ROADBLOCKS IN GENDER BIAS MEASUREMENT FOR DIACHRONIC CORPORA *

Saied Alshahrani Esma Wali Abdullah Alshamsan Yan Chen
Jeanna Matthews
Department of Computer Science
Clarkson University, Potsdam, NY, USA
{alshahsf,walie,alshamar,cheny3,jnm}@clarkson.edu

ABSTRACT

The use of word embeddings is an important NLP technique for extracting meaningful conclusions from corpora of human text. One important question that has been raised about word embeddings is the degree of gender bias learned from corpora. Bolukbasi et al. [1] proposed an important technique for quantifying gender bias in word embeddings that, at its heart, is lexically based and relies on sets of highly gendered word pairs (e.g., mother/father and madam/sir) and a list of professions words (e.g., doctor and nurse). In this paper, we document problems that arise with this method to quantify gender bias in diachronic corpora. Focusing on Arabic and Chinese corpora, in particular, we document clear changes in profession words used over time and, somewhat surprisingly, even changes in the simpler gendered defining set word pairs. We further document complications in languages such as Arabic, where many words are highly polysemous/homonymous, especially female professions words.

***K*eywords** word embedding · gender bias · NLP · Arabic · Chinese · profession words · diachronic

1 Introduction

Natural Language Processing (NLP) plays a significant role in many powerful applications such as speech recognition, text translation, and autocomplete and is at the heart of many critical automated decision systems making crucial recommendations about our future world. Word embedding systems are widely used to represent text data as vectors

* *Citation:* **Saied Alshahrani, Esma Wali, Abdullah R Alshamsan, Yan Chen, and Jeanna Matthews. 2022. Roadblocks in Gender Bias Measurement for Diachronic Corpora. In Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change 2022. Association for Computational Linguistics.**

and enable NLP computation. Systems such as Word2Vec [2], GloVe [3], and BERT [4] ingest large corpora of human text and can be used to learn semantic and syntactic relationships between words.

At the same time, it has been demonstrated that these systems learn a wide variety of societal biases embedded in human text including racial bias, gender bias, and religious bias [5, 6]. In a widely cited paper, Bolukbasi et al. [1] demonstrated that a system trained with a corpora of Google News would complete the word comparison "man is to computer programmer as woman is to what?" with the response "homemaker" suggesting an alarming level of gender bias when used in tasks such as sorting resumes for computer programming jobs. Chen et al. [7] extended these techniques beyond English to eight other languages (Chinese, Spanish, Arabic, German, French, Farsi, Urdu, and Wolof) and applied them to Wikipedia corpora in each of these languages. They documented persistent gender bias and lack of representation in the modern NLP pipeline.

NLP research often uses large, modern datasets like Google News and Wikipedia. Developers of a wide variety of NLP-based applications begin with large pre-trained models that are also based on large corpora of human text [8]. These pre-trained models also largely reflect the speech/writing of modern English speakers producing digital text. The speech/writing of speakers of the more than 7,000 languages spoken worldwide is often under-represented [9]. Similarly, historical speech/writing is often under-represented despite the fact that historical speech/writing is often considered foundational to cultural identity. Investments in multilingual NLP and processing of diachronic corpora are essential if we want our NLP-based automated decision making systems to more widely reflect foundational cultural norms and identity from around the world.

The inspiration for this paper was to re-examine Bolukbasi et al.'s [1] popular NLP-technique for quantifying gender bias from the perspective of applying it to diachronic corpora in Arabic and Chinese. Specifically, Bolukbasi et al.'s [1] method begins with identifying a set of profession words and a set of highly gendered word pairs (defining set). In this paper, we explore the degree to which these words might change over time. We document ways in which this method is fundamentally fragile for diachronic corpora because of the way these sets of words would change over time.

In Section 2, for background, we elaborate on Bolukbasi et al. [1] and Chen et al.'s [7] multilingual extensions and some other relevant related work. Section 3 describes our experience with two different diachronic Arabic corpora, especially the impact on changes in profession set words over time. In Section 4, we discuss changes in some defining set words in Chinese using the Google Ngram Viewer. We conclude and discuss future work in Section 5.

2 Background and Related Work

Bolukbasi et al. [1] pioneered a method for quantifying the amount of gender bias learned in by word embedding systems and many researchers have built on their techniques including Chen et al. [7] who observed substantial hurdles in extending the techniques beyond English. In this paper, we build on both Bolukbasi et al. [1] and Chen et al.'s [7] work to examine additional hurdles that would arise when attempting to apply these techniques to diachronic corpora.

Bolukbasi et al.'s [1] original method is based on two sets of words. The first set (the defining set) consists of 10 highly gendered word pairs (she-he, daughter-son, her-his, mother-father, woman-man, gal-guy, Mary-John, girl-boy, herself-himself, and female-male) and the second (profession set) consists of 327 profession words such as nurse, teacher, writer, engineer, scientist, manager, driver, banker, musician, artist, and chef. They used the difference between the defining set word pairs to define a gendered vector space and then evaluated the relationship of the profession words relative to this gendered vector space. Ideally, profession words would not reflect a strong gender bias. However, in practice, they often do. According to such a metric, the word doctor might be male biased or the word nurse female biased based on how these words are used in the corpora from which the word embedding model was produced.

