Computational Paleography of Medieval Hebrew Scripts*

Abstract

This short paper presents ongoing work on ERC Synergy XXX international multidisciplinary project on computational analysis of medieval manuscripts. We focus on clustering Ashkenazi square script using a dataset of 206 pages from 59 manuscripts. Collaborating with expert paleographers, we identified ten critical features and trained a multi-label CNN, achieving high accuracy in feature prediction. We assume that it is possible to computationally predict the sub-clusters known to paleographers and those that are yet to be discovered. Using PCA and χ^2 feature selection, we identified visible clusters. Moving forward, we aim to enhance feature extraction with deep learning algorithms and provide computational tools to ease paleographers' work, enabling new methodologies for analyzing Hebrew scripts and refining our understanding of medieval Hebrew manuscripts.

Keywords

Medieval Hebrew manuscripts, computational paleography, convolutional neural networks, image clustering, recurrent neural networks

1. Introduction

We are engaged in an international effort that aims to develop a revolutionary approach to manuscript studies by combining traditional, digital, and computational paleographic methods to refine and potentially rewrite our understanding of Hebrew scripts, particularly their geographical variation in scribal practices. Using computational tools, we will analyze these manuscripts from paleographic, codicological, linguistic, and literary perspectives. This approach aims to help us understand crucial issues related to the manuscripts' materiality, textuality, transmission, and the historical and intellectual context of their creation and readership. By combining traditional philology with machine learning, computer vision, and computational linguistics, we will process large amounts of textual and paleographical data that traditional philology cannot handle. See Figure 1.

The principal aims include:

- Develop OCR (optical character recognition) systems to convert manuscript images into searchable text.
- Implement text mining algorithms to compare a large corpus of texts and identify quotations, paraphrases, borrowings, allusions, and other intertextual relationships.
- Train machine learning models to perform handwriting analysis and predict the geographical and temporal origins of each manuscript.
- Design natural language processing (NLP) algorithms to extract and analyze linguistic features for improved textual searches and historical context placement.

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

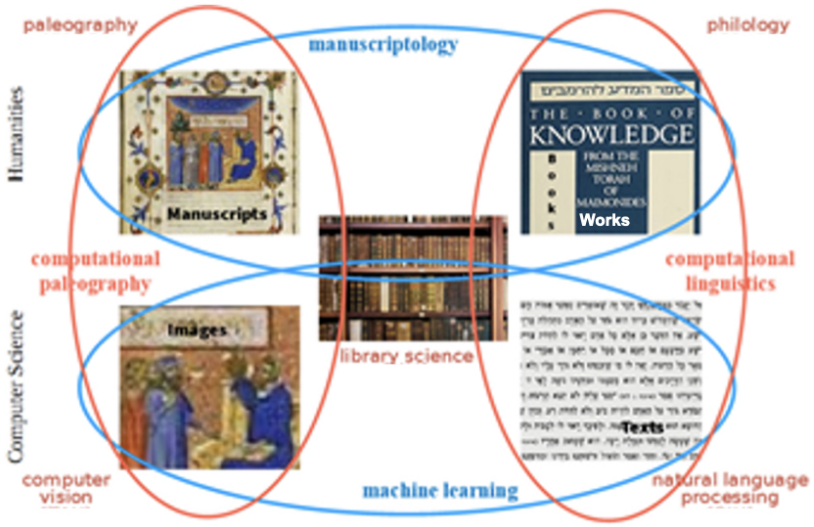

Figure 1: Synergy in computational manuscriptology.

- Integrate traditional and computational methodologies for paleographic, philological, and textual analysis.

2. Computational Paleography Tasks

Accessing the textual and non-textual information of manuscripts is valuable only if we can understand the texts in their specific context of place and time. Among all medieval Hebrew manuscripts, only about 3,500 are dated, featuring colophons (scribes' notes) or other identifying marks. The primary methods to establish manuscript provenance are paleography (the study of handwriting) and codicology (the study of the physical aspects of books). The only existing database of dated medieval Hebrew manuscripts, SfarData (https://sfardata.nli.org.il), focuses primarily on codicology.

