are still researching on many languages which are controversial of which family they belong to, or still figuring out the classic linguistic question that 'is it a language or a dialect'. Therefore, the language family tree might also have to be adjusted and updated overtime.

Despite the uncertainties of the accuracy of information, language family tree is able to act as everyone's introduction of language distance and relationships. The visual concept of a thick tree trunk spread out into thinner branches works effectively as an analogy of language families and their relationships which to some extent, further suggests the distance among languages. For example, the branches (the languages) which share the same trunk (the language family) has closer distance compared with the branches derived from other trunks.
It seems that language family tree works finely at explaining the historical evolution of languages. However, it also raises questions like: how close the two languages are if we see them right next to each other in the diagram? In what way they are close? To answer this, we have to investigate linguistic relationships and distance among languages by looking at multiple aspects of language composition, for example, phonetic and syntax aspects in this thesis.

3.4 Lexical distance diagram

Besides language family tree, another linguistics design work that visualize linguistic distance is created by Ukrainian linguistics professor Kostiantyn Tyshchenko—'lexical distance diagram' which was introduced in 2013. It was one of the first and profound attempts to visualize language distance by linguists. The diagram shows languages spoken in Europe and their lexical distance with language families color-coded. The closer that distance, the more words they have in common in
the chosen word list; the further the distance, the harder the mutual comprehension would be since they have less identical words.

However, this diagram solely emphasizes the written parts of the languages. Without taking other linguistics features like grammar, syntax, phonetic which are important for intelligibility into consideration, we could only say that it shows partial interface of the whole extremely complex language system. In addition, some distance among languages are not as clear in the diagram and could be misleading by using symbols to omit the real distance. Despite these inaccuracies, it is definitely a fresh attempt to visualize language relationships, but it might have to be more carefully designed and clarified.

Fig 3.4.1 Kostiantyn Tyshchenko: lexical distance diagram

3.5 IPA Chart

If we try to visualize not just lexical distance, but also the phonetic parts of the languages, the International Pronunciation Alphabet chart, as known as the IPA chart could be a great data material for phonetics visualization. The initial
The idea of IPA chart was first to create several standardized phonetic tables which include all the sounds of the languages founded in the world in the late 19th century by Phonetic Teachers' Association (Kenyon 1951, 311—320). The version we see today was developed from the model in International Phonetic Handbook in 1989. Due to the frequency of use and simplicity, I would only discuss the pulmonic consonants and the vowels without special marks rather than all the symbols displayed in this document.

The standard IPA chart revised in 2018 presents vowels with the x—axis indicating the position in human's mouth when pronouncing; they—axis suggests the openness or closure level of human lips when pronouncing. Similar to the vowel chart, the chart of pulmonic consonants is also with the x—axis representing the positions of human's mouth involved when pronouncing, and with the y—axis showing the ways to pronounce. Since it is designed to be standardized, it would be
appropriate to be analyzed in global languages researches.
However, when first time looking at the IPA chart, a lot of the symbols on the table are not as instinctual enough that allows people to have a clear clue of how to pronounce them. In the 'Conclusion and Discussion' section, I would further evaluate the advantages and the disadvantages of using phonetic symbols from IPA chart in my case study analysis.

Fig 3.5.1 IPA Chart, 2018 version

3.6 Visible Speech

The IPA chart today is composed of more than 150 linguistic symbols to "represent" sounds in the world as accurately as it can. However, like I mentioned previously, the symbols in the chart have already developed based on an independent system with little relation of how the sound is produced and how it sounds like. About 20 years before the first IPA chart was introduced, visible speech was another phonetic visualization to connect with the sounds and their visual representations of the organs involved when pronouncing.

Visible Speech, also as known as physiological alphabets, is a writing system invented in 1867 by Alexander Melville Bell, the father of Alexander Graham Bell who owns the patent of the first practical telephone (Bruce 1995, 149). In his book, Visible Speech: The Science of Universal Alphabetics, he illustrates a visual diagram of human articulator with symbol elements that compose each sound in English language and potentially, all the sounds in world languages. He further argues that this new phonetic notation system could help not only the deaf to perceive and produce sounds and eventually be able to communicate, but also foreign language learning
since it is simply established by a fairly straightforward visual representation of pronouncing sounds (15—16).

The invention of visible speech is a well—designed system attempting to use simplified, logic and instinct—driven
symbols to visualize how a sound might "look" like when human pronounce it. Even though the idea of visible speech was not widely put into practice then, Bell's theory inspired the later invention of recording sound pattern in real time which is known as spectrogram today.

Fig 3.6.1 Alexander Melville Bell: Visible Speech—initial visual clues for sounds

Fig 3.6.2 Alexander Melville Bell: Illustration of Visible Speech

3.7 Spectrograms

A spectrogram is a visual representation of sound frequencies and intensity. The y—axis shows the frequency of sounds. The higher the frequency is, the higher the pitch of the sound is, and vice versa. The x—axis, on the other hand, shows the change of frequencies over time. Furthermore, the shades of color suggest the intensity of sounds. In other words, its scale from red, yellow, green to blue and purple represents where occurs the dominant frequency marked by relatively reddish colors.