Bolukbasi et al. [1] uses these two sets of words to compute a gender bias metric for each word and from there to express the gender bias of a corpora. Specifically, each word is expressed as a vector by Word2Vec and then the center of the vectors for each defining set pair is calculated. For example, to calculate the center of the definitional pair woman/man, they average the vector for "woman" with the vector for "man". Then, they calculate the distance of each word in the definitional pair from the center by subtracting the center from each word in the pair (e.g., "woman" - center). They then apply Principal Component Analysis (PCA) to the matrix of these distances. PCA is an approach that compresses multiple dimensions into fewer dimensions, ideally in a way that the information within the original data is not lost. Usually, the number of reduced dimensions is 1-3 as it allows for easier visualization of a dataset. Bolukbasi et al. [1] used the first eigenvalue from the PCA matrix (i.e. the one that is larger than the rest). Because the defining set pairs were chosen to be highly gendered, they expected this dimension to be related primarily to gender and therefore called it the gender direction or the g direction. Finally, the g direction is a vector, and there is a vector representing each word. Therefore, they used cosine similarity between the vector for each word, w, and the g direction vector as the measure of gender bias for that word. For a corpora or other collection of words, one can average the gender bias of words contained in the corpora as a measure of gender bias in the corpora using the equation of Bolukbasi et al. [1] for the direct gender bias of an embedding:

$$\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |\cos(\vec{w}, g)|^c$$

where N is the given gender neutral words, and c is a parameter that determines the strictness in measuring gender bias.

Chen et al. [7] extended the Bolukbasi et al.'s [1] method to eight languages besides English - Chinese, Spanish, Arabic, German, French, Farsi, Urdu, and Wolof. In order to do so, they first made modifications to the defining set to make it more translatable across the 9 languages. For example, they dropped pairs like she-he, her-his, gal-guy, Mary-John, herself-himself, and femalemale because of problems in translation for some languages and adding pairs like queen-king, wifehusband, and madam-sir. Second, they observed that the Bolukbasi et al.'s [1] method cannot be applied directly to languages such as Spanish, Arabic, German, French, and Urdu that primarily use grammatically gendered nouns (e.g., escritor/escritora in Spanish vs. writer in English). They solved this problem using a weighted average of the number of occurrences of each variant of the professional word (male, female, or neutral) multiplied by the gender bias score for that variant.

In this work, we build on both Bolukbasi et al. [1] and Chen et al. [7] and focus on the unique challenges that arise when applying these techniques to diachronic corpora. Specifically, we examined changes in both the profession set and defining set over time in Arabic and Chinese. Certainly, professions have changed drastically over that amount of time and so a method based on profession set words like Bolukbasi et al.'s method will have substantial challenges. We explored this using corpora including a database of Arabic poems spanning 11 eras from the Pre-Islamic period (before 610) to modern day. While we saw less change over time in the usage of the simpler defining set words than in the profession set words, we did observe some interesting changes in even the defining set words over time, especially in Chinese. In the process of this work, we also documented further complications in languages such as Arabic, where many words are highly polysemous/homonymous, especially female professions words.

Wevers [10] also used word embeddings to examine gender bias over time. They used a collection of Dutch Newspaper articles spanning over four eras (1950-1990), training four embedding models per newspaper, one per era, using the Gensim implementation of Word2Vec to demonstrate how word embeddings can be used to examine historical language change. They observed clear differences in gender bias and changes within and between newspapers over time. Slight shifting of bias was observed in some themes like shifting towards female bias in themes related to sexuality and leisure (mostly seen in newspapers with religious background). Shifting towards male bias in themes related 'money', 'grooming', and negative emotions, especially in newspapers with a liberal background, was also observed.

Rudolph and Blei [11] developed dynamic embeddings building on exponential family embeddings to capture the language evolution or how the meanings of words change over time. They used three datasets of the U.S. Senate speeches from 1858 to 2009, the history of computer science ACM abstracts from 1951 to 2014, and machine learning papers on the ArXiv from 2007 to 2015. They demonstrated how words like Intelligence, Iraq, computer, Bush, data change their meaning over time. They observed that the dynamic embeddings provided a better fit than classical embeddings and captured interesting patterns about how language changes. For example, a word's meaning can change (e.g., computer); its dominant meaning can change (e.g., values); or its related subject matter can change (e.g., Iraq).

Xu et al. [12] demonstrated the characterization of the semantic weights of subword units in the composition of word meanings. They used a subword-incorporated or a word embedding model variant for the evaluation and revealed interesting patterns change in multiple languages. Their training datasets consist of Wikimedia dumps for 6 Languages (up until July 2017) consisting of Chinese and other Indo-European languages like English, French, German, and Italian. The results revealed major differences in the long-term temporal patterns of semantic weights between Chinese and five Indo-European languages. For example, in Chinese, the weights on subword units (characters) show a decreasing trend, i.e., individual characters play less semantic roles in newer words than older ones whereas the opposite trend was observed in other languages. Therefore, Chinese words are treated more as a whole semantic unit "synthetically", while words in Indo-European languages require more attention into the subword units "analytically". These results provide evidence towards word formations to the linguistic theories. For example, the notion of "word" in Chinese is always changing: Modern Chinese has multiple characters as a whole semantic unit opposite to its older counterpart.

Time Periods	Number of Books	Vocab Size	Token Size
Books Before Islam	3	16,460	39,255
Books Before 1900	2,820	2,075,505	566,366,883
Books After 1900	773	1,335,027	136,870,579
Duplicate Books	11	-	-
Unknown Books	2,931	-	-
All Shamela's Books	6,527	2,520,372	703,276,717

Table 1: Measurements of Shamela Library dataset in terms of the number of books, vocabulary size (unique words), and token size (all words) for each time period. We did not train a GloVe model on the unknown books alone or the duplicate books and therefore are not reporting vocab size and token size.

The semantic weight carried by a single character is decreasing over time. This is strong evidence in support of the claim that Chinese has been evolving towards more detailed multisyllabic words from concise and monosyllabic words.

