Klingon, Elvish, German, English, and Math

Olin College of Engineering Linearity I Rebecca Jordan & Marie-Caroline Finke

In Linearity, we are learning about Singular Value Decomposition and its appplications. Inspired by the paper "Singular Vectors' Subtle Secrets", we decided to extend a cryptographical analysis of letter frequency in English to other languages. Klingon is a language invented for the science-fiction television show Star Trek, and Sindarin is an Elvish dialect invented by J.R.R. Tolkien for his fantasy world of Middle-Earth. Both languages have been developed to a degree where people can speak and write them fluently. We are interested to see how the letter patterns of these invented languages compare to those of natural languages, such as German and English.

Singular Value Decomposition

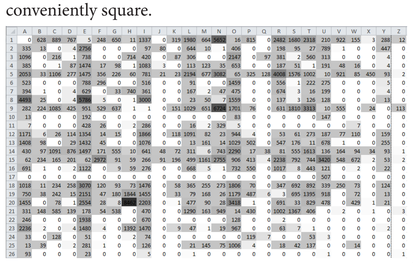
Singular Value Decomposition is a method of decomposing a matrix of any shape into a series of rank one matrices (matrices with only one linearly independent column.) The sum of the rank one matrices is equal to the original matrix.

The first step is finding three specific matrices, U, Σ, and V that decompose the given matrix A such that $A = U\Sigma V^T$. This is very similar to the $A = PDP^{-1}$ decomposition, where P is the horizontal concatenation of the eigenvectors of A, and D is a diagonal matrix of the eigenvalues associated with each eigenvector. Since it is not possible to find the eigenvectors of a nonsquare matrix, instead the eigenvectors of square matrices AA^T and A^TA are found.

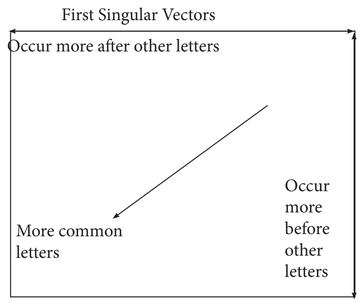
The nonzero eigenvalues of AA^T and A^TA are the same. Their associated eigenvectors u_i and v_i are related by $u_i = Av_i$. The columns of U, an NxN matrix, are the eigenvectors of AA^T in order of decreasing associated eigenvalues, and the columns of V, an MxM matrix, are the associated eigenvectors of A^TA. The singular values of A are the square roots of the eigenvalues of AA^T and A^TA. Σ is a diagonal matrix composed of the singular values of A and 'filled in' with zeros to have size MxN. The singular value $\sigma_i = \sqrt{\lambda_i}$. The best rank 1 approximation of A can then be calculated as $\sigma_i u_i v_i^T$. The best rank 2 approximation is $\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$, and so on.

How We Use It

We use a Python script to generate an adjacency matrix (such as the one below for English) showing how frequently one letter followes another in a large sample text. This matrix is conveniently square.

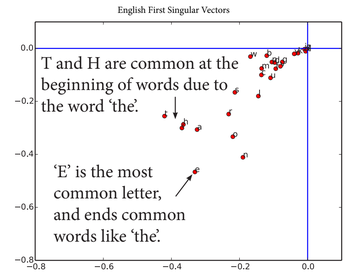


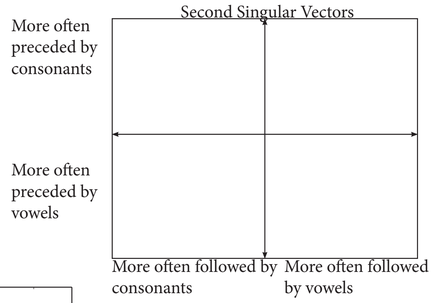
We then find U, Σ, and V for the adjacency matrix. The coordinates for a letter on the first graph are (u_{1n}, v_{1n}), where u_i is the first column vector in U, v_i is the first column vector in V, and n is the index of the letter in the alphabet (so A would be 0, B would be 1, etc.) The second graph is the same except for using (u_{2n}, v_{2n}), from the second columns of U and V.



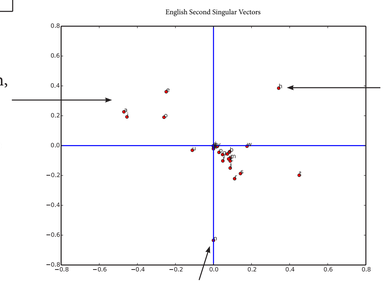
English

The most common letters in English are usually ETAOIN SHRDLU. Our method of analyzing the text ignores single-letter words, so A and I are underrepresented.





In both English and German, vowels usually fall between two consonants and consonants fall between two vowels.



In a first-rank approximation, all the rows are scalar multiples of v_1^T and all the columns are scalar multiples of u_1. Therefore, the ith element of u_1 increases with the frequency that the ith letter in the alphabet precedes other letters, and the ith element of v_1 increases with the frequency that the ith letter in the alphabet follows other letters. Plotting these two values on the Cartesian plane as (u_1, v_1) shows both the relative frequency of the letters, with more common letters landing further from the origin, and the frequency of those letters beginning or ending a word, with letters occurring mostly at the beginning of words landing further from the x-axis and letters occurring mostly at the end of words landing further from the y-axis.