A Spectrogram could be one of the answers to what sounds would "look" like. Speech sounds are, by the phonetic nature, way more complicated than the symbols in visible
speech system. A spectrogram is originally most common to be used to identify and research single phonetic sound in computational linguistics. In the recent year, we can see some creative ways that interact with it through music and sound recordings from nature. Even though there's not much for multi—languages comparison studies, the fact that spectrogram is able to catch the subtle details of sounds more accurately would make it a useful tool to analyze sound patterns across different languages.

Fig 3.7.1 Spectrogram of the spoken words “nineteenth century” in English

To compare languages, we have to look at the syntactical aspect besides lexical and phonetic features. In the section, I will mainly focus on the word frequency and the structure of sentences.

3.8 Word Tree: structure of the sentences

To look at the structure of sentences, the word tree generator below made by Jason Davies would be one of the great tools to investigate. Words are composed in different orders in different languages. By looking at the same word, we could quickly tell what parts of speech are ordered accordingly and compare it with different languages.

One of the limitations is that most word tree tools are available to limited languages. To compare, for example, European and Asian languages, would be a challenge since we need to pull out the parse of the Asian languages first. Even some separate packages in a few non—European languages
In the case study, the sentence structure is going to be analyzed in a fairly different approach compared to this which will be discussed in detail in the next section.
4.1 Visualizing Languages—What are we visualizing?

Language, by its nature, is an extremely complicated human invention that has been developed for thousands of years. As one of the major methods of human communication, it could be examined from spoken and written forms. The case study would discuss visualizations of phonetics relationships with a small part of syntax to compare. By investigating how we pronounce words and compose sentences, this section would explore the possibilities of visualizing phonetic relationships. The visual tools I used in this case study are human speech diagrams, histograms of vowels distribution, notation design system including color notation belts, phoneme frequency pyramids, sound spectrograms and plotting charts.

As for the reason that this document does not include the writing system: unlike the spoken form of language (which includes sounds of words and sentence structure) has
standard forms like phonetic symbols on IPA chart or parts of speech to evaluate, a written system is more of individually distinct from one and other. Except some European and East Asian languages share similar systems, a lot of languages' writing systems in the world do not have much relation with one and another and therefore, get more challenge to visualize their relationships.

Fig 4.1.1 Alice in Wonderland: Written forms of Arabic, Korean and Russian

4.2 Visualizing Phonetics—Introduction of Phonemes in Languages

Phonemes is a small unit of speech sound that distinguish one word from another in languages. In this case study, phonemes will be marked as phonetic symbols with visual transcriptions based on the IPA chart. An example text from the first paragraph of Alice's Adventures in Wonderland and its phoneme transcription will look like this:

Alice's Adventures in Wonderland

1 Alice was beginning to get very tired of sitting by her sister on the bank,
2 and of having nothing to do:
3 once or twice she had peeped into the book her sister was reading,
4 but it had no pictures or conversations in it,
5 'and what is the use of a book,'
6 thought Alice, 'without pictures or conversation?'

æləsəz ædvɛnʧərz ɪn wʌndərlænd

1 æləs wʌz bɪgɪnɪŋ tu gɛt vɛri tæʊɜrd əv sɪtɪŋ bæɪ hɜr sɪstər ɒn ðə bæŋk,
2 ænd əv hævɪŋ nʌθɪŋ tu du:
3 wʌns ɔr twaɪs ʃi hæd pipt ɪntu ðə bʊk hɜr sɪstər wɜズ rɛdɪŋ,
4 bʌt ɪt hæd ɜə nʊ pɪktʃəz ɔr kənvərsɛʃənz ɪn ɪt,
Each phonetic symbol represented is produced by distinct positions of human speech organ. Ranging from front, middle to back sounds, language speech sounds could be summarized by the distributions of phonemes' positions. In this section, I will discuss the role of vowels and consonants' distribution of pronouncing positions in phonetic relationships across languages respectively.

4.3 Visualizing Vowels

In this human speech organ diagram, I align the vowel position chart from IPA into our oral cavity: from the front of the speech organ to the back; and up to down to suggest how big our mouth has to open when pronouncing a specific vowel. Therefore, instead of an absolute point in the mouth, the position of vowels discussed here is more of a psychological concept of positions.
Here is an example of the first paragraph of Alice's Adventures in Wonderland which summarizes the vowel distributions in ten languages which would be discussed repeatedly in this section which several visualization approaches: English, Spanish, French, Italian, German, Danish, Russian, Portuguese, Chinese and Japanese according to their phonetic symbol transcriptions.

The concept of vowel's 'backness' or 'position' of human speech organ is fairly straight—forward. Starting from lips to the back of the human oral cavity. Unlike consonants which have distinct places of pronouncing (for example, /b/ is a bilabial sound which specifically involves two lips), the pronouncing positions of vowels are not as absolute but more of a relative concept. Try to pronounce /i/ sound and /u/ sound which is one of the most front and most back sound respectively based on IPA vowel diagram. The spaces between these two sounds are categorized as mid—front, mid and mid—back accordingly.
With examples might be more clear: English and Spanish are both Germanic languages, the distribution of vowel positions seem to show some common between the two: vowels are distributed in all positions from front to back in human speech organ. However, Russian as a language from a different branch of language family tree, also has a relatively even pattern like English and German. On the other hand, another distinct pattern here is highly concentrated on front and back side which is shared by Spanish, Italian and Japanese. For Spanish and Italian, it seems easier to predict this distribution since both languages belong to Romance languages; however, Japanese as an East Asian language seems to also share this very similar pattern with a minor difference of more back vowels than the rest two.