3 Changes in Arabic Over Time

Building on both Bolukbasi et al. [1] and Chen et al. [7], we consider how the sets of profession words required by the Bolukbasi et al.'s method would need to change over time in Arabic. We begin by describing two diachronic datasets that we used and how we processed these datasets, then we describe the changes in the profession word usage over time.

3.1 Datasets and Methodology

In this paper, we use two Arabic datasets: Shamela Library (المكتبة الشاملة) that is released by Shamela Library Foundation [13], and Arabic Poem Comprehensive Dataset (APCD) by Yousef et al. [14]. Shamela Library is a free project that collects thousands of Islamic religious and other related sciences books. APCD is a collection of Arabic poems spanning 11 eras, from the Pre-Islamic (before 610) to the Modern age (1924 - Now). Arabic NLP researchers commonly use these two datasets to study Arabic classics.

We processed the Shamela Library dataset version of 6,538 Arabic books (6,527 unique books after removing duplicates) in Microsoft Word format (1997-2004).[2] The books in this corpora were not labeled according to the publication dates. Thus, to study the language change over time in the Arabic language, we further classified Shamela's Arabic books into three different time periods based either on their publication date or the authors' date of death when publication date was not available. We identified books written before Islam or before 610 (only three books), books written before 1900 (2,820 books), and books written on or after 1900 (773 books). We were not able to identify publication dates or the authors' dates of death of the remaining 2,931 books due to not having any; Table 1 summarizes some key attributes of this dataset.

[2]We contribute the scripts we wrote to process these corpora and overcome several challenges with the data. For example, one challenge we faced was correctly converting back and forth between the Arabic Windows-1256 to the Unicode (UTF-8) encoding schemes. The Arabic books were written in an old version of Microsoft Word (1997-2004), which caused encoding scheme conversion errors, resulting in unreadable characters by native Arabic speakers or even NLP tools. Scripts can be found here: `https://github.com/Clarkson-Accountability-Transparency/gBiasRoadblocks`

Eras	Poetic Verses	Vocab Size	Token Size
Pre-Islamic (before 610)	21,907	60,082	204,450
Islamic (610-661)	2,942	12,388	24,461
Umayyad (661–750)	63,776	119,533	610,563
Between Umayyad and Abbasid	24,077	65,220	221,058
Abbasid (750–1258)	234,494	252,339	2,156,195
Andalusian (756–1269)	111,011	151,503	1,024,653
Fatimid (909–1171)	124,129	172,460	1,171,842
Ayyubid (1174–1252)	112,350	152,165	1,061,503
Mamluk (1250–1517)	164,780	198,748	1,550,669
Ottoman (1517–1924)	159,576	186,795	1,492,132
Modern (1924 - Now)	778,723	462,478	7,146,135
All APCD's eras	1,797,765	736,576	16,663,658

Table 2: Measurements of Arabic Poem Comprehensive Dataset in terms of number of poetic verses, vocabulary size (unique words), and token size (all words) for each era.

We also processed the APCD, an Arabic poetry dataset that is collected mainly from the Poetry Encyclopedia (الموسوعة الشعرية) that is released by Abu Dhabi Department Poetry and Diwan (الديوان) [15]. Unlike Shamela, this dataset was already labeled by era, making it a good choice for studying language change over time. It has, before preprocessing, approximately 1,831,770 poetic verses labeled by their meter, the poet's name, and the era they were written in. One drawback of this corpora is that it is relatively small. Table 2 summarizes some key attributes of this dataset.

We then produced a total of 16 GloVe models [3] from the three time periods of Shamela, the 11 eras of APCD, all Shamela, and all APCD.[3] Each GloVe model is a context-independent model that produces a one-word vector (word embedding) for each word even if that word appears in the context a few times unlike BERT and ELMo [4, 17]. Each GloVe model provides vocabulary size, token size, and word vectors. It is important to note that before training GloVe models, it was necessary to preprocess the two datasets using Linux/Unix command-line utilities like `tr` (for translating or deleting characters), `sed` (for filtering and transforming text), `iconv` (for converting between encoding schemes), and `awk` (for pattern scanning and language processing), along with CAMeL tools [18], an open-source python toolkit for Arabic NLP, to dediacritize the Arabic diacritical marks and remove unnecessary characters.

3.2 Modern and Historical Professions

We began with a consideration of how the profession sets used in Bolukbasi et al. [1] and Chen et al. [7] would need to change over time. First, we identified 50 modern profession words that we expect would simply not exist in the older time periods/eras in Shamela and APCD datasets.[4] For example, the profession of electrician would not have existed before the advent of electricity. Second, we identified 50 historical profession words that we think exist in older time periods/eras in Shamela and APCD datasets but which are much less common in modern times.

[3] Bolukbasi et al. [1] used Word2Vec to generate word embeddings, and in this paper, we chose GloVe instead because GloVe performs better than Word2Vec in the Arabic language [16]. See GloVe Models Properties in Appendix A.

[4] The full list of the used defining set, modern profession set, and historical profession set words and their word counts, vocabulary sizes, and word frequencies for all Shamela's books and all APCD Arabic poems are listed in Appendices B, C, and D.

As in Chen et al. [7], we further categorized each word based on gender. In Arabic, most profession words have a male variant and a female variant in which the spelling is changed slightly based on gender, for example female pilot (طيَّارة) and male pilot (طيَّار). Linguistically, many professions that would be extremely uncommon for men or women do have a male or female version of the word (e.g., it is rare for a woman to have the profession chamberlain/head of staff (حَاجِب), but there is a female word for that profession). However, in some cases, either the male or female version does not even exist linguistically (e.g., there is no male word of midwife (قَابِلة) profession). There are also more rare neutral words, like musician (موسِيقَار), that is used for both genders with no spelling changes.