However, for a project studying document images, reliance on paleography is essential. Paleographers on our team have worked on developing more precise regional and chronological classifications by analyzing extensive, well-defined manuscript samples and examining the correlations between their local textual features and scripts. We make use of HebrewPal (https://www.hebrewpalaeography.com), an ongoing effort to build a comprehensive database of Hebrew paleography. Processing this data involves synergetic collaboration between traditional paleographers and computer scientists.

As the computational paleography team, we are currently working on solving the problem of finding subgroups among the Ashkenazi square script documents. Medieval Hebrew manuscripts are described by paleographers according to their script mode (square, cursive, and semi-cursive) and their geographical type. The six geographical types are Oriental (Egypt, Palestine, Syria, Lebanon, Iraq, Iran, Uzbekistan, and Bukhara, Eastern Turkey), Sephardic (the Iberian Peninsula, Provence and Languedoc, North Africa, and Sicily), Italian, Ashkenazi (France and England, the Holy Roman Empire, Central and Eastern Europe), Byzantine (Greece, the Balkans, Western Asia Minor, and regions surrounding the Black Sea), and Yemenite (Figure 2). This level of codicological classification for Hebrew manuscripts has been done successfully by computational means [1, 2, 3].

Figure 2: Medieval Hebrew script types in square mode.

Within certain script type-modes, there are distinct subclusters. In rare cases only have these subclusters been relatively well-studied. This is, for example, the case of the Ashkenazi square script, which has been well-studied and clustered [4, 5, 6, 7]. However, even the most experienced paleographer is most familiar with the manuscripts they work with frequently, and no human memory can retain thousands of examples of a script. Moreover, the variations within some script type-modes are very subtle. Therefore we are working to develop computational methods to identify, through computer sciences, clusters, and sub-clusters within different script types yet to be discovered (or those not discoverable) by paleographers.

3. Data

The Ktiv project at the National Library of Israel has led a significant digitization campaign of Hebrew-character manuscripts from collections worldwide. It has accumulated more than 80% of the extant manuscripts, making tens of thousands of manuscripts accessible via a unified catalog. The Friedberg Genizah Project has added images and metadata of approximately 350,000 fragments from medieval book and document depositories, known as "genizot". This digital corpus serves as the source material for our project. For the clustering task, we used high-resolution pages from well-preserved manuscripts.

We have prepared a dataset for the Ashkenazi square clustering problem, called the "ASC dataset." The dataset is publicly available online at [URL Placeholder]. It contains 206 images, each depicting part of a page from 59 manuscripts, with approximately four pages from each manuscript (Table 1). Additionally, it includes an annotation file for the bounding boxes of the main text regions and text lines. The samples are unlabeled, but it is known that 17 manuscripts

Figure 3: Sample page images from the ASC dataset.

are from Germany and 11 are from France, while the origins of the remaining 31 manuscripts are unknown (Figure 3). All the manuscripts are written in Ashkenazi Square script, and we aim to discover potential subclusters within these manuscripts based on their script types, which have slight variations.

Table 1
Statistics of the ASC Dataset

	Germany	France	Unknown	Total
Manuscripts	17	11	31	59
Pages	62	35	109	206
Text regions	136	61	260	457
Text lines	4413	1799	8080	14292

4. Methods and Results

Our preliminary work focused on clustering medieval manuscripts written in Ashkenazi square script using the ASC dataset. Conventional computational methods, such as the bag-of-words approach, struggle to identify the intricate features necessary for effective paleographic clustering, as the frequency of occurrence of paleographical features varies even within the same script type. To address this, we had expert paleographers identify ten critical features that they use in their analyses of this script type. The ten features identified in this way are vocalization marks, left (end of line) justification, vertical stretch, strings, short descenders, fishtails, left slant, biting, nesting, and shading (Figure 4).