Now it is clear that the sounds of a language could be summarized and compared like the chart above, here comes a crucial question: what is the importance of discussing the backness of the vowels? I recorded the different sounds in languages and found out that the feature of vowels as more distinctly audible sounds seems to be shared by most languages in the world. Even though this tendency could not be found in consonants (whose audio level is primarily determined by the manners of pronouncing), when we hear languages, we hear a lot of vowels. For example, when a person shouts out a word, the parts that can be prolonged and emphasized are vowels. Therefore, vowels seem to be an appropriate element to analyzed.
Furthermore, the backness of vowels could also suggest the audio level (dB) among different vowels sometimes. Take front vowel: /a/ and back vowel /ɑ/ as example. /a/ not only has higher sound frequency distribution, but also is louder than /ɑ/ on the waveform.

4.4 Visualizing Consonants

Compared to vowels, consonants position distributions might be relatively tricky. Sounds happen at the front, middle or back of human speech organ do not have an obvious tendency. However, the positions of where consonants happen are also important to be included while discussing phonetic relationships. A native speaking of X language which involves lots of consonants primarily in either front or middle parts of speech organ. When they learn to speak Y language which has mostly back sounds but very few front and middle sounds, the learner might feel more distant to this language, and vice versa.

Therefore, I would include consonants, too, to the case study. Before starting to compare languages, I would like to introduce how the consonant positions work in a color notation system. Based on International Pronunciation Alphabet chart
(IPA chart), I categorize "Bilabial" (made with both lips), "Labiodental" (made with the lips and teeth) and "Dental" (made with the tongue against the upper teeth) sounds into the FRONT category which will be marked with pink; "Alveolar" (made with the tongue close to or touching the ridge behind the teeth on the roof of the mouth), "Post—alveolar" (made with the tongue near or touching the back of the alveolar ridge), "Retroflex" (made with the tip of the tongue curled back toward the hard palate), "Palatal" (made with the middle part of the tongue against the hard palate) and "Velar" (made with the back part of the tongue against the soft palate) sounds into the MIDDLE category which will be marked with gray; and finally "Uvular" (made with the back of the tongue against or near the uvula), "Pharyngeal" (was made primarily in the pharynx) and "Glottal" (made with obstructing airflow in the glottis) sounds into the BACK category which will be marked with black.
4.5 Color Notation System and Phonemes

Based on the positions in human articulator involved when pronouncing phonemes, I encode colors to suggest the backness of each phoneme. To emphasize difference, the front sounds and back sounds are encoded with pink and black respectively so that the distribution of these sounds can stand out for further analysis. To warm up, let us look a simple example book title of text: Alice's Adventures in Wonderland, in four languages: English, Spanish, Chinese and Japanese.

This color notation system provides a new way to visualize sounds: even without knowing how to pronounce the phonemes, readers could still further examine the sounds' position distribution patterns during a speech.
In this system, I mark front and back sound as pink and black respectively and the rest is shown in gray which can help readers spot the patterns more easily. Just from the book title, Alice's Adventures in Wonderland, there are already some patterns could be observed. Spanish has not only the longest sentence with the most phonemes, but also has the most front vowels. In contrast, Japanese, even with similar vowel distribution pattern with Spanish, it has more back sounds than the other. As for English, it has more middle sounds than any other three languages and Chinese, on the other hand, shows a fairly balanced color distribution here.

However, the book title might not be enough to really identify clear patterns. By looking at the first paragraph of Alice's Adventures in Wonderland in more languages, the patterns might be easier to spot.

Now, I would like to list four language groups which share similar patterns. The first section of every visualization is the phonetic symbol transcription with frequency distribution pyramids below showing pronouncing positions from the front to back and the most frequent phonetic symbols from the bottom; and finally, the color notation representation.

To read the visual representations better, here are a few points to look at: how are the phonemes distributed in different positions, the number of phonemes, the shape of phoneme frequency pyramid (tall or short, sharp or flat) and the color distribution. Noted that the first pyramid graphs have four position categories in order to tell more patterns and second pyramid graphs integrate middle—front, middle and middle—back all into the 'middle' category that is marked as gray for simplicity of color notation representations.

Before going through each language, I would like to summarize the common characteristics of the languages which set them different from other groups of languages.
Group 1: English, Danish, French, Chinese

Group 1 has primary middle sounds but also has a good amount of front and back sounds. Except for Chinese which has slightly more back sounds than front sounds, phonemes are distributed at middle > front > back in this group. Sharp and relatively taller pyramids could be observed here which suggest a wider diversity of phonemes in all positions with no extreme pattern having an overwhelming amount of any specific phonemes. Among the four groups, this group has the most balanced distribution pattern showing when pronouncing these languages, various parts of speech organ are involved pretty evenly.
English has a medium number of phonemes in this sample text and has primary middle sounds but overall, has a relatively even position distribution of vowels and consonants alike. No extreme pattern with one or two phonemes significantly more than the rest is observed.

Fig 4.5.1 Color Notation System of Phonetic Representation: English

**Danish**

Danish has a medium number of phonemes with an even distribution. However, the relatively shorter pyramids suggesting a little lower variety of phonemes with several concentrated patterns of /a/, /æ/ and /u/.
French

French has the tallest pyramid of front sounds in this group which shows a high variety in front sounds. On the other hand, it has the shortest pyramids in the middle sounds. Unlike other languages in the group one, even it also has the most sounds happening in the middle human articulator, its diversity of middle sounds is the lowest compared with the rest four languages. No extreme pattern with one or two phonemes.
significantly more than the rest is observed.