In the APCD dataset, we found, as expected, that there are some modern professions that occur noticeably only in the modern era of the Arabic poems, but do not appear at all in the previous historical eras, such as the male engineer (مُهَنِدِس) that occurs 17 times, and the neutral profession of an electrician (كَهرَبائِي) that occurs only four times in the modern age, indicating that those modern professions are increasingly appearing in the modern age of the Arabic poems and confirming that Arabic native speakers (i.e., Arabs) still use the poems as an effective way to document the Arabic language changes over time.

On the other side of history, in the Shamela dataset, we found that a few historical professions frequently occur in the time periods before 1900 but not significantly after 1900. Some professions reflect essential shifts in legality. For example, one profession that is fortunately no longer legal or acceptable is male slaver (نَخَّاس). Fortunately, the male slaver profession appears much less often (only 12 times) in the time period after 1900, while it appears unpleasantly 118 times before the 1900 time periods. As another example, male chamberlain/head of staff (حَاجِب) appears 9,518 before the 1900 time periods, but only appears 914 times in the time period after 1900, showing that this male profession/position is on its way to extinction.

3.3 Polysemous/Homonymous Professions

The Arabic language is one of the most morphologically rich languages, with a high level of orthographic ambiguity, causing native speakers to use the optional diacritical marks to differentiate between two words [19].[5]

We noticed in the Shamela Library dataset that a few modern profession words change their connotations over time, and many profession words have alternate meanings due to the Arabic's orthographical ambiguity. We also found that this was especially true of female profession words. For example, the word (مُدَرِّسَة) for female teacher also means a school building (مَدْرَسَة), another word (طَيَّارَة) for a female pilot also means an airplane (طَيَّارَة). In all these cases, this complicates the use of both word counts and word embeddings in tracking the relative uses of profession words over time. One homonymous example is the female trader (مُتَدَاوِلَة) profession. The same word (مُتَدَاوَلَة) also means

[5] In our preprocessing, we removed the optional diacritical marks as is generally recommended for Arabic NLP as a first step to reducing some data sparsity [18]. Unfortunately, removing diacritical marks increases the orthographic ambiguity, but retaining them would lead to a high degree of variance for the same word because the placement of diacritical marks varies with the grammatical placement of the word in a sentence. It is a difficult tradeoff for Arabic NLP that other researchers are attempting to tackle with advanced techniques, such as stemming and lemmatization [20, 21].

[7] English translations of the word clusters are automatically generated using Google Translator API that is included in the deep-translator Python model (`https://deep-translator.readthedocs.io`).

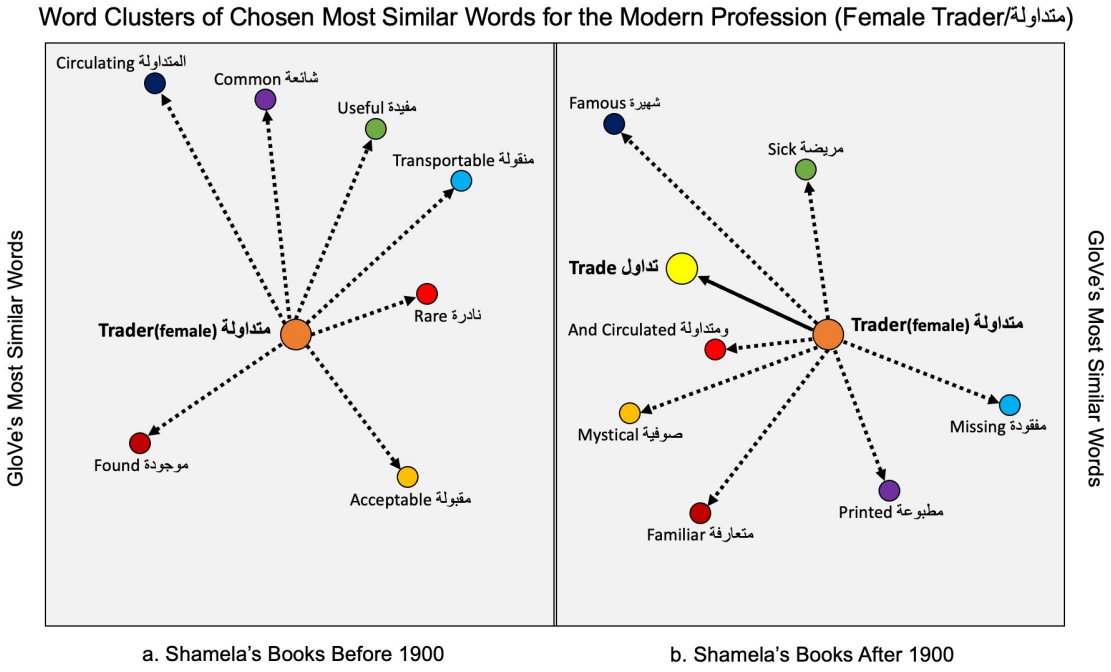

Figure 1: a. A word cluster of chosen GloVe's most similar words of the female profession trader (مُتَدَاوِلَة) in Shamela Library dataset in the time period before 1900, demonstrating that its word cluster is including different words with different meanings due to its homonymy. b. A word cluster of chosen GloVe's most similar words of the female profession trader (مُتَدَاوِلَة) in Shamela Library dataset in the time period after 1900, illustrating that a new related-trading activity word joining the profession word cluster, (trade/تَدَاوُل)[7]

common, famous, familiar, or circulating to describe a current news event. We see this alternate meaning dominate the usage of the word, complicating any attempt to study the prevalence of females engaged in this profession. Interestingly, we see evidence of change over time in the usage of this word. To investigate the semantic meaning of related words to the trading activity, we studied GloVe's most similar words (calculated based on the cosine similarity between two word vectors) for this profession word in two time periods of the Shamela Library dataset: before 1900 and after 1900. As shown in Figure 1a, before 1900, none of most similar words reflect the trading profession word (مُتَدَاوِلَة). However, in Figure 1b, after 1900, we see a word related to trading activity (trade/تَدَاوُل) appear in the most similar words of GloVe model. Thus, the connotation of the female trader (مُتَدَاوِلَة) profession is changing over time to more often reflect the actual profession of female trader (مُتَدَاوِلَة) and not just the alternate meaning of current news events.

