Turkish Vowel Harmony and Disharmony Exemplified Through Neural Networks

Matthew Buzzell
Indiana Univeristy
matabuzz@iu.edu

Abstract

Vowel harmony is a very prominent phonological alternation in Turkish. Given its relevancy, we must understand how it functions within today's ecosystem. With the global push towards artificial intelligence and generative language models, it becomes ever more important to understand how phonological alternations are patterned within linguistic models. Specifically, this paper will focus on how vowel harmony is modeled in Turkish within a sequence-to-sequence recurrent neural network (RNN).

However, before modeling Turkish vowel harmony within an RNN it is essential to understand how vowel harmony functions in Turkish. To this end, I will first describe the traditional autosegmental account of Turkish vowel harmony. This will include an analysis for [BACK] vowel harmony and [ROUND] vowel harmony. I will then describe Turkish vowel disharmony as it appears in both Turkish roots and opaque suffixes. This description will include the specific restrictions on vowel disharmony as it pertains to the marked and unmarked vowels.

After this explanation, I will describe the specifications of the experiment. This will include a description of the experiment, the specifications of the training/test data, and a summary of the analysis code to be run on the output of the RNN. Additionally, I will include the results from the experiment and an analysis of those results.

By the end of this paper, I will show the accuracy RNN models have when constructing novel Turkish words. Moreover, these constructed Turkish words will not only show the acumen neural networks possess in learning phonological alternations, but also their flexibility to learn the restrictions on these alternations. On top of this, all these phonological rules and exceptions will be learned without hand-crafted formulas, but solely based on a simple input of a series of Turkish words.

1 Turkish Orthography

As a general policy, authors on Turkish phonology tend to use Turkish orthography to formulate their papers. Given this, before covering Turkish vowel harmony I will quickly go over Turkish orthography, specifically how the orthography relates to IPA. Additionally, I will use Turkish orthography in place of IPA as that will ensure my notation is consistent with the figures and quotes I use from other authors.

Turkish has eight vowels in its vocabulary: [a, e, i, o, u, ı, ü, and ö]. It is important to recognize that the vowel [ı] is different from the vowel [i] despite looking similar. In terms of terminology, [ı] is usually referred to as the "dotless i". Moreover, these vowels can be broken up into three different categories of four, exemplified by Figure 1. This figure can be related to a prototypical IPA vowel chart to understand where each of these letters falls, but for the purposes of vowel harmony, this chart will suffice.

Having described the vowels of Turkish and their corresponding features, we can now describe Turkish vowel harmony using Turkish orthography.

2 Vowel Harmony

Vowel harmony is a phonological phenomenon that has been widely studied by linguists such

Orthography	[-BACK]		[+BACK]	
	[-ROUND]	[+ROUND]	[-ROUND]	[+ROUND]
[+HIGH]	[i]	[ü]	[ı]	[u]
[-HIGH]	[e]	[ö]	[a]	[o]

Figure 1: A chart showing the vowels in Turkish distributed by their phonological features [BACK], [ROUND], and [HIGH].

as Robert Kirchner (**?**), John Goldsmith (**?**), and Öner Özçelik (**?**), but for our primitive autosegmental notation of vowel harmony, we will turn to Clements and Sezer (**?**).

Vowel harmony in Turkish is broken up into two categories: [BACK] vowel harmony and [ROUND] vowel harmony. [BACK] vowel harmony occurs when the [BACK] feature from one vowel spreads to another vowel. Using Figure 2 as an example, the Turkish word for 'stalk' is [sap]. When adding the suffix /-In/ the word becomes [sap-ın][1]. This is because the [+HIGH] vowel in the suffix takes the [+BACK] feature from the vowel in the root word ([a]) due to vowel harmony. As for [ROUND] vowel harmony, it is a bit more restrictive than [BACK] vowel harmony. [ROUND] vowel harmony occurs when the [ROUND] feature spreads from one vowel to another [HIGH] vowel. Referring once again to Figure 2, the word for 'end' is [son]. When adding the suffix /-In/ the word becomes [son-un]. This is because the suffix vowel is [+HIGH], meaning that the [+ROUND] feature from the vowel in [son] can spread to the suffix vowel. Conversely, when adding the suffix /-lEr/ to the word [son] the word becomes [son-lar]. The suffix in this case does not have the [+ROUND] feature because the vowel in the suffix /-lEr/ is not [+HIGH], meaning that the [+ROUND] feature cannot spread.