Fig 4.5.4 Color Notation System of Phonetic Representation: French

Chinese

Chinese has fairly even distributions with relatively more sounds happening at the mid—back and back positions of human speech organ. In the color notation visualization, more black color could be observed below. It is hard to pinpoint one position that has significantly more phonemes than the other parts. Shorter pyramids also indicate a little lower diversity of phonemes.
Group 2: Spanish, Italian, Japanese, Portuguese

Group 2 has significantly more front sounds compared with the rest two groups of languages. Flat and relatively shorter pyramids could be observed here which suggest a lower diversity of phonemes in all positions with several extreme patterns having an overwhelming number of any specific phonemes especially happen on vowels. Among the four groups, this group has the most front sounds and vowels showing when pronouncing these languages, front parts of speech organ are involved the most actively. With many front vowels could be found in this group which potentially could make this group the most audible language group among the four groups. Generally, the length of sentences in this group is
the longest of the rest four groups too.

Spanish

Spanish has a lot of front, mid/mid—front sounds with one of the fewest back sounds in this group which contributes little black and much more of pink in the color notation representation below. It shows more extreme distribution which mainly focuses on front vowels. Some specific front vowels stand out: /a/, /i/, /e/ and back vowel /o/. A relatively flat pyramid suggesting a lower variety of phonemes overall.
Italian

Alone with Spanish, it has obvious patterns of the most front sounds and the fewest back sounds. Even in this group, its phoneme distribution pattern is the closest to Spanish in this sample paragraph. Both languages show shorter and flat
Japanese has very wide and short pyramids even in this group which indicates a lower diversity of phonemes and more extreme distribution patterns here. The patterns concentrate on front vowels /a/, /i/, /e/ and back vowels /o/, /ɯ/ which lead to the widest and shortest pattern that is distinct from other groups. The reason of being categorizing into this group is because of the bigger number of front sounds. Even though it has a lot of front vowels, but with relatively more back vowels, it might balance out the audible level of the language.
In other words, Japanese might not be as audible as the languages above in this group.

Portuguese

Portuguese has a clear pattern of more front, mid/mid—front sounds. It has relatively taller pyramids in front and back sounds which make it slightly different from Spanish, Italian and Japanese. However, it is still not difficult to see that it has some extreme patterns of much more front vowels /e/, /ɛ/, /a/, middle vowel /ə/ and back vowels /ɒ/, /u/ and /ɑ/. Compared with vowels, consonants in Portuguese has relatively lower diversity which is shared commonly in this group of languages.
Group 3: Russian, German on its own

Russian

Russian has slightly different patterns that set it from any other groups above. First, Russian has the least number of phonemes in general which could be tell by shorter and narrow phoneme pyramids. Second, it has very few front sounds.
and back sounds with most sounds happening in the middle of human speech organ. With the color notation system, a significant amount of gray could be observed here clearly.

Fig 4.5.10 Color Notation System of Phonetic Representation: Russian

German

German has slightly fewer phonemes than English but still
keeps a relatively balanced pattern with harp and tall pyramids indicating a variety of phonemes.

Fig 4.5.2 Color Notation System of Phonetic Representation: German

With simple groupings, we could see some patterns and surprising findings from it: Group 1 of English, German, Danish, French, Chinese are two Germanic languages, one Scandinavian language, one Romantic language and one East Asian language. Germanic, Romantic and Scandinavian languages are both Northern/Western European languages which might not too surprising that share some phonetic features. Yet Chinese, as the only Asian language in this group, still has similar pronouncing forms.

Group 2 includes Spanish, Italian, Japanese and Portuguese which three of them are from the Romantic language family and the rest is Japanese which is not even a European language. Even Japanese is the only Asian language here, it does share incredible similarity with Spanish and Italian based on this small piece of sample text. This pattern could be taken as one of the evidence that even some languages are spoken in countries that are far from each other, their phonetic patterns could still be similar.

Russian and German, as the only language in the last group, show pretty similar patterns as the group 1 but with fewer back sounds. It is interesting to see that even these two languages are from different language families but have similar pronouncing positions during speech. However, it might need bigger text data to further prove it. I would discuss future possibilities of including more languages and bigger data for further analysis in the Discussion and Future Work section.
4.6 From characterization to visualizing distance

After grouping the languages based on shared or similar patterns, now the question is how to visualize the distance among them.

Taking from the same data of phoneme transcriptions, I counted the frequencies of sound positions in front, middle, back position respectively. The hypothesis is that if two languages have similar distribution of positions of sounds, then their phonetic relationship could be closer compared with the ones don’t.