3.4 Illegal Professions

In the religion of Islam, some professions are forbidden, for example, all types of usury, and serving, selling, or drinking alcohol. We examined a set of illegal/religiously forbidden profession words in Islam across the 11 ages of the Arabic poems, such as male usurer (مُرَابِي), female usurer (مُرَابِية), male bartender (سَاقِي), and female bartender (سَاقِيَة). Specifically, we closely focused on the diachronic semantic meaning change of the bartending profession words in the

parallel eras of the APCD dataset. Interestingly, we found that bartending profession words in the early ages of the Arabic poems like Pre-Islamic, Islamic, and Umayyad only point to providing water to people but not serving wine even though the wine does exist. Those bartending profession words are polysemous and could carry other meanings like the male bartender (سَاقِي) could have a meaning of the phrase 'my leg' (سَاقِي), while the female bartender (سَاقِيَة) could have as well the meaning of 'a water creek or an aqueduct' (سَاقِيَة).

To thoroughly investigate the occurrence of those profession words regarding their correlation with water – the allowed/halal drink, and the wine — the forbidden/haram drink in Islam, we manually analyzed the Arabic poems of each age and decided whether that word occurrence is a water-related meaning, wine-related meaning, or other unrelated meanings to both of the drinks. Figure 2a shows that the male bartender (سَاقِي) profession word started to appear in the Arabic poems as a profession of serving alcohol generally, wine exclusively, as a symbol of love, passion, and adoration for women from the age of between Umayyad and Abbasid until the Modern age.

One example of that is when the Abbasid Arabic poet, Abu Bakr Al-Sanobi (أبو بكر الصنوبري), said in his famous poem, the Pole of Pleasure in the Descriptions of Wines (قطب السرور في أوصاف الخمور): "O bartender of wine, do not forget us, O Goddess of Oud, spur singing (أيَا سَاقِي الخَمرِ لا تَنسنا – ويا ربَّة العُودِ حُثِّي الغِنَا)." Another example of that in another age, the Ottoman age, is for the Arabic poet, Abdul Ghani Al-Nabulsi (عبد الغني النابلسي), said in this romantic poem, Bartender O Bartender (ساقي يا ساقي): "Bartender O bartender, Give me some of his remaining wine (سَاقِي يَا سَاقِي – اسقِينِي مِن خَمرِهِ البَاقِي).

Similarly, in Figure 2b, the female bartender (سَاقِيَة) started to appear as a profession of serving wine from the age of between Umayyad and Abbasid until the Modern age as same as the male bartender (سَاقِي) profession word, except they did not appear in the two ages of Ayyubid and Ottoman. While the female and male bartender (سَاقِي و سَاقِيَة) surprisingly appeared in correlation with wine in the Arabic poems despite its religious forbiddance, both of the two profession words also refer to water-related words. For example, the female bartender (سَاقِيَة) refers to the 'water creek or aqueduct.' One example to show that is when the Modern Arabic poet, Rashid Ayoub (رشيد أيوب), said in his poem: "I sat in the meadow alone at the water creek, in which the water echoed the sound of my melodies", جَلَستُ فِي الرَّوضِ وَحدِي عِندَ سَاقِيَةً يُرَدِّدُ المَاء فِيهَا صَوت أَلحَانِي.

4 Changes in Chinese Over Time

Although our primary focus in this study has been on Arabic, we found interesting evidence of change over time in Chinese as well. Classical Chinese (before 1900) uses a vocabulary and grammar that differs significantly from modern Chinese. We were surprised to find evidence not just of changes in professions over time, but also changes in defining set words. As we found in the diachronic corpora in Arabic, we expected changes in profession words over hundreds of years, but thought that the more fundamental defining set words like woman/man, girl/boy and madam/sir would not change substantially.

