Natural Language Counting using Non-Corpus Based Models for the Voynich Manuscript

Brian Kitano

May 6, 2018

1 Abstract

Automated language identification and segmentation is a well established field of study. [1] However, most existing methods rely on corpus-based training to identify a summary statistic, which is then compared to the document in question. In addition, little investigation has been done into the analysis of multilingual documents. [2] In this paper, we describe a preliminary approach to determining the number of languages present in a document without any prior assumptions of how many or which languages may be present. We conducted experiments to test the approach by generating a dataset of multilingual documents using the Cross Language Dataset [3], which achieved visually compelling indications of success. We then applied the method on the Voynich Manuscript [4] in order to determine how many languages are present; our results were inconclusive. Finally, we discuss problems, possible improvements, and further avenues of analysis for our method.

2 Introduction / Motivation

The Voynich Manuscript (VMS) is a 15th century manuscript written in an unidentified script and it has eluded cryptographers, medievalists, linguists and hobbyists for its 600 year long history. There are hundreds of claims of decipherment, ranging from Pahlavi [5] to Nahuatl [6]. A dominant majority of these hypotheses are predicated on the manuscript being in a single language. However, Currier (1976) identifies two statistically distinct languages "A" and "B", with perhaps as many as a dozen different "hands," or unique writing styles of the script. [7] Currier also segments the folios of the VMS into groups based on his hypothesized language identifications; these folio groups are often referred to distinctly in proceeding literature.

Since the creation of a number of computer-readable transcriptions for VMS beginning in the 1940s, approaches based on computational linguistsics have proliferated, which have lead to surprising insights on the statistical structure of the VMS. [4][8][9] These studies have shown that the VMS conforms to some statistical phenomena that natural languages exhibit, such as word distribution, but also contains aberrant structure, like a particularly low character entropy or the non-random distribution of words with respect to their position on the page. [8]

Since Currier, a number of scholars have attempted to verify his results by identifying statistical differences between all of the folios, correlating them to his "A" and "B" folio group distinction, and claiming that these differences are indicative of the groups being in different languages or dialects. For example, Reddy and Knight (2011) use a bigram Hidden Markov Model to segment words into one class or another; they find that folios contain a large proportion of one class over another, and these follow Currier's A/B grouping closely. [9] VoynichAttacks (2017) calculates the unigram distribution of each folio, projects the distribution onto three dimensions using t-SNE, and then colors the vectors based on Currier's identifications. [10] Similarly, Zandbergen (2018) determines the bigram distributions for each folio, projects the vectors down using PCA, and colors them the same way. [8]

In all of these cases, the authors attempt to validate their statistical methods by showing that they align closely with Currier's identifications. However, a more robust form of validation would be to first perform the analysis on constructed documents which are in two languages, and then apply the test to the VMS. We thus look to existing approaches on natural language identification and segmentation, and then adapt them to perform on the VMS.

3 Related Work

The most relevant previous work attempts to solve the related but distinct task of identifying languages present within a document. Dunning (1994) states the fundamental observation which enables these methods to perform: "language understanding is not required for language identification." [1] From this starting point, a number of different approaches have been developed to identify languages and segment words based on which language they are in.

Dunning (1994) uses a series of n-gram Markov model to determine the probability of a letter given the sequence of letters previously seen. The models are first trained on a known corpus, and then applied to a document in an unknown language. It then determines the probability of a string l appearing in the document given that the document is in some language A. Finally, it classifies the word based on a which language maximizes the likelihood of the string's appearance. It is a robust process, which performs well on both short (20 bytes) and long (500 bytes) of text, with accuracy above 90%. For very large training sets ($>$ 50K bytes) on long strings (100 bytes), the accuracy is almost always above 99%.