First, I put all ten languages in the scatter plot chart: with x-axis as the count of back sounds, y-axis as the count of front sounds and the size of the circles as the count of middle sounds.
sounds. It would look like this:

**Fig 4.6.1 Sound Position Distribution**

Take one step forward, I could start to visualize the "phonetic distance" among these languages by connecting them like below:
Group 1: English, Danish, Chinese, German, Russian

Subgroup: French (close to Group 1 but have more front and less middle sounds)

Group 2: Spanish, Italian, Portuguese

Group 3: Japanese

Different from the phoneme frequency pyramids, this visualization suggests of the ‘distance’ among languages. Take Japanese, German and Russian as examples: from the phoneme frequency pyramids, Japanese seems to have a similar pattern with the languages (Spanish, Italian and Portuguese)
in the Group 2; however, in this plotting chart which emphasizes on front sounds and back sounds, the number of back sounds in Japanese stands out to the left of the visualization which is more isolated from the rest languages than what the phoneme frequency pyramids could tell.

Another example is German and Russian. Even though from the phoneme frequency pyramids and color notation system in the previous section indicate a more distinct pattern from the rest, their plotting seem not to be too far from English, Danish and Chinese.

From this scatter plot visualization, not only the languages which are close to one another are shown, the languages that are far from each other could also be observed. For instance, if we simply look at the pronouncing positions during this speech paragraph, Spanish, Japanese and Portuguese are the ‘most far’ languages to Russian and German.

This scatterplots of distance are made to show the quantity aspect of pronouncing positions of languages. Some languages like Japanese tends to have more syllables to convey the same meaning of a sentence. This is also one of the phonetic features of a language that sets it apart from others, therefore suggests a farer distance in this example.

To further discuss each position—front, middle and back, I use size to indicate the count of sounds not only for the middle sounds but for all three categories for comparison. The results are displayed in the order from more to less and are colored with the same grouping colors as the scatterplots.

The representation shows certain alignment with the previous diagram: Spanish, Portuguese and Italian and their sub—group—French (red) are grouped either together or very close; similar pattern could be observed in the English, Chinese, Danish, German and Russian group (gray). Japanese, by itself, is closer to the red group in terms of front sounds
Fig 4.6.6 Phonetic Relationship Distance—Count as size in three positions
count; while is distant from them in terms of middle and back sounds count.

4.7 Color notation in speech organ

By marking different colors based on the backness of pronouncing sounds but still keeping the representation as sentence—based structure is one way to visualize phonemes; by counting the frequencies of sounds based on pronouncing positions is another way to visualize ‘phonetic system’ Another possible way to explain the similar concept is tracking dots on the pronouncing positions in human speech organ.

First, I count the frequencies of each phoneme of Alice’s Adventures in Wonderland in ten languages. And then correspond each of them on the relative positions according to IPA chart. The more frequent the speech positions involve, the darker the dot color would be.
Using the title 'Alice's Adventures in Wonderland' as an example again, some patterns could be observed here: Spanish, Italian and Portuguese are characterized into same group again due to the obvious darker focus in the front of human speech organ.

Fig 4.7.1 Color notation of phonetic relationships with human speech organ

Russian, French, Japanese and Chinese are grouped into 'the most balanced' category which have a relatively even
distribution pattern. The last group shows quite similar tendency with the most balanced group but with darker dots located in the middle of human articulator.

To take a closer look in each group: the points to see for patterns are: mostly front or middle or back sound, with or without obvious patterns, balanced or scattered and so on.

Group 1: Spanish, Italian, Portuguese (most front)

The result echoes back to the color notation and pyramid visualization: this group of languages has the most phonemes and the most front sounds compared to other groups. Spanish and Italian shared one of the most similar patterns with the darkest dots in the front.

Group 2: Russian, French, Japanese, Chinese (most balanced)

This group has the most balanced pattern whose phonemes involve all front, middle and back of human articulator.
Group 3: English, German, Danish (most middle)

English, German and Danish are again, assigned to the same group with the shared pattern of having the most middle sounds. This group is similar to group 2 but with slightly less front and back sounds.

4.8 Sound Spectrograms and Languages

To visualize speech sounds, the sound frequency pattern of languages could be another aspect to investigate. By recording sound spectrograms, it is allowed to include more languages (whose phonetic transcriptions are not as accessible as the ten languages in this case study) for further analysis. In this part, besides the ten languages used in the color/dot notation system, I add 'Arabic' and 'Korean' to try to show more diverse languages. I would discuss future possibilities of including more languages for further analysis in the Discussion and Future Work section.

Fig 4.8.1 Sound Spectrograms Comparison

In this section, I use 'Alice' of the translations in different languages as an example to explore the relationships of speech frequency with distinct phonemes. Each phonetic symbol has a specific frequency pattern that is shown by sound spectrograms. The front vowels /æ/ and /a/ account for a lower frequency distribution while the fricative sounds like /s/ and /z/ tend to show a higher frequency distribution. However, in the spectrogram patterns of Japanese and Korean seem not to have such high—frequency end which might be due to the following back vowel /u/ lowers the frequency distribution.

Besides consonants, to echo back to the similar concept previously mentioned, the sound spectrogram can also tell where the loud and quiet vowels are.

With a combination of two loud vowels, /a/ and /i/ in Chinese contributes one of the most noticeable red waves in the diagram. Contrast to that, relatively quiet and more back vowel /ɯ/ (pronounced as a loose /u/) frequently appears in
Japanese show a much lower mark.

Group 1: English, French, German
Group 2: Italian, Portuguese, Russian, Danish
Asian languages like Arabic, Chinese, Japanese, Korean have relatively distinct patterns from each other

4.9 Visualizing Syntax Structure
Aside from the phonetics, the other important part of the spoken language is the sentence structure. How sentences are composed is largely determined by parts of speech and word order which would be the main focus in this section. Due to *Alice in Wonderland* is a piece of literature work, the sentences which are translated into different languages might have different sentence structures that influence how the sentences sound.