4.1. Predicting Paleographical Features

We trained a multi-label VGG-19 network [8] model to predict the presence of these features on a given page image. Treating this as a multi-label problem allowed us to account for the coexistence of multiple labels, as their spatial location and frequency of occurrence are not crucial for paleographical definition.

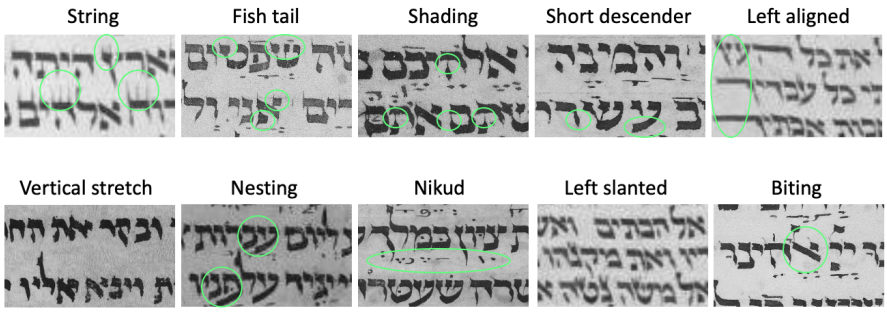

Figure 4: Figure showing all ten features identified in the dataset. Each image patch contains a specific feature, highlighted by green circles.

To prevent overfitting, we employed the regularization technique of early stopping. This method halts training once performance on the validation set ceases to improve, thereby preventing the model from memorizing the training data. Consequently, we stopped the training when the validation loss reached 0.12, achieving the model with the best validation performance (Figure 5).

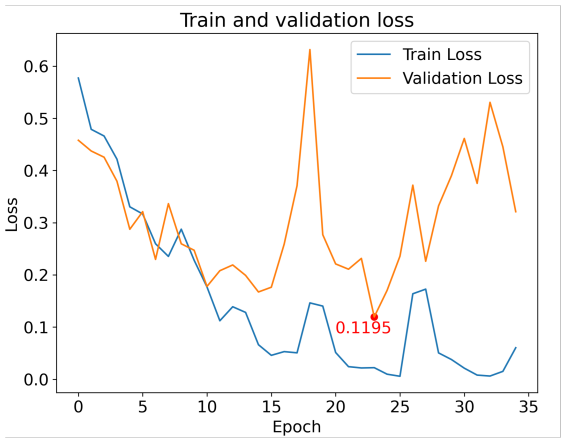

Figure 5: Training and validation loss across epochs, demonstrating the application of early stopping to achieve the model with the best validation performance when the validation loss reached 0.12.

To evaluate model performance on out-of-vocabulary data, we split the dataset at the manuscript level (Figure 6). Out-of-vocabulary testing involves evaluating the model on data that contains features or patterns not seen during training. This is crucial for ensuring that the model can generalize well to new, unseen data, simulating real-world scenarios where new manuscripts are encountered. Splitting at the manuscript level ensures that entire manuscripts, rather than individual pages, were held out as a test set. This method provides a more robust assessment of the model's generalization capabilities by ensuring the evaluation is based on completely unseen data.

The prediction performance on the out-of-vocabulary test set, as shown in the bar graph, demonstrates that our model can effectively automate the tasks performed by a paleographer,

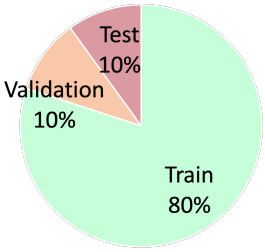

Figure 6: Pie chart showing the split percentages of the dataset at the manuscript level, used for out-of-vocabulary testing to ensure the model's generalization capabilities on unseen data.

achieving accuracy levels of 98%. The performance graph (Figure 7) shows three types of F1 scores: macro average, micro average, and weighted average. The macro average F1 score calculates the F1 score for each class individually and then takes the average, giving equal importance to all classes, regardless of their size. This means every class contributes equally to the final score. The micro average F1 score combines the contributions of all classes to calculate the F1 score, treating every individual prediction equally. This approach is more influenced by the classes with a larger number of samples. The weighted average F1 score calculates the F1 score for each class and then takes the average, weighted by the number of samples in each class. This gives a balanced view by considering the size of each class, ensuring that larger classes have more influence on the final score.