Souter (1994) expands on Dunning's work, and shows that between unique character string identification, frequent word recognition, and bigram/trigram based recognition, trigraphs achieve the highest test accuracy in identifying a text when trained on 100K byte corpuses. In addition, they show that a 100%

accuracy rate can be achieved when the bigram model is tested on samples of 200 characters or more, and when the trigram model is tested on samples of 175 characters are more. Furthermore, they show that a bigram model needs to learn the frequencies for 75% of the possible bigrams in order to perform optimally, while trigram models only need to learn 25-50% of the possible trigrams. When either model learns the frequency of rare n-grams, the accuracy decreases. [11]

Perhaps the most relevant paper is Rehurek and Kolkus (2009), which develops an TF-IDF based model to identifying and segmenting multilingual documents on the web. [2] The model creates a 'graded' metric for every word in a document, called 'relevance.' Positive relevance corresponds to the word belonging to the language, a zero relevance implies no correlation, and a negative relevance corresponds to the word belonging to a language which is not part of the family of languages which have calculated models. While this model is well suited for identifying when languages are not accounted for, it has a high overhead, as building a training model for languages requires large corpula.

4 Methods

The method we present is an hodge-podge of the previous methods discussed, but applied in an unsupervised way, designed to require no training data (since the VMS has no comparable manuscripts).

Given a document D, we randomly extract k samples of n-length substrings. For each substring, we calculate the bigram frequency. This bigram matrix is also the posterior distribution of a first order Markov Model, thus we treat it as a posterior probability distribution $\hat{P_k}$, which is an estimate of the true bigram distribution P_A, where A is a language. Since these are estimated over substrings that will not contain every possible bigram, we add a small smoothing value ($s = .001$) for all bigrams, thereby assigning a non-zero probability to unseen bigrams.

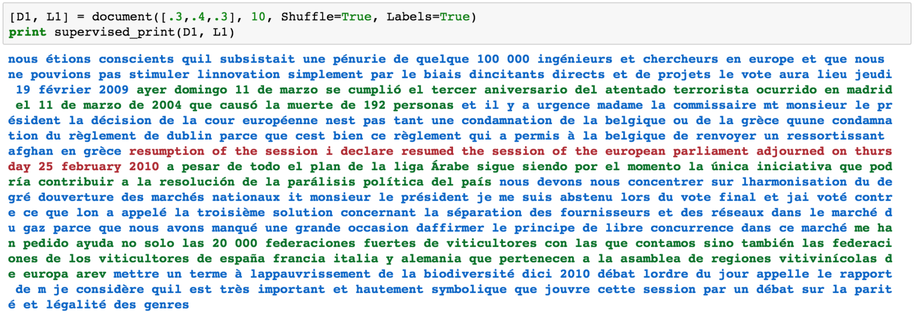
We assume that if two substrings are from the same language, then their bigram distributions should resemble each other. From this assumption, we use the symmetric KL Divergence semimetric H for probability distributions, which measures the difference between two distributions. H is given by the formula

$$H(P_A, P_B) = \sum_{x \in X} P_A(x) \log_2 \frac{P_A(x)}{P_B(x)} + \sum_{x \in X} P_B(x) \log_2 \frac{P_B(x)}{P_A(x)},$$

where X is the set of possible bigrams. Our estimate of H is $\hat{H}$, which is given by

$$\hat{H}(\hat{P_i}, \hat{P_j}) = \sum_{x \in X} \hat{P_i}(x) \log_2 \frac{\hat{P_i}(x)}{\hat{P_j}(x)} + \sum_{x \in X} \hat{P_j}(x) \log_2 \frac{\hat{P_j}(x)}{\hat{P_i}(x)}.$$