4.9.1 Color Notation System and Parts of Speech—Book Title
The similar notation system that encodes different parts of speech with colors would be used here to visualize sentence order and structure. Here is the sample sentence from the book title: Alice’s Adventures in Wonderland in 24 languages:

Alice’s Adventures in Wonderland

With similar color notation, I assigned red as nouns, gray as function words like proposition or articles, and black as verbs. With a focus on content words, red and black are more notable for observing patterns. Besides color, the subject noun ‘Alice’ is also marked with a black outline for reviewing where the subject is in this phrase in different languages. Most languages listed above have subject put on the first except Hindi and Japanese. In Arabic, the subject seems to be put on the last in written form, but it is supposed to read from right to left in this language. Therefore, in this phrase I will count that the subject is put on first like other languages.
Group 1: English, Dutch, German, Swedish, Danish, Malay

Group 2: Spanish, Portuguese, French, Italian

Group 3: Russian, Czech, Polish, Romanian, Arabic, Thai, Indonesian, Vietnamese

Group 4: Japanese, Korean

Group 5 (have their own patterns do not fall in groups above): Turkish, Chinese, Korean, Japanese, Hungarian, Hindi

After grouping these languages, several interesting patterns could be found here. First, the sentence structure in this sample book title seems to follow their language family roots. Group one is composed of Germanic languages, Scandinavian languages and Malay which is famous for having many loan words from English, Dutch, Arabic and other languages because of its history (cite). The second group is all Romanic languages which tend to have a longer sentence with more articles and propositions which contain the meaning of "the" and "of". The third group is Slavic languages and surprisingly, some southeastern Asian languages. They both tend to separate "wonderland" as two words. This pattern might be a pure coincidence, or these two regions somehow had influenced each other in terms of composing sentences. Group four is Japanese and Korean which both are from Eastern Altaic language family. Their sentence structure is slightly different (Japanese has "'s" to show possession), but their general order is "strange/land/('s)/Alice".

The second interesting finding is that there is a group of languages has their own distinct patterns which are different enough from putting any of them together. Group four has Turkish, Chinese, Hungarian, Hindi which makes certain sense since they are all from different language families.

Especially in Hindi is quite different from the rest sentence structure discussed above. In Hindi, "Alice in Wonderland" translated word by word would be: The/wonderland/in/Alice. In this particular phrase, it seems that it is natural and grammatical to put objects in the front while subjects at last.
4.9.2 Color Notation System and Parts of Speech—Whole Sentence

After familiarizing the notation concept used in sentence structure, I would further analyze the whole sentence which is quoted from the first chapter of Alice's Adventures in Wonderland, "And what is the use of a book," thought Alice, "without pictures or conversation?" Here I have listed several languages as examples below to illustrate my points.

English
"And what is the use of a book," thought Alice, "without pictures or conversation?"

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

German
«Und was nützen Bücher,« dachte Alice, «ohne Bilder und Gespräche?»

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

Danish
"Og hvad fornejelse har man af en bog uden billeder, og hvor personerne ikke snakker med hinanden?" tænkte Alice.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

Italian
- e a che serve un libro, pensò Alice, - senza dialoghi né figure?

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

Spanish
«¿Y de qué sirve un libro sin dibujos ni diálogos?», se preguntaba Alicia.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

Portuguese
"e de que serve um livro", pensou Alice, "sem figuras nem diálogos".

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

French
“La belle avance,” pensait Alice, “qu’un livre sans images, sans causeries!”

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

Russian
Что толку в книжке,— подумала Алиса,— если в ней нет ни картинок, ни разговоров?

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Prop./Article</th>
</tr>
</thead>
</table>

is quoted from the first chapter of Alice's Adventures in Wonderland, "And what is the use of a book," thought Alice, "without pictures or conversation?" Here I have listed several languages as examples below to illustrate my points.

Same as the last example, I assign red as nouns, gray as function words like proposition or articles, and 'Alice' as pink
with a black outline to emphasize where the main subject is.

Fig 4.9.2.1 Color Notation System of Syntactical Representation
—Whole Sentence

When analyzing the whole sentence is a bit more challenging. Alice in Wonderland, after all, is a piece of literature work. Compared with the word—by—word precise translations, the translators seem to focus on the beauty and naturality of language use more. Patterns are harder to find because the sentence order might be completely different. This situation well explains why language analysis could be complicated and problematic to find rules.

However, if looking closely, there are still some interesting points that worth discussion:

First, I could roughly categorize two big groups that show distinct sentence order tendencies. On the left are all European languages that put verb before the agent noun—"thought Alice" in English and other seven languages; compare to the right side, three East Asian Languages that tend to put agent noun before the action verb. Two kinds of order (verb—agent or agent—verb) might be grammatical in both categories; however, there might be a reason that the verb—agent order is used in all eight European languages in this case: it sounds more natural in the context.

The second interesting pattern would be the negation—noun order. All languages have the order as negation (‘no’, or ‘without’ in English) in front of the adjective in a sentence except Japanese and Korean. Both languages would prefer a ‘noun—negation’ (images no, conversations no) order which is quite unique in sentence structure. Even as one of the East Asian languages, Chinese, on the other hand, is translated as the order that closer to the rest European languages.

