CMPT318
Eleanor Lewis
301300766

Character Frequency Analysis on the Inscriptions of Aphrodisias Dataset

The Inscriptions of Aphrodisias (IAph) are a publicly available set of over 1500 documented inscriptions collected from the site of the ancient Greek town Aphrodisias, situated in what is now Turkey. The inscriptions are found on a variety of structures including the town's temples, theatre, and monuments, and are by a variety of authors in contexts including liturgical, funerary, and informal - what would now be called 'graffiti.' The majority of the inscriptions are in Greek, from about 200 BCE to about 600 CE, with a few inscriptions dated earlier or later. The collection is thus a sizable sample (>200,000 characters) of Koine and early medieval Greek. Since about a third of the inscriptions are confidently dated to within 100 years, and since all the inscriptions are from a single small town, regional differences in dialect are minimized as a source of variation, and the collection is here approached as a possible window on language change over time. Greek in the Koine period is known to have undergone a vowel shift; therefore, this analysis focuses on vowel and diphthong frequency in dated inscriptions. Using the bag-of-words approach, the collection is examined for signs of change over time, and vowel and diphthong frequency is tested as a possible means of dating sufficiently large text samples.

The IAph collection was downloaded as a zipped folder of xml files in 'epidoc' format, a custom format developed by classicists to standardize digital representations of text sources such as inscriptions and papyri. IAph is a relatively early example of the use of 'epidoc' format, which seems not to have been fully standardized at the time. Additional problems during the extraction phase stem from human error in applying the format and from typos. These files are a rich source of other types of information, but extraction to pandas DataFrame is difficult due to variable individual file structure (in terms of the presence/absence/number of specific tags). Future work with collections in this format could begin by 'methodizing' the extraction of all content to Spark or pandas DataFrame, and could look at sorting by tags such as location or context (liturgical, funerary, … etc., as mentioned above). Here, information extracted was limited to the Greek text of the inscription (with no distinction between clear and unclear characters; this will be addressed in the conclusion) and its earliest and latest possible date (or "no value" if unknown). Files containing no text were excluded from the initial DataFrame. Rows containing no date were separated from the working data and saved in their own DataFrame for later use. Rows containing inscriptions with possible date range greater than 100 years were similarly separated and saved. Remaining were 547 inscriptions with date known to within 100 years. This set was used for feature extraction and training.

To ensure sample pools of adequate size, the inscriptions were sorted into categories defined by the closest hundredth year. Since information was available for both ‘earliest possible’ and ‘latest possible’ dates, analysis was done and results are shown for both categorizations.

Consonants and stray English characters were removed from the Greek text strings and replaced with whitespace using a language library for ancient Greek in Python, greek-accentuation. The same library was used to remove accents from the remaining vowels. Accents are generally not present in the original inscriptions; they are added by later scholars to aid in pronunciation. Unfortunately this is not universally true - some types of accents were occasionally included in text at the time and might have been in the original inscriptions. Further work with this type of collection might usefully stretch to OCR on the images of the inscriptions that are sometimes associated with the epidoc files, bypassing human transliteration.

At this stage the text strings consisted of vowels clustered as they are in the original text. First, analysis was done on vowel frequency without regard to vowel clusters. This process and the results are shown in the notebook ‘analyze_vowel_frequency.ipynb’. As noted, both ‘earliest possible’ and ‘latest possible’ date categorizations are used. Also shown are the DataFrames containing the raw character counts and the character frequencies. An unexpected result here was the relatively steady increase in frequency of the omicron character (‘o’). Linear regression was performed on this character; p-values for ‘earliest possible’ and ‘latest possible’ categorizations were 0.599 and 0.526 respectively. Further attempts to characterize vowel frequency change were not made at this time, both to avoid p-hacking and because these results are not needed for training a classifier.

Second, the analysis was done on vowel frequency with diphthongs counted separately from single vowels. This was done by tokenizing the text strings on whitespace and then separating the vowels clusters into their component parts using the greek-accentuation library’s ‘syllabify’ method, based on the assumption that the vowel clusters were all ‘legal’ and that diphthongs could therefore be legitimately distinguished from individual vowels this way. The character frequencies are plotted and the raw count and frequency DataFrames are shown in the notebook ‘analyze_vowel_cluster_frequency.ipynb’.

The data was transformed using the Tfidf transformer to put more weight on more variable features. The multinomial NaiveBayes classifier was trained on the data and scored on the test set. In both cases, with and without tfidf transformation, the score was about 30%. By increasing minimum test string length to 400, the score was increased to about 40%. This is a frustrating outcome; possibly longer test strings and a bigger text sample would improve scores, but as it stands this is not a useful result.

In addition to the possibilities with this type of dataset mentioned above, further work specifically with the text strings might include analysis of consonant and consonant cluster frequency, as well as n-gram analysis on syllables and n-gram-like analysis of vowel sequences within words (that is, syllable sequences within words, omitting consonants).

Finally, it is a truism in archaeology that ‘three stones make a wall.’ That is, given how little data there often is, aggressive extrapolation is common and even expected. However, it is important to note, at the risk of undermining even these negligible results, that bias is present in the epidoc source material even before this analysis begins. As noted above, where characters in the original inscriptions are damaged, unclear, or missing, researchers may include, with the rest of the Greek text, their guess as to what was there originally. This guess may be informed by the researcher’s predictions based on the date of the inscription which may be known from other sources (the age of the object inscribed or events noted in the inscription, for instance). Also, when dates are not known, they may be guessed partly based on the researcher’s beliefs about what qualities of the text are typical of what dates. So, there is a risk of developing a model for prediction based on ‘data’ that is already a prediction. Further work should be done with care to avoid this.

References

Joyce Reynolds, Charlotte Roueché, Gabriel Bodard, Inscriptions of Aphrodisias (2007), available <http://insaph.kcl.ac.uk/iaph2007>, ISB 978-1-897747-19-3.

Aphrodisias - A Roman city in modern Turkey, http://aphrodisias.classics.ox.ac.uk/, accessed Aug 8, 2017

Scikit-learn developers, The Bag of Words representation, http://scikit-learn.org/stable/modules/feature_extraction.html#the-bag-of-words-representation, accessed Aug 8, 2017

McLean, B. H. An Introduction to Greek Epigraphy of the Hellenistic and Roman Periods from Alexander the Great down to the Reign of Constantine. Ann Arbor, University of Michigan Press, 2002.