We also assume that since $H \geq 0$ for any pair of distributions, that the non-zero difference between two bigram distributions from the same language is due to the natural variance of the language, and is minimal compared to any



```
[D1, L1] = document([.3,.4,.3], 10, Shuffle=True, Labels=True)
print supervised_print(D1, L1)
```

```
nous étions conscients quil subsistait une pénurie de quelque 100 000 ingénieurs et chercheurs en europe et que nous
ne pouvions pas stimuler linnovation simplement par le biais dincitants directs et de projets le vote aura lieu jeudi
 19 février 2009 ayer domingo 11 de marzo se cumplió el tercer aniversario del atentado terrorista ocurrido en madrid
 el 11 de marzo de 2004 que causó la muerte de 192 personas et il y a urgence madame la commissaire mt monsieur le pr
ésident la décision de la cour européenne nest pas tant une condamnation de la belgique ou de la grèce quune condamna
tion du règlement de dublin parce que cest bien ce règlement qui a permis à la belgique de renvoyer un ressortissant
afghan en grèce resumption of the session i declare resumed the session of the european parliament adjourned on thurs
day 25 february 2010 a pesar de todo el plan de la liga Árabe sigue siendo por el momento la única iniciativa que pod
ría contribuir a la resolución de la parálisis política del país nous devons nous concentrer sur lharmonisation du de
gré douverture des marchés nationaux it monsieur le président je me suis abstenu lors du vote final et jai voté contr
e ce que lon a appelé la troisième solution concernant la séparation des fournisseurs et des réseaux dans le marché d
u gaz parce que nous avons manqué une grande occasion daffirmer le principe de libre concurrence dans ce marché me ha
n pedido ayuda no solo las 20 000 federaciones fuertes de viticultores con las que contamos sino también las federaci
ones de los viticultores de españa francia italia y alemania que pertenecen a la asamblea de regiones vitivinícolas d
e europa arev mettre un terme à lappauvrissement de la biodiversité dici 2010 débat lordre du jour appelle le rapport
 de m je considère quil est très important et hautement symbolique que jouvre cette session par un débat sur la parit
é et légalité des genres
```

Figure 1: A multilingual document with shuffled language order.

other language. Formally, this assumption is

$$H(A, A) \leq H(A, B)$$

for any languages A and B.

Finally, we plot the frequencies of all the samples of $\hat{H}$ taken across the document. We expect that if all the substrings are in the same language, then the histogram of $\hat{H}$ will be unimodal. Otherwise, if the substrings are not all in the same language, then we expect to see a multimodal distribution for $\hat{H}$. More specifically, if there are n languages present, we anticipate $\binom{n}{2} + 1$ modes in $\hat{H}$, one for each pair of languages and one for the difference between every language and itself, which we assume will be minimal.

5 Results

The performance of our model was first tested and validated on a specially constructed dataset, then applied to the VMS.

We are able to generate multilingual documents of variable length by drawing sentences in English, Spanish, and French from the Cross Language Dataset. [3] We model each document as a multinomial distribution of languages, and then draw sentences from each language based on those proportions, an approach much like Bag-of-Words models, but in our case a Bag-of-Sentences model. [12] In addition, we are also able to shuffle the order of sentences in our generated documents. If we don't shuffle the sentences, then the all the sentences in one language will appear consecutively; otherwise, they will alternate randomly, like in Figure 1.

We test the method on five different corpula containing five different kinds of documents: trilingual shuffled and unshuffled, bilingual shuffled and unshuffled, and monolingual (shuffling monolingual documents doesn't change the language order). All the documents were 30 sentences long, and each corpula contains 20 documents. Each test initally took $k = 50$ samples of $n = 175$ length substrings

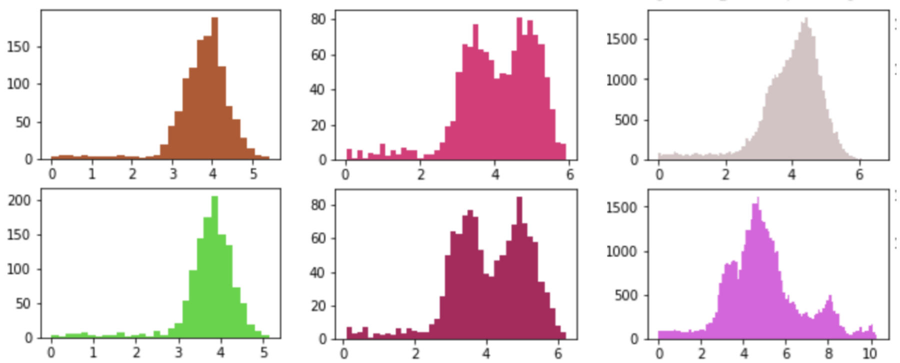


Figure 2: Sample $\bar{H}$ from monolingual (left), bilingual (middle), and trilingual (right) unshuffled documents.

from every document; however, for trilingual documents it became impossible to discern the multimodal distribution without higher sample sizes.