The visualizations can be more complete by including as many languages as possible if the time and data permitted. However, this approach could be taken as the initial attempt to represent part of the syntactical relationships across
languages and acts as one of the reference methods for future syntax analysis of world languages.

It is surprising that so many sentence structure information could be found with a simple color notation system. Without looking into the jargons of linguistics, the combination of color blocks seems to enable people to perceive how a language could be structured and compare the differences effectively.

05 Conclusion

With the case study of using color notation system illustrating linguistic relationships across ten languages, in this chapter I would like to discuss how exactly this visualization system could help readers comprehend phonetic and some syntax relationships more effectively, what are the limitations of this system, and future opportunity to improve it.

First, though out this color notation visualization, only three colors: pink, gray and black are used. Even though this system is based on phonetics and syntax theories which have many complicated concepts, I carefully simplify the categories, keep the most important information and leave out the secondary one. For example, vowels happen in front, mid—front, middle, mid—back and back oral cavity are simplified as front (presented by pink color), middle (presented by gray color) and back (presented by black color). Similar to vowels, consonants which involve even more positions than vowels are also simplified into the same three categories. Without taking much time to remember many colors and what each means, this notation system only needs readers to remember three colors and their meanings which are fairly straightforward.

Besides simplicity, this design also enables readers to find out the parts that stand out the most more effectively. For instance, a color pattern from some language which shows a lot more gray than other languages could be easily distinguished. No matter if it is a more balanced color pattern, or a single color stands out, the use of simple colors enhances the clarity for better understanding this system.

In addition to simplicity and clarity, the concept of front
(presented by pink color), middle (presented by gray color) and back (presented by black color) is repeatedly used throughout the phonetic parts of the case study. Even though in sound spectrograms, the use of colors is different (in which pink to gray shades represents intensity from high to low), the abstract concept of leveling is in somewhat similar here. Therefore, this form of repetitively allows readers to take less time when are first introduced to the system and therefore shortens the learning curve. When introducing complex and unfamiliar concepts, it would be less painful for readers if they do not have to take too much time and efforts to learn the technical terms and the definition of theories before looking at the main picture.

On the other hand, the design of human speech organ diagram which corresponds to real human nasal and oral cavity, would also help increase readers' mental memory of phonetics concepts. Even though this static word document cannot interact with readers, the animated videos presented in my theory exhibition would allow people to enhance their muscle memory by "karaoking" with real speech sounds happen in different parts of speech organ in real time. The static screenshots of its final results in this case study section still provide a general picture of where the sounds happen in different languages when trying to pronounce a sentence with the same meaning and allow readers to compare one language to another and eventually be able to group and find patterns of their relationships.

With the limit of length and data source of this research, only ten languages are included in the case study. However, once establishing the rules and structure of the color notation system, it would be relatively easier to apply on more languages with the same system without extra learning which would be discussed more in the "Discussion & Future Work" section.

After concluding the advantages of each method in the case study, I would also like to explain why I use several methods to analysis basically the same piece of data. In the
case study, it presents five different approaches to look at phonetic and syntactical relationships of languages. First, phoneme frequency in different positions with pyramid shapes; second, color blocks notation of sentences; third, color dots notation in speech organ; fourth, the sound spectrograms showing sound frequency distributions, and the last, color blocks notation of sentence structure with parts of speech. The raw text data might not be big, yet by looking at linguistic relationships from multiple angles allow readers to compare and see different aspects of languages. Since language is one of the most complicated human inventions ever. Any research on language has to be very careful. It would be dangerous to only see one single theory or visualization showing language relationships and make assumptions based on that. In the related work of lexical relationship would work as an example here to demonstrate this point: only use a vocabulary list as the sole data and map out lexical network can, therefore, only show parts of the relationship. Language relationships are often associated with serious and sensitive social issues like national identity, history and even territory politics. Therefore, the theories and visualizations of language have to be extra cautious during the process of composing. Using
one approach might inevitably have bias but using multiple different perspectives to look at the same data might be a way out for representations that are closer to reality.

06 Discussion & Future Work

In the discussion section, I would like to discuss if the phonetic symbols are the proper data that are used for visualization in the thesis, are there other possibilities to represent sound better, how the color notation systems of phonetic relationships could potentially benefit language learning, the limitations on data acquiring and the last, what technology and graphic design approaches might be useful to generate with bigger dataset for more accurate analysis.

6.1 Concern of Data

First, in this document, I use phonetic symbols on IPA chart to represent small segments of sounds which would be described as phonemes in linguistics. IPA chart is so far the only standardized phonetic symbol table that is widely used for phonetics analysis. However, when first time looking at the table, a lot of the symbols are not as instinctual enough that allows people to have a clear clue to pronounce them. In English, the phonetic symbol transcriptions tend to be relatively more familiar to English speakers compared with other European languages and Asian languages. It needs more knowledge and familiarity of IPA chart in order to read non—English phonetic transcriptions.

**English**
æləsəz ædəvɛntʃərz ɪn wʌndərlænd

**Russian**
priklютɕ eɲiɭ ø li iʃ fstrɐnʲ e ɕtʊdɨ es

**Japanese**
ɸɯᵝ ɕiŋino kɯᵝnʲino aɾʲisɯᵝ
Therefore, on one hand, phonetic symbol is a proper source of data due to its standardization; but on the other hand, it would be even better if the symbols themselves are more instinctual or give more hints of pronouncing them.