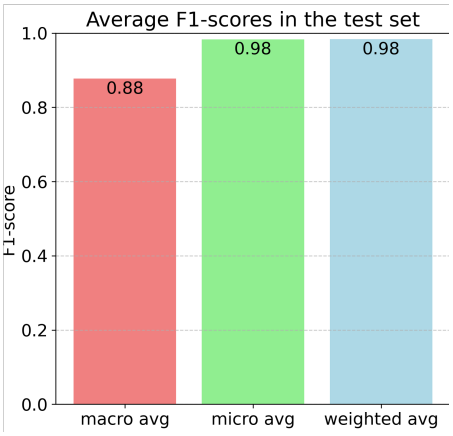

Figure 7: Bar chart showing the average F1 scores for the prediction performance on the out-of-vocabulary test set, demonstrating the model's effectiveness in automating a paleographer task with an accuracy level of 98%

During training, we monitored the prediction accuracy for each of the ten labels to identify the ease or difficulty of learning specific features (Figure 8). For example, the "left slanted" feature took longer to learn but eventually reached high accuracy, indicating its non-binary nature and frequent occurrence in each letter. Conversely, features such as "nesting," "shading," and "string" features took longer to learn and resulted in lower accuracies due to their gradual

values and less frequent appearances.

Figure 8: Feature-wise training F1 scores through epochs, showing the learning progress for each of the ten labels. The "left slanted" feature, despite taking longer to learn, eventually achieved high accuracy, while features such as "nesting," "shading," and "string" features exhibited lower accuracies due to their gradual values and less frequent occurrences.

4.2. Exploring Subclusters

To identify the subclusters, we performed a brute-force search to find the feature combinations that lead to the most cohesive subclusters. Principal Component Analysis (PCA) was used to visualize the samples in 2D and identify potential clusters (Figure 9). By systematically testing all features or selected features, we found that χ^2 feature selection led to visible clusters. This feature selection process highlighted visible clusters based on the selected features (strings, left slanted, vertical stretch, and nesting), addressing the challenge faced by paleographers who can easily identify individual features on a single page but struggle to simultaneously remember and analyze these features across multiple pages to discern grouping patterns. The clustering algorithm mainly successfully grouped manuscripts of known provenance and suggested some meaningful grouping of other manuscripts. The results can be further improved by enlarging the dataset.

5. Conclusion and Future Work

This approach addresses the challenge faced by paleographers who can easily identify individual features on a single page but struggle to simultaneously remember and analyze these features across multiple pages to discern grouping patterns. Thus we end up with well-defined clusters plus insights into the features driving these formations, surpassing the limitations of traditional paleographic analysis.

Looking forward, we aim to explore methods to identify discriminative features that go beyond those defined by paleographers. We assume that a script type S' possesses n distinct paleographical features that are absent in a baseline script type S (Ashkenazi square script,