For unshuffled documents, depicted in Figure 2, the method works very well at deriving the multimodal distributions for $\hat{H}$, representing the differences between the bigram distributions for each pair of languages. Furthermore, in the bilingual and trilingual histograms, the first mode is at the same value as the single mode in the monolingual distribution, which reinforces the hypothesis. However, for trilingual documents, the modes are more difficult to discern.

The model performs poorly on shuffled documents, whose histograms are depicted in Figure 3.

Finally, for the VMS, we generated histograms taken from documents of similar length to the VMS, as well as the VMS itself, and are presented in Figure 4.

6 Discussion

We attribute the failure of this test on shuffled documents to the weakness of the test's ability to capture high fidelity bigram distributions in sampled substrings. Note the qualitative differences between substrings sampled from shuffled versus unshuffled trilingual documents in Figures 6 and 7. Ideally, our substrings will capture and preserve the fidelity of bigram distributions within a single language, as this will maximize the differences between two bigram distributions. The substrings from the unshuffled document are all mainly in one language, and will only be in two languages at the point where our document transitions from one language to the other. This property is non-existent in the substrings sampled from shuffled documents; we can therefore expect the fidelity of the bigram distributions to the languages which they contain to be relatively low.

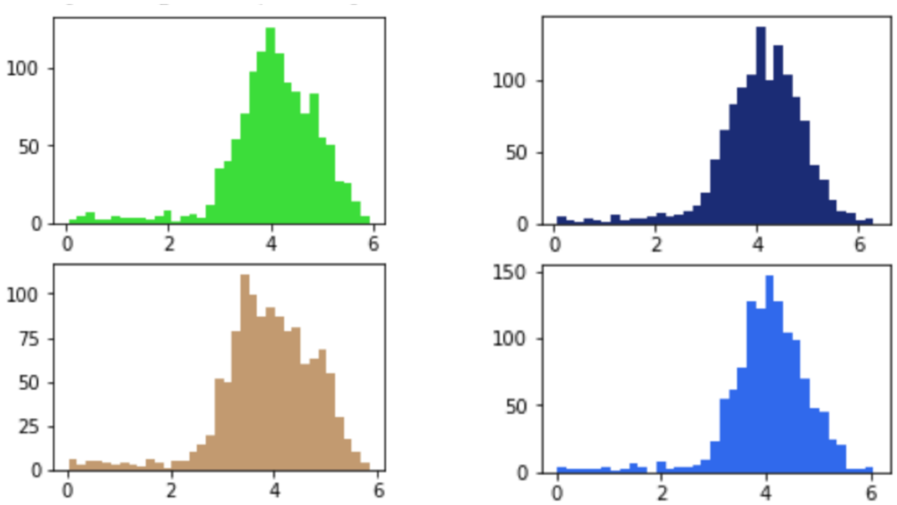


Figure 3: Shuffled histogram samples for bilingual documents (left) and trilingual documents (right).

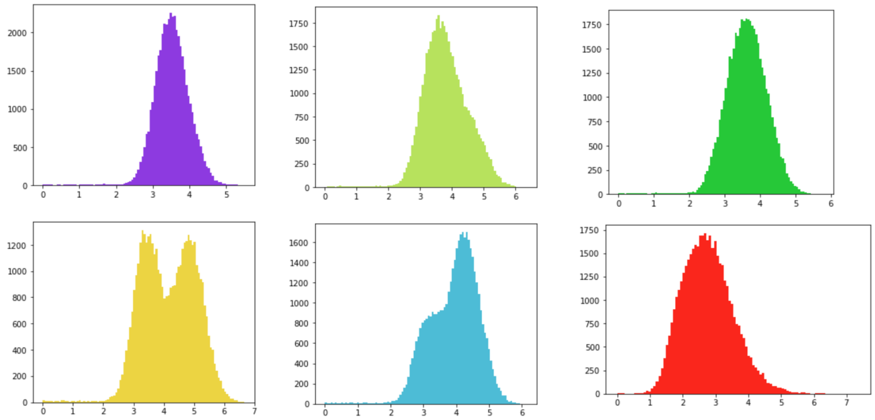


Figure 4: Histograms of similar length from each document type and the VMS. From top-left clockwise: monolingual, bilingual shuffled, trilingual shuffled, VMS, trilingual unshuffled, bilingual unshuffled.