Besides phonetic symbols on IPA chart, what else could be the possible and appropriate substitutes? Romanized letters might be one of the possibilities. When seeing the phonetic symbol of three languages above, it is not straightforward of how to pronounce them. However, if we use Romanized letters, it would be a lot easier to pronounce for anyone who are familiar with English alphabets.

English
Alice's Adventures in Wonderland

Russian
Priključenija Alisy v strane čudes

Japanese
Fushigi no koku no Arisu

It is not difficult to find that the Romanized letter transcriptions are a lot easier to pronounce for whoever is familiar with English compared with the phonetic symbols. Then here comes the question: why not just using Romanized letter for the case study? The only downside of Romanized letters would be the transcription is easy to pronounce for English speakers, but it certainly loses accuracy of the real pronunciations through the transcription. Take 'Alice' in Japanese for example: in Romanized letter transcription, it is transcribed as 'Arisu'. One of the obvious inaccuracies of pronunciation is that instead of the /r/ sound, it is always pronounced as a flap /ɾ/ sound which is more of a sound between /l/ and /r/ in Japanese. This inaccuracy is able to be represented with phonetic symbols from IPA chart but not Romanized letters. Therefore, using Romanized letter has its pros and cons: it is easy to imagine how the sentences of foreign language sound like, but with the drawback that the transcription is not accurate compared
with the real pronunciations. In the case study, I tried out different approaches and finally decided the accuracy of pronunciation is the primary concern of the thesis. Instead of having inaccurate data but more understandable result representations, I would prefer using relatively more accurate data but with more explanations of what I have found.

6.2 Limitations of Visualization

After justifying the reasons that I choose phonetic symbols as the main data source, I would discuss how the color notation systems of phonetic representations could potentially help language learning. Through the visualizations in the case study, it can tell that some languages have more balanced positions involved when pronouncing, while some focus on only several parts of speech organ. When people learning to speak another language, it is relatively easy for them to pronounce the sounds that already exist in their mother tongue. To contrast, it is naturally more difficult when the speakers of the most—front—sound languages want to pronounce back sounds. As long as learner recognize these differences, it would give them some clue of how to practice the difficult sounds, for example, practice to produce sounds from the back of the articulator.

However, to effectively utilize the notation systems for language learning or other purposes, it is also important to recognize the limitations of data acquiring. Like I previously mentioned in the 'Research Background' section, I mainly acquire the phonetic symbol transcriptions through a language learning website called EasyPronunciation.com which provides resources for learning pronunciations of different languages online. The website targets language learning which might be likely influenced by political, cultural and economic power. The more powerful a country is, it naturally draws more foreign learners to learn its language. It is challenging for me to find Middle Eastern, African, Southeastern Asian, and Pacific language open sources translated in English. I am aware that with only ten languages which mostly are European languages is definitely not enough to have an overlook on world language
relationships. A lot of languages from the regions I mentioned above have a great number of speakers are not included in my color notation system which I would try to include them into the sound spectrograms and syntax analysis.

6.3 Future Opportunities

Aside from the limitation of data source, one of the parts that the case study could further research in the future is that using machine learning to handle bigger text data and identify the linguistic relationships across languages more efficiently. With bigger text data pool would help avoid bias resulted from smaller datasets and eventually pulls out even more accurate representations. Besides, if the machine learning could combine to real—time data acquisition, it would not only efficiently represent relationships of languages, but also be able to reflect how language relationships have changed overtime. With globalization, do languages become more similar to one another?

The other possibilities of linguistic relationships research could be focused on dialects and accents. For example, why some stereotype says that American English is somehow ‘louder’ than British English? With little research on the differences in their pronunciation rules, some clue might be able to explain it. Take ‘clock’ for instance, ‘clock’ is pronounced as /kläk/ in American English and /kloʊk/ in British English. The vowel in American English is a front vowel /ä/ which is naturally more audible than back vowel /ɒ/ in British English. Would this be a legit reason to explain the stereotype that American English is louder? It might need more evidence and research on bigger context and data to prove. One way to dig in further on this topic could be combing recording data and machine automated analyzing. First, we could record multiple native speakers from different regions but speak one language. Take English again, as the example: record same text of speech from English native speakers from the UK, Australia, South Africa and the US; and then assign machine to learn "the differences" among all recordings to either establish a scoring
system or other ways to determine the distance.

In addition to technology and various topics that worth discussing, it would be very interesting to explore more graphical design in the implement of notation system. Not just colors, textures, traces animation or other patterns could also potentially help visual understanding of the concept which makes the visualization more dynamic and easy to compare their relationships.

As for the syntactical aspect of language, word frequency is another important point that worths further analyzing besides word order in sentence structure. In order to convey meaning, sentences are composed of both content words (like noun, verb, adjective) and function words (like article, preposition) in many languages. However, different languages have their own way to construct a sentence: some might have way more function words than another, or some only use very few function words repeatedly. The visualizations to illustrate this difference and its distance would be another important piece of this big world language puzzle.
Bibliography

Bell, Alexander Melville. *Visible Speech, the Science of Universal Alphabetics: or, Self—Interpreting Physiological Letters, for the Writing of All Languages in One Alphabet*, Illus. by Tables, Diagrams, and Examples. Simpkin, 1867.