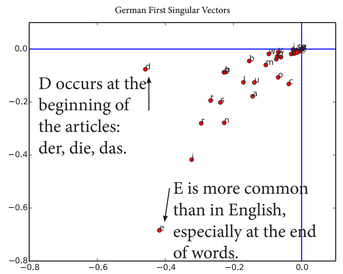
The best rank two approximation of the original adjacency matrix A is $\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$, so our graphs of the i^{th} values of the second vectors of U and V show each letter's deviation from the point indicated by the first graphs. This turns out to show what letters are most commonly preceded by a vowel or a consonant.

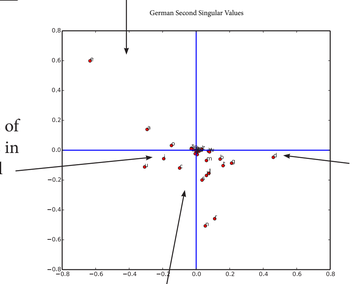
The vowels are identified as the letters that surpass the average frequency per letter for the most letters. For example, in English the letter A occurs before the letter B significantly more often than the average letter occurring before B. In fact, A has a higher frequency than the average for being in front of 18 letters, more than half. A similar pattern can be seen for all vowels. After using this method to separate vowels from consonants, the second singular value is able to sort what type of letter a letter is preceded and followed by. In English and German vowels are typically preceded and followed by consonants and thus show up in quadrant II of the secondary singular value graph. A letter like h, which in English is often preceded by t and followed by e, shows up in quadrant I. Consonants show up in quadrant IV as they are preceded and followed by vowels and letters that are preceded by vowels and followed by consonants show up in quadrant III.

One notable exception to this is in Elvish, where vowels occur in quadrant IV, consonants in quadrant II, and h, also preceded by t and followed by vowels, in quadrant III. This is most likely because the most common letter is n, a consonant, and all other letters are defined by their propensity for appearing around it. Other consonants do not appear near it, and are sorted into the same quadrant, while vowels are sorted into the quadrant opposite. This results in a graph flipped over the y=x axis.

German

German, like English, has a few very common letters, including vowels, that show up in articles and other common words.



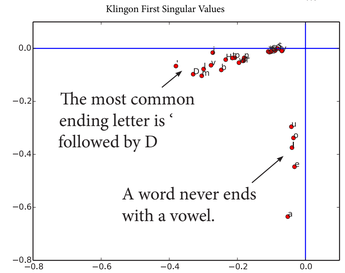


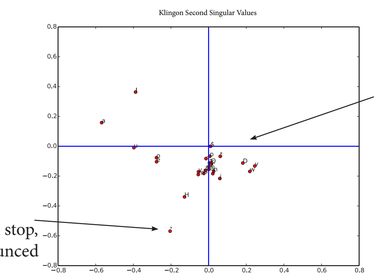
One distinct difference between German and English is German's strings of consonants without vowels in between. English's diagonal trend is not as clear in German.

The creators of Klingon may have been trying to copy the stereotypically 'harsh' sound of German by creating strings of consonants, resulting in a similar pattern on these graphs.

Klingon

Klingon has the most artificial separation between vowels and consonants, representing the strict rules used to create it.





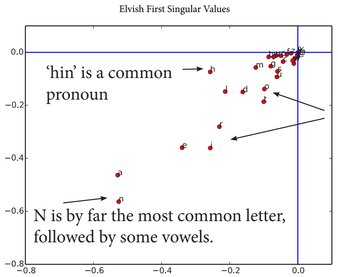
'I' and 'O' are single-letter words and therefore underrepresented.

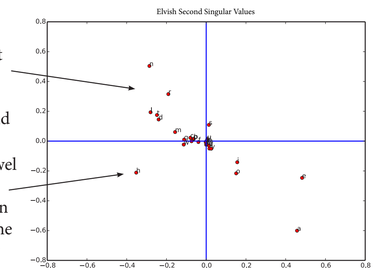
The letter ' , which represents a glottal stop, can only be pronounced after a vowel.

While Klingon was constructed to sound harsh, Elvish was constructed to sound graceful. Elvish has many sounds in common with English to ensure it sounds pleasant to English-speakers, while Klingon is meant to alienate English-speakers.

Elvish

Elvish shows much less separation between beginning and ending letters. Its distribution is more natural than Klingon's, but not as varied as English or German.





Conclusion

This Singular Value Decomposition analysis revealed some fascinating patterns in these languages. In the first graphs we can easily identify the most common letters, and in the second we can see patterns of vowels and consonants. Between these two, we can see a graphical representation of what the language sounds like. German and Klingon have combinations of consonants, while English and Elvish follow a stricter pattern of alternating consonants and vowels. The influence of common articles, pronouns, and suffixes causes outliers like 'h' in English and Elvish.

The combination of these patterns reveals key differences between the two constructed languages. J.R.R. Tolkien was known for his careful incorporation of naturalistic elements and irregularity into his constructed languages, and it shows on the graphs. Elvish has letter patterns and outliers similar to those of English, but different enough that they clearly constitute another language, just like German's. This was probably meant to give Elvish a pleasant but foreign sound to English-speakers. In comparison, Klingon's rules are very strict. The language was constructed primarily by fans of Star Trek, and while some variance was introduced by different contributors, it is a less natural and more prescriptive language. It was constructed to sound harsh and alien, so its letter patterns and even alphabet are very different from those of English. Its letter patterns more closely resemble those of German, which English-speakers tend to regard as harsh-sounding.