For the VMS, we feel confident in our ability to rule out that it is a perfectly 'sorted' document. If it is in multiple languages, then those sentences are sufficiently shuffled as to evade detection. Alternatively, it may be in multiple dialects, which would account for the right skew but overall unimodality of the distribution. However, we do not feel confident in our ability to rule out anything beyond sortedness.

6.1 Hyperparameters and Stochasticity

While the method is not depedent on a training corpus, it is heavily reliant on high fidelity substring samples that are representative of a language. The parameters which determine how well the modes can be derived are the number of samples taken from the document k, and the length of these samples n.

At first glance, if we increase the number of samples, then we consequently decrease the variance of the estimate for P_i, and thus make the Gaussian distributions representing the difference of distributions in $\hat{H}$ sharper. However, if the length of our samples is too long relative to the size of the document, then by increasing the number of samples taken we risk inflating the frequency of some bigrams by overcounting them repeatedly. More specifically, since the algorithm samples characters appearing towards the center of the document more than characters appearing on the ends, we can imagine that as the limit of the number of samples approaches infinity, the bigrams that appear on the ends of document effectively disappear, and the bigrams appearing in the center of the dominate. This would obviously inhibit our ability to determine high fidelity sample bigram distributions from a document.

If we increase the length of the substrings we sample, then for monolingual documents our results would improve, since the sample variance would decrease. However, on a multilingual document, by increasing the length of the samples we would increase the probability of our substring being in more than one language, which would lead to a mixed bigram distribution of low fidelity to either language.

On the converse, if we decrease the length of the substrings we sample, then for multilingual documents we lower the probability of our substring being in two languages, but at the cost of increasing the variance within the language's bigram distribution. This balance between short and long substrings can be expressed formally as

$$\sigma_{TOT} \sim \sigma_W + \sigma_B$$

and strongly resembles the sums of squares identity in the analysis of variance for a one-way layout:

$$SS_{TOT} = SS_W + SS_B,$$

where SS_{TOT} is the sum of squares for an entire model, SS_W is the sum of squares within each group, and SS_B is the sum of squares between each group. [13]

We also note that unlike the previous methods, our model is stochastic, since it randomly samples substrings from the document. This was intended

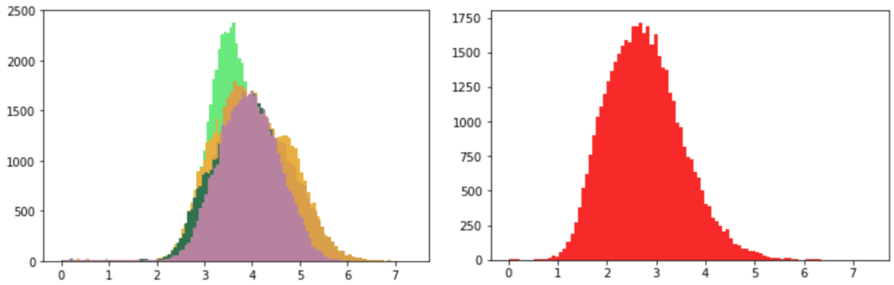


Figure 5: Histograms from generated documents (Latin alphabet $l \approx 45$) on the left; VMS (alphabet $l = 22$ on the right.

to reduce bias in the model, but a more deterministic approach could improve performance.