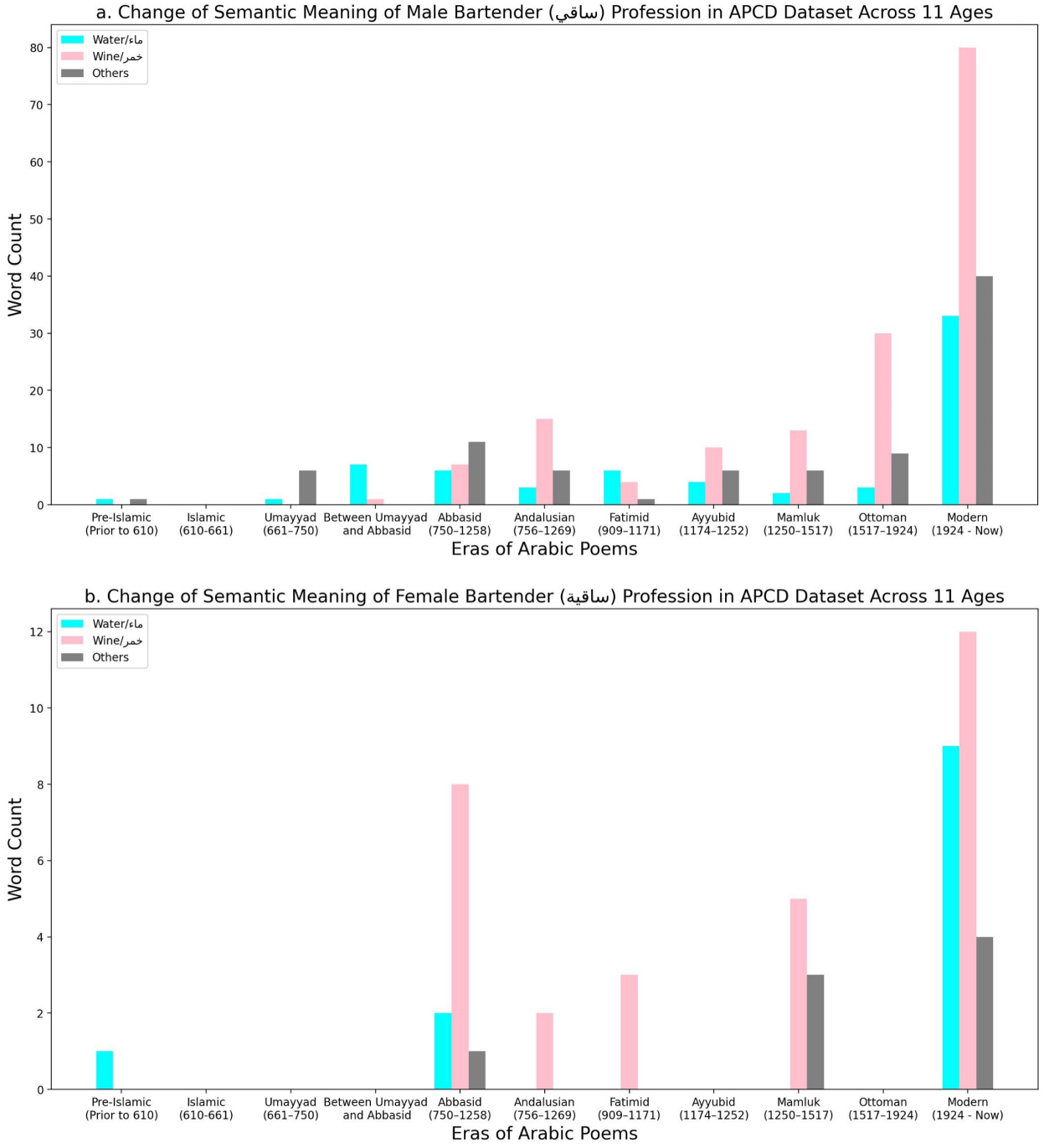

Figure 2: a. A word count of the occurrence of the male bartender (سَاقِي) across the 11 ages of the Arabic poems in the APCD dataset, showing the related meanings of the profession word like serving water, wine, or could be entirely meaning something that entirely unrelated to the profession word's meaning of serving drinks. b. A word count of the occurrence of the female bartender (سَاقِيَة) across the 11 ages of the Arabic poems in the APCD dataset, showing the related meanings like serving water, wine, or could be entirely meaning something that entirely unrelated to the profession word's meaning of serving drinks.

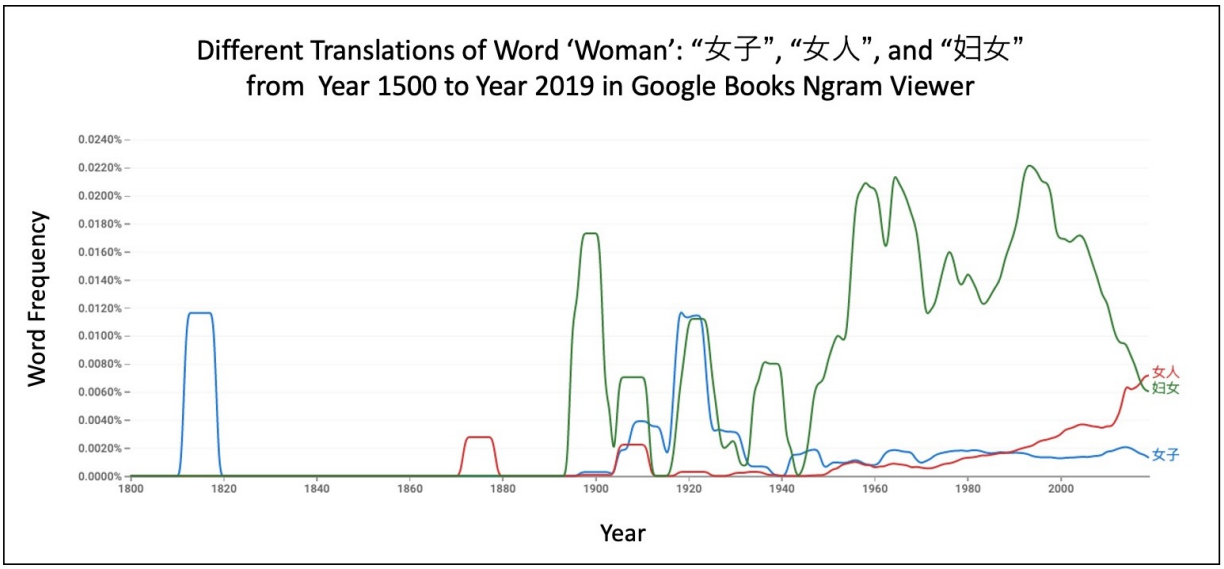

Figure 3: A timeline of word frequencies of different translations of word 'woman': "女子", "女人", and "妇女" that were found in multi-sources printed between 1500 and 2019 using Google Books Ngram Viewer.