This accounts for a primitive autosegmental analysis of Turkish vowel harmony. Consequently, for our RNN to be considered accurate, it must generate words that follow these rules. However, not all words in Turkish follow vowel harmony.

[1]/I/ is a Turkilogical shorthand for a [+HIGH] vowel, whereas /E/ is a Turkological shorthand for a [-HIGH] vowel.

3 Vowel Disharmony

While vowel harmony is the general rule of thumb when it comes to Turkish vowels, there are a variety of roots and suffixes that violate vowel harmony. However, these violations are not unrestricted. The first restriction to vowel disharmony is that it is restricted to the vowels [a, e, i, o, u]. These vowels are usually referred to as the unmarked vowels. Conversely, the vowels [ı, ü, ö] are usually referred to as the marked vowels. Words that violate vowel harmony and contain a marked vowel are frequently regularized. As per Figure 3, the Turkish word for 'communism' is [komünizim] which violates vowel harmony ([o] is [+BACK] whereas [ü] and [i] are [-BACK] and [o] and [ü] are [+ROUND] whereas [i] is [-ROUND]). However, this word is regularized to [kominizim] where the [ü] is changed to an [i]. Notice that the regularized word still violates vowel harmony ([o] is [+BACK] and [+ROUND] whereas [i] is [-BACK] and [-ROUND]), but it does not contain a marked vowel. However, it is not always the case that the marked vowels are regularized to not violate vowel harmony. For example, the words for 'hope' ([ümit]) and 'typhus' ([tifüs]) are perfectly acceptable in Turkish even though both of these words violate vowel harmony ([ü] is [+ROUND] whereas [i] is [-ROUND]). This lends itself to the generalization that marked vowels do not occur in words that violate [BACK] vowel harmony. Additionally, the vowel [ı] is slightly more restrictive than the other marked vowels. This is because there are no Turkish roots that contain a [ı] and violate [ROUND] vowel harmony.

On top of this, Turkish has "opaque suffixes" that also violate vowel harmony. The reason these suffixes violate vowel harmony is because they contain prespecified vowels. As an example from

'rope'	'girl'	'face'	'stamp'	'hand'	'stalk'	'village'	'end'	
ip-in	kız-ın	yüz-ün	pul-un	eļ-in	sap-ın	ķöy-ün	son-un	(gen.sg.)
ip-ļer	kız-lar	yüz-ļer	pul-lar	eļ-ļer	sap-lar	ķöy-ļer	son-lar	(nom.pl.)

Figure 2: Examples of vowel harmony between a root and the genitive singular suffix /-In/ (the 2nd row) and the nominative plural suffix /-lEr/ (3rd row).

komünizim ~ kominizim	'communism'
mersörize ~ merserize	'mercerized'
püro ~ puro	'cigar'
ķüļot ~ ķiļot	'panties'
nüzuļ ~ nüzüļ	'paralysis'
nüfus ~ nufus	'population'
kupür ~ küpür	'denomination, clipping'
motör ~ motor	'engine, motorboat'
šoför ~ šöför	'driver'
bisķüvit ~ büsķüvüt	'biscuit'
sövaļye ~ sovaļye	'knight'

Figure 3: A list of Turkish words that undergo regularization.

Figure 4, for the suffix [-Iyor], the second vowel [o] is prespecified, meaning that it will always surface as the vowel [o]. Consequently, this leads to vowel disharmony as in the Turkish word for 'I am coming' which is [gel-iyor-um]. This violates vowel harmony as the first two vowels [e] and [i] are [-BACK] whereas the final two vowels [o] and [u] are [+BACK]. Additionally, the vowel in the final suffix obeys vowel harmony with the rightmost vowel in the opaque suffix. However, looking over Figure 4, the generalization that marked vowels do not occur in opaque suffixes is made apparent.

Summarizing these generalizations for vowel disharmony tells us that: first, roots that are [BACK] vowel disharmonic will not contain the marked vowels; second, roots that are [ROUND] vowel disharmonic will not contain the vowel [ı]; and finally, opaque suffixes that violate vowel harmony will not contain the marked vowels. While our neural network must generate Turkish words that follow vowel harmony, for our neural network to be considered accurate, it must be robust enough to recognize these restrictions on vowel harmony.