6.2 Spelling Variation

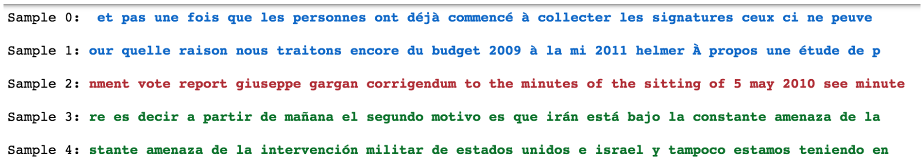
Would this approach be able to identify spelling variations, such as British versus American spelling conventions, as opposed to completely distinct languages, given that the sentences belonging to each dialect were unshuffled? Without testing, we believe that posterior bigram frequencies would capture the significant differences between the dialects, like 'ou' and 'ur' having a high appearance in words like 'favourite' and 'colour' in British but not in English, but would not be significant enough to sufficiently increase the distribution differences over the entire space of bigrams. This would then go undetected in the histogram plotting, still yielding a single mode.

6.3 Indirect Comparisons

The effect of alphabet size may induce unwarranted shrinkage on the histograms generated from the method. If one language has a smaller alphabet size than another, then the expected value of the psKLD would shift left, since there are less degrees of freedom to create divergence. As a result, languages with smaller alphabet sizes will appear to have less differences, which would make identifying modes in $\hat{H}$ more difficult. We see this consequence in the VMS histogram (alphabet size $l = 22$), whose left skew rises at 1, whereas the histograms for all of the generated documents ($l \approx 45$) rises at 2 in Figure 5.

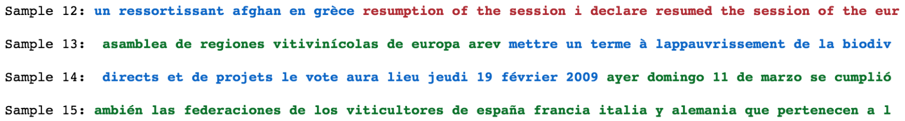
7 Future Directions

We feel that we took a number of shortcuts in order to be able to determine preliminary results within a semester, and believe that there are a number



```
Sample 0:  et pas une fois que les personnes ont déjà commencé à collecter les signatures ceux ci ne peuve
Sample 1: our quelle raison nous traitons encore du budget 2009 à la mi 2011 helmer À propos une étude de p
Sample 2: nment vote report giuseppe gargan corrigendum to the minutes of the sitting of 5 may 2010 see minute
Sample 3: re es decir a partir de mañana el segundo motivo es que irán está bajo la constante amenaza de la
Sample 4: stante amenaza de la intervención militar de estados unidos e israel y tampoco estamos teniendo en
```

Figure 6: Sampled substrings from an unshuffled document.



```
Sample 12: un ressortissant afghan en grèce resumption of the session i declare resumed the session of the eur
Sample 13:  asamblea de regiones vitivinícolas de europa arev mettre un terme à lappauvrissement de la biodiv
Sample 14:  directs et de projets le vote aura lieu jeudi 19 février 2009 ayer domingo 11 de marzo se cumplió
Sample 15: ambién las federaciones de los viticultores de españa francia italia y alemania que pertenecen a l
```

Figure 7: Sampled substrings from a shuffled document.

of different avenues that would lead to fruitful results, many of which were mentioned in previous work in the field.

7.1 Higher Order n-grams

Dunning (1994) uses a model which computes over all n-gram models up to a specified n, and then calculates word likelihood based on a collaborative process across all n-gram windows. In addition, Souter (1994) finds that using trigrams performs optimally over bigrams, requires shorter substrings to accurately classify words, and requires a smaller fraction of the possible trigrams in ordre to classify with high accuracy. We thus think the best direction to go in would be to adapt the model for higher order n-gram models.

7.2 Cross Validation

The sensitive interplay between the length of substrings and the number of substrings we sample from each document can be addressed via cross-validating over the parameter space, which for n is bounded by the length of the document, and for k can be bounded by the overcounting rate. This method could be used to maximize the difference between distributions, and thus lead to larger separations between modes. in $\hat{H}$.