In Chinese, the word 'woman' can be translated in many ways, including "女子", "女人", and "妇女". The word "女子" was popularly used in ancient times, but its usage has decreased in modern writing. In Figure 3, we used Google Books Ngram Viewer to chart the word frequencies of the different translations of the word 'woman': "女子", "女人", and "妇女" found in sources printed between 1500 and 2019 in Google's Books corpora in English, Chinese, French, German, Hebrew, Italian, Russian, or Spanish [22]. This shows us that as languages evolve over time, defining sets, like profession sets, may also have to evolve to measure gender bias using methods like the Bolukbasi et al.'s method [1].

Besides using Google Books Ngram Viewer, we also assembled a small collection of works that might be considered "classics" in Chinese spanning the period 475 BC - 1992, for example 司马迁 (Records of the Grand Historian) by Qian Sima, 萧红 (Tales of Hulan River) by Hong Xiao, and 论语 (The Analects). We found that roughly half of the profession words used by Chen et al. [7] did not appear, and that also two of the defining set words "boy" and "madam" used did not appear. Interestingly, Google Books Ngram Viewer showed that the word 'madam' was used very frequently between 1905 and 1910, but our small classics corpora did not include texts written in that time period. Again, these results indicate that as languages evolve over time, profession sets and even defining set words would have to evolve to measure gender bias.

5 Conclusion and Future Work

In order for NLP to reflect the rich multilingual, multicultural, and historical heritage of human text, it is essential that NLP techniques be extended beyond modern digital English text to multilingual and diachronic corpora. In this paper, we have explored the challenges of applying an important technique for measuring the gender bias learned by word embedding systems to diachronic corpora. We also have shown how techniques like those pioneered by Bolukbasi et al.

[1] and extended by Chen et al. [7] have fundamental limitations when analyzing corpora spanning large periods of time. We showed that their technique based on analyzing the gender bias of profession words would have difficulty because professions change drastically over hundreds of years. Interestingly, we also documented changes in defining and profession set words over time and also challenges with polysemous/homonymous profession words especially female profession words in Arabic.

In this paper, we have focused mostly on identifying the problems with techniques applied successfully to measure gender bias in modern corpora like Google News or Wikipedia. In the future work, we plan to focus more on modifying profession sets and defining sets over time to overcome these problems. Our results indicate that as languages evolve over time, defining sets and profession sets would have to evolve to measure gender bias.

In this study, we focused on Arabic and Chinese, but we would like to extend our work to more languages. Adding an English corpora may be our next step. Although we like to actively focus on languages besides English, English can serve as an important comparison point because so much of the modern NLP tool chain has been optimized for English. We may be able to study the impact of changes in profession sets and defining sets over time with fewer complicating factors. We would also like to experiment with different advanced Arabic NLP techniques like stemming and lemmatization [20, 21] and see how applying such techniques could improve the results and reduce Arabic's orthographical ambiguity or even other Arabic NLP-related current issues like correcting spelling errors, especially in Arabic dialects, where there are no official orthography rules [23].

6 Acknowledgments

We'd like to thank the Clarkson Open Source Institute for their help and support with infrastructure and hosting of our experiments. We'd like to thank Abigail Matthews and Thomas Middleton for their help and support in writing and reviewing the manuscript.

References

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[3] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[6] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

[7] Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. Gender bias and under-representation in natural language processing across human languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 24–34, 2021.

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? New York, NY, USA, 2021. Association for Computing Machinery.

[9] Esma Wali, Yan Chen, Christopher Mahoney, Thomas Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Neefe Matthews. Is machine learning speaking my language? a critical look at the nlp-pipeline across 8 human languages. *arXiv preprint arXiv:2007.05872*, 2020.

[10] Melvin Wevers. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. *arXiv preprint arXiv:1907.08922*, 2019.

[11] Maja Rudolph and David Blei. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011, 2018.

[12] Yang Xu, Jiasheng Zhang, and David Reitter. Treat the word as a whole or look inside? subword embeddings model language change and typology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 136–145, 2019.

[13] Shamela Library Foundation. Shamila library dataset. *https://shamela.ws/page/download*, 2012.

[14] Waleed A. Yousef, Omar M. Ibrahime, Taha M. Madbouly, Moustafa A. Mahmoud, Ali H. El-Kassas, Ali O. Hassan, and Abdallah R. Albohy. Arabic poem comprehensive dataset. *https://hci-lab.github.io/ArabicPoetry-1-Private/PCD*, 2018.

[15] Diwan. Poetry dataset. *https://www.aldiwan.net/*, 2013.

[16] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, 112:340–349, 2017. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

[17] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations." arxiv preprint. *arXiv preprint arXiv:1802.05365*, 2018.

[18] Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032, 2020.

[19] Michael Grosvald, Sarah Al-Alami, and Ali Idrissi. Word reading in arabic: Influences of diacritics and ambiguity. In *36th West Coast Conference on Formal Linguistics*, pages 176–181. Cascadilla Proceedings Project, 2019.

[20] Youssef Kadri and Jian-Yun Nie. Effective stemming for arabic information retrieval. In *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*, 2006.

[21] Hamdy Mubarak. Build fast and accurate lemmatization for arabic. *arXiv preprint arXiv:1710.06700*, 2017.

[22] Marzieh Karch. How to use the ngram viewer tool in google books. In *https://www.lifewire.com/google-books-ngram-viewer-1616701*, 2021.