4 Neural Network Architecture

To properly assess the accuracy of a neural network to generate Turkish words I ran an experiment where the output of a sequence-to-sequence RNN trained on Turkish morphemes was evaluated against specific phonological criteria. However, before explaining the results of the experiment, I will describe the specifications of the neural network. To train and test my model I used the Kenet Turkish treebank from Universal Dependencies, which is the largest treebank in Turkish consisting of 178,700 tokens. From this treebank, I used the surface forms of words which can be found in the second column of the .conllu files. After collecting the training data, I created a neural model that takes in the surface forms of words and trains a neural network that models Turkish words. The neural network was a sequence-to-sequence RNN model called Char RNN PyTorch created by Nikhil Barhate. The neural network is a character-level language model using multi-layer LSTMs in PyTorch. When the RNN is trained it will predict the next letter in the sequence making it an ideal candidate to analyze whether a neural network can learn to predict the vowel harmonic sequence. Additionally, once trained it can generate a new text sequence which can then be analyzed for vowel harmonicity.

5 Results

After training and running the test program, the results extracted were as shown in Figure 5. A total of 70,983 characters were generated, of which, 39 characters were unique. Additionally, of the 105 words generated only 9 words violated vowel

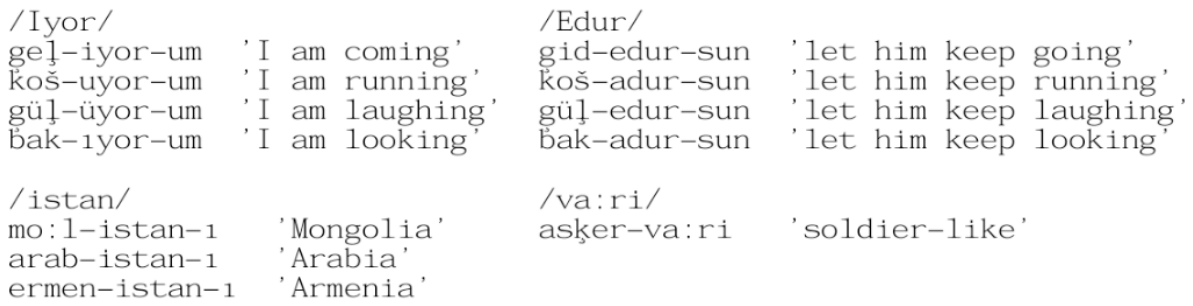

Figure 4: A list of varying opaque suffixes and their different use cases.

harmony. This statistic shows a 91.43% accuracy rate when it comes to generating words that adhere to Turkish vowel harmony. However, per our criteria, there are exceptions to vowel harmony that are acceptable in Turkish. To accurately assess whether the RNN is accurately generating Turkish words these vowel disharmonic sequences need to be evaluated.

The generated words that violated vowel harmony were [hikumluğunun], [ihtimale], [başhekimi], [telefonu], [üstününun], [noktacıklar], [teperşü], [söylemeyi], and [tutuyorlar]. Despite violating vowel harmony, most of these words are still consistent since their vowel disharmonic sequences are licit in Turkish. To determine whether these generated words were licit in Turkish I spoke with a native speaker of Turkish, Onat Zeybek Kuşkonmaz, to determine what the roots of these generated words might be, as this is an important criterion for determining whether a vowel disharmonic sequence is acceptable.

Starting first with the word [hikumluğunun], this word violates [BACK] and [ROUND] vowel harmony as [i] is both [-BACK] and [-ROUND] whereas [u] is both [+BACK] and [+ROUND]. However, according to the Turkish native speaker, the root of this generated word would be [hikum]. This word then becomes perfectly acceptable in Turkish. The suffixes in this word don't violate vowel harmony since they are consistent with the rightmost vowel in the root. Additionally, the root is acceptable since, while it does violate [BACK] and [ROUND] vowel harmony, it does not include any of the marked vowels.

Next, the word [ihtimale] violates [BACK] vowel harmony as [i] and [e] are both [-BACK] vowels whereas [a] is a [+BACK] vowel. However, if this generated word is the root then it would not violate the restrictions of vowel disharmony since it does not contain any marked vowels[2].