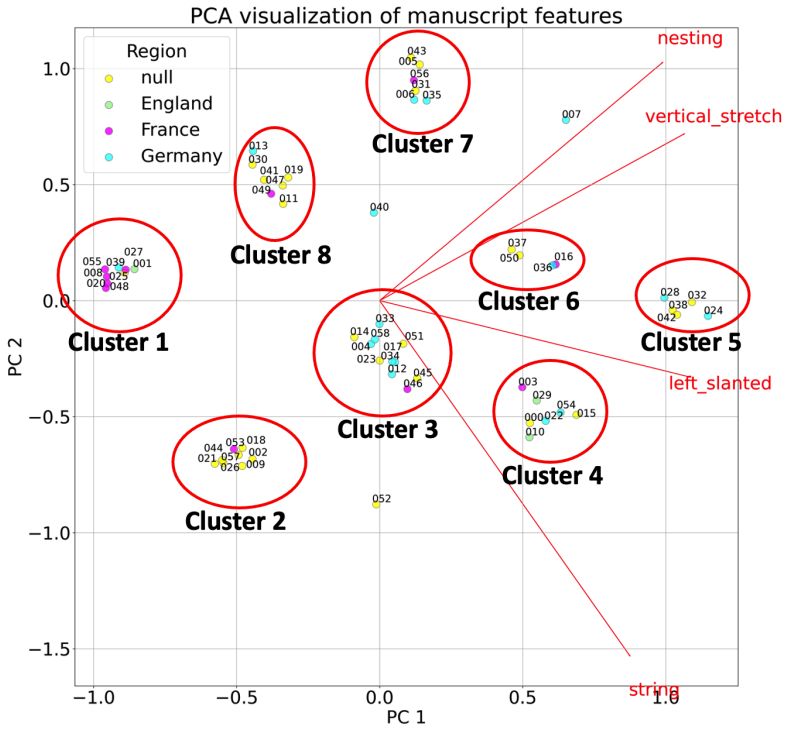

Figure 9: 2D PCA visualization of manuscripts based on χ^2 selected features, highlighting the formation of visible clusters using the identified features (strings, left slanted, vertical stretch, and nesting). Each dot is labeled with the identifier of the corresponding manuscript.

in our case). We train a multi-label CNN to predict the presence of all n features in images of script S', while predicting the absence of these features in images of the baseline S. Using gradient-weighted class activation mapping (Grad-CAM), we visualize the spatial locations of these n features within the images of S'. This approach enables us to identify characteristics that may not be immediately apparent to human experts, furthering our understanding of these script types.

To further enhance the representation of handwriting style features, we will incorporate another deep learning architecture. Specifically, we will train a sequence-generating recurrent neural network (RNN) on the ordered sequence of contour tip points from letter strokes. The hidden state vectors from the RNN will then be used as embedding vectors, which are expected to capture stylistic features of the handwriting.

References

[1] L. Wolf, N. Dershowitz, L. Potikha, T. German, R. Shweka, Y. Choueka, Automatic paleographic exploration of Genizah manuscripts, in: F. Fischer, C. Fritze, G. Vogeler (Eds.), Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the