7.3 Model Convergence

In Souter (1994), the question of how much training data is required for a model to converge is explored. If we are able to create a general heuristic for the minimal size of a training corpus required to converge to within a certain threshold of a true n-gram distribution for a language, then we can limit the length of the sample substrings taken, thus lowering σ_B without sacrificing σ_W.

7.4 Other Language Families

We are not sure whether these tests will work on other alphabetic language families, such as Arabic, Hebrew, or Japanese Hiragana. Considering that Reddy and Knight (2011) point to the VMS as being an abjad script, it is worthwhile pursuing this direction further.

7.5 Gaussian Mixture Modeling

We assume that variation within and between language bigram distribution, and thus the differences between them, are distributed normally. While this is a tenuous assumption, especially since psKLD is not a true metric but instead a semimetric, it may still be a potentially interesting approach to model $\hat{H}$ as a Gaussian mixture with $\binom{n}{2} + 1$ Gaussians, and assess the parameters and fit. In addition, if we are able to extract a general set of Gaussian distributions for known pairs of languages, then we could in principle be able to isolate these Gaussians within documents that contain more languages, then remove them to amplify the Gaussians that belong to pairs which contain unknown languages.

7.6 Segmentation

Although this may be premature, since the bigram distributions extracted from the documents are probability distributions, we could compute the mean bigram distribution $\bar{P}_i$ for the ith language, and then compute the maximum likelihood for each word over the space of all the languages, following the method of Dunning (1994). This would work alongside a Hidden Markov Model approach like in Reddy and Knight (2011).

8 Conclusion

Here, we presented a method for determining the number of languages present in a document with no prior assumptions as to which or how many languages might be present, as well as no need for training data. Given these constraints, the preliminary results are strong for certain kinds of documents, while weak for others. We then apply the method on the Voynich Manuscript, for which the model was designed; the results are inconclusive (as always), but seem to limit the possibility of it being in multiple languages in separate sections. Finally, we note that this method is more of a proof-of-concept, and can be made more robust and useful by adding standard statistical tools, and by incorporating results from similar previous work.

9 Special Thanks

The author would like to thank Bob Frank of the Yale Linguistics Department for extremely helpful suggestions and advice, tolerating the author's lack of experience in NLP, and entertaining his curiosity. In addition, the author gratefully acknowledges Claire Bowern for an incredibly well-curated semester filled with fantastic lectures, and where the author could fulill the lifelong dream of studying a book written by aliens.

References

[1] Ted Dunning. *Statistical identification of language.* Computing Research Laboratory, New Mexico State University, 1994.

[2] Radim Řehůřek and Milan Kolkus. Language identification on the web: Extending the dictionary method. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 357–368. Springer, 2009.

[3] Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, and Didier Schwab. A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož, Slovenia, May 2016.

[4] Marcelo A Montemurro and Damián H Zanette. Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis. *PloS one*, 8(6):e66344, 2013.

[5] J Michael Herrmann. The voynich manuscript is written in natural language: The pahlavi hypothesis. *arXiv preprint arXiv:1709.01634*, 2017.

[6] Arthur O Tucker and Jules Janick. Plants as the rosetta stone of the voynich codex©. `https://hort.purdue.edu/newcrop/voynich/unraveling-voynich-chap-5-rosetta-stone.pdf`.

[7] Prescott H. Currier. Papers on the voynich manuscript. `http://www.voynich.nu/extra/img/curr_main.pdf`, 1976.

[8] René Zandbergen. Voynich.nu. `www.voynich.nu`, 2018.

[9] Sravana Reddy and Kevin Knight. What we know about the voynich manuscript. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 78–86. Association for Computational Linguistics, 2011.

[10] VoynichAttacks. Using t-distributed stochastic neighbor embedding (tsne) to cluster folios. `voynichattacks.wordpress.com/author/jjbunn/`, 2017. Computational Attacks on the Voynich Manuscript.

[11] Clive Souter et al. Natural language identification using corpus-based models. *HERMES-Journal of Language and Communication in Business*, 7(13):183–203, 1994.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[13] John Rice. *Mathematical statistics and data analysis*. Nelson Education, 2006.