Finally, the word [tutuyorlar] presents our first case of an opaque suffix. According to the Turkish native speaker, the root of this word is [tut] with the suffixes [-uyor] and [-lar]. The suffix [-uyor] is an opaque suffix where the [o] is prespecified. Because of this, we see that the first vowel in the suffix ([u]) is vowel harmonic with the vowel in the root ([u]) and the vowel in the second suffix ([a]) is harmonic with the second vowel in the opaque suffix ([o])[3]. Therefore this word is consistent with the restrictions on vowel disharmony.

To summarize, only one word generated by the neural network that violated vowel harmony also violated the restrictions on vowel disharmony. This means that the neural network achieved a 99.05% accuracy rate when generating novel Turkish words that are consistent with vowel harmony and disharmony.

6 Conclusion

Given the global impact of generative language models and neural networks, it becomes ever more important to demonstrate their efficacy on a variety of different inputs. Throughout this paper, I have explained how vowel harmony works specifically in Turkish and showed the array of different criteria a generative neural model would need to abide by to sufficiently generate accurate Turkish morphemes. Additionally, I ran an experiment where a sequence-to-sequence RNN model was trained and tested on Turkish words to see if it

[2]According to my Turkish native speaker, the final vowel [e] in this word may be a suffix, in which case, this word would violate the restrictions on vowel disharmony.

[3]The vowel in the suffix [-lar] is not a high vowel which is why the [ROUND] feature is not spread from the opaque suffix to the following suffix.

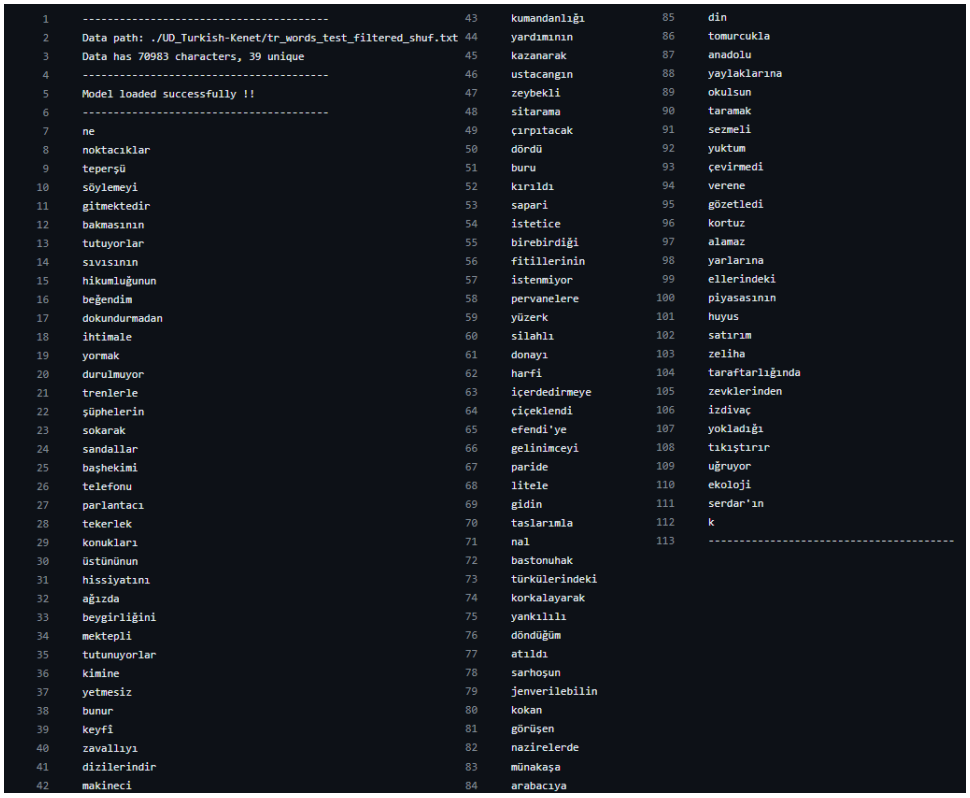

Figure 5: The output from the RNN after running the test program.

could accurately generate Turkish vowel harmony and disharmony. Through this test, we found that the RNN generated accurate Turkish vowel harmony and disharmony at a percentage of 99.05%. This demonstrates the robustness of neural models and how valuable they can be in future Turkish morpheme generation. Additionally, this experiment could be extended to a vast array of differing types of vowel harmony as well as different phonological alternations.

Acknowledgments