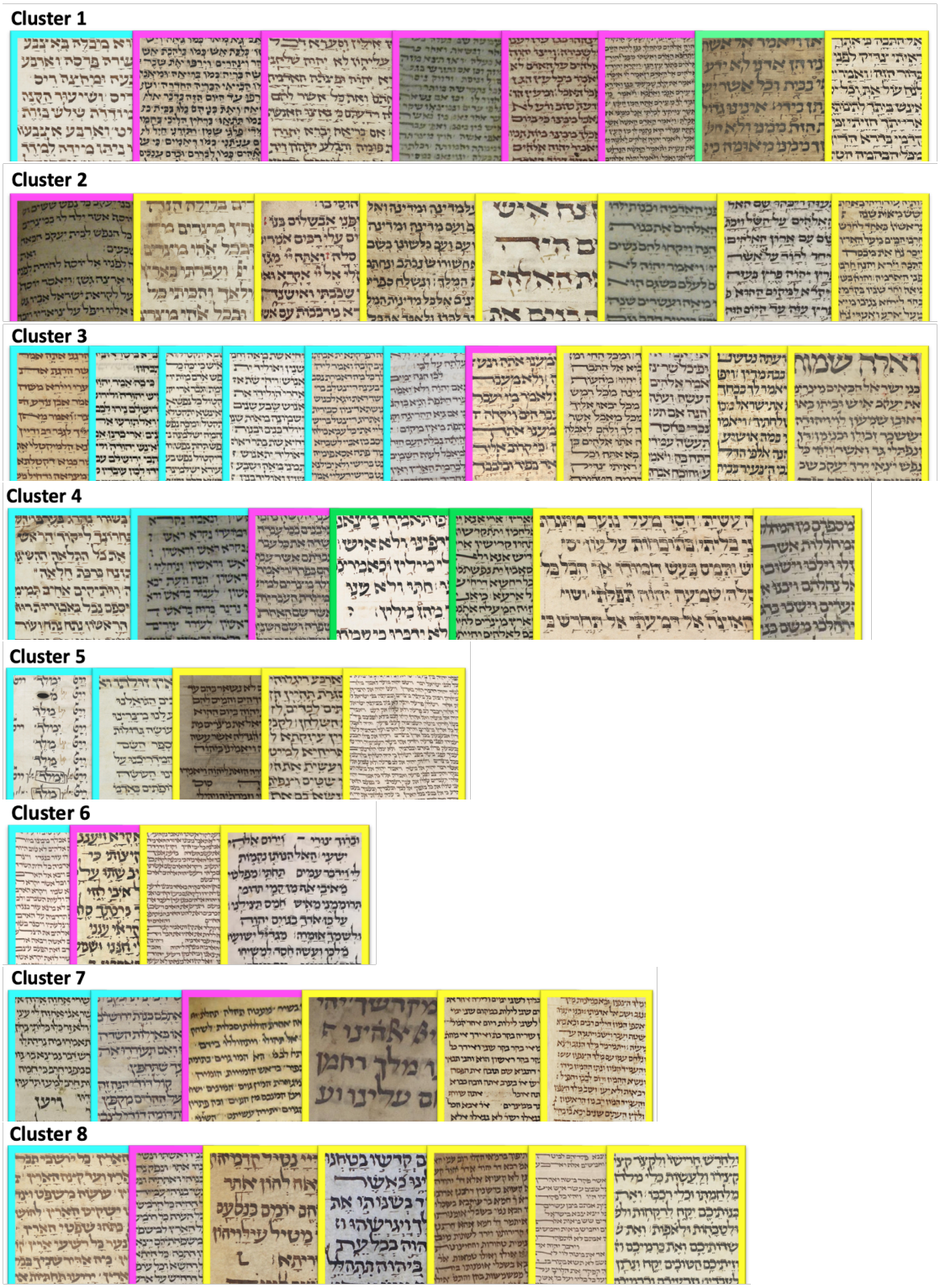

Figure 10: Sample patches from the manuscripts in each of the subclusters. Frames are color-coded: cyan for Germany, magenta for France, green for England, and yellow for unknown.

Digital Age, volume II of *Schriften des Instituts für Dokumentologie und Editorik*, Books on Demand, Norderstedt, Germany, 2011, pp. 157–179.

[2] L. Wolf, L. Potikha, N. Dershowitz, R. Shweka, Y. Choueka, Computerized paleography: Tools for historical manuscripts, in: 18th IEEE International Conference on Image Processing, 2011, pp. 3545–3548. doi:10.1109/ICIP.2011.6116481.

[3] B. Madi, N. Atamni, V. Tsitrinovich, D. Vasyutinsky-Shapira, J. El-Sana1, I. Rabaev, Automated dating of medieval manuscripts with a new dataset, in: Workshop on Computational Paleography (WCP), Athens, Greece, 2024.

[4] M. Beit-Arié, E. Engel, Specimens of mediaeval Hebrew scripts, volume 3, Israel Academy of Sciences and Humanities, 2017.

[5] E. Engel, Calamus or Chisel: On The History of the Ashkenazic Script, volume 28 of *Studies in Jewish History and Culture*, Brill, Leiden, The Netherlands, 2010, pp. 183–197. doi:10.1163/ej.9789004179547.i-398.39.

[6] E. Engel, Between France and Germany: Gothic characteristics in Ashkenazi script, in: N. de Lange, J. Olszowy-Schlanger (Eds.), Manuscrits hébreux et arabes: Mélanges en l'honneur de Colette Sirat, 2014, pp. 197–219.

[7] J. Olszowy-Schlanger, The early developments of Hebrew scripts in north-western Europe, Gazette du livre médiéval 63 (2017) 1–19.

[8] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).