[23] Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Appendices

Appendix A: GloVe Models Properties

Time Period/Era	Word Vectors Size	Training Parameters
Shamela Before Islam	2.7 MB	
Shamela Before 1900	1.89 GB	
Shamela After 1900	1.08 GB	Learning Rate (-eta 0.5)
All Shamela	2.2 GB	Training Iterations (-iter 20)
Pre-Islamic Era	15.1 MB	Weighting Function Cutoff (-x-max 10)
Islamic Era	1.3 MB	Number of Threads (-threads 8)
Umayyad Era	43 MB	Weighting Function Exponent (-alpha 0.75)
btw Umayyad & Abbasid	15.4 MB	Memory Limit in GB (-memory 8)
Abbasid Era	122.1 MB	Number of Context Words (-window-size 15)
Andalusian Era	68.5 MB	Minimum Word Count (-min-count 1)
Fatimid Era	76.3 MB	Dimension of Word Vector (-vector-size 256)
Ayyubid Era	69.3 MB	Output Format (-binary 2)
Mamluk Era	94.4 MB	Set Verbosity (-verbose 2)
Ottoman Era	88.6 MB	Word Vector Output (-model 2)
Modern Era	279.9 MB	
All APCD	482.8 MB	

Table A1: GloVe models properties: time period/era, word vectors size, and training parameters.

Appendix B: Defining Set Words

	Before Islam			Before 1900			After 1900			All Books		
English (Arabic) Word	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF
woman (امرأة)	22	16460	0.00134	136198	2075505	0.06562	22861	1335027	0.01712	159081	2520372	0.06312
man (رجل)	16	16460	0.00097	410841	2075505	0.19795	58266	1335027	0.04364	469123	2520372	0.18613
daughter (ابنة)	12	16460	0.00073	30595	2075505	0.01474	4748	1335027	0.00356	35355	2520372	0.01403
son (ولد)	11	16460	0.00067	164806	2075505	0.07941	30992	1335027	0.02321	195809	2520372	0.07769
mother (أم)	32	16460	0.00194	378336	2075505	0.18229	80522	1335027	0.06031	458890	2520372	0.18207
father (أب)	0	16460	0	14068	2075505	0.00678	3147	1335027	0.00236	17215	2520372	0.00683
girl (فتاة)	2	16460	0.00012	1524	2075505	0.00073	3384	1335027	0.00253	4910	2520372	0.00195
boy (فتى)	4	16460	0.00024	13738	2075505	0.00662	3413	1335027	0.00256	17155	2520372	0.00681
queen (ملكة)	0	16460	0	2287	2075505	0.0011	2547	1335027	0.00191	4834	2520372	0.00192
king (ملك)	10	16460	0.00061	137141	2075505	0.06608	26987	1335027	0.02021	164138	2520372	0.06512
wife (زوجة)	1	16460	6.00E-05	15939	2075505	0.00768	4848	1335027	0.00363	20788	2520372	0.00825
husband (زوج)	2	16460	0.00012	39889	2075505	0.01922	5909	1335027	0.00443	45800	2520372	0.01817
madam (سيدة)	0	16460	0	2433	2075505	0.00117	1512	1335027	0.00113	3945	2520372	0.00157
sir (سيد)	2	16460	0.00012	27450	2075505	0.01323	9885	1335027	0.0074	37337	2520372	0.01481

Table B1: Defining set words and their word count (WC), vocabulary size (VS), and word frequency (WF) for Shamela's books through the four-time periods (before Islam, before 1900, after 1900, and all Shamela).

	Pre-Islamic Era			Islamic Era			Umayyad Era			btw Umayyad & Abbasid			Abbasid Era			Andalusia Era			Fatimid Era			Ayyubid Era			Mamluk Era			Ottoman Era			Modern Era			All Eras		
English (Arabic) Word	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF
woman (امرأة)	0	60082	0	0	12388	0	4	119533	3.00E-05	0	65220	0	9	252339	4.00E-05	3	151503	2.00E-05	8	172460	5.00E-05	7	152165	5.00E-05	17	198748	9.00E-05	3	186795	2.00E-05	51	462476	0.00011	102	736576	0.00014
man (رجل)	33	60082	0.00055	2	12388	0.00016	77	119533	0.00064	47	65220	0.00072	376	252339	0.00149	61	151503	0.0004	120	172460	0.0007	77	152165	0.00051	161	198748	0.00081	101	186795	0.00054	646	462476	0.0014	1701	736576	0.00231
daughter (ابنة)	60	60082	0.001	2	12388	0.00016	62	119533	0.00052	24	65220	0.00037	100	252339	0.0004	29	151503	0.00019	62	172460	0.00036	27	152165	0.00018	55	198748	0.00028	46	186795	0.00025	273	462476	0.00059	740	736576	0.001
son (ولد)	14	60082	0.00023	2	12388	0.00016	29	119533	0.00024	16	65220	0.00025	129	252339	0.00051	29	151503	0.00019	73	172460	0.00042	45	152165	0.0003	104	198748	0.00052	92	186795	0.00049	327	462476	0.00071	860	736576	0.00117
mother (أم)	325	60082	0.00541	63	12388	0.00509	1094	119533	0.00915	285	65220	0.00437	1992	252339	0.00789	883	151503	0.00583	1121	172460	0.0065	728	152165	0.00478	1178	198748	0.00593	1465	186795	0.00784	7177	462476	0.01552	16311	736576	0.02214
father (أب)	11	60082	0.00018	1	12388	8.00E-05	73	119533	0.00061	26	65220	0.0004	176	252339	0.0007	49	151503	0.00032	89	172460	0.00052	84	152165	0.00055	107	198748	0.00054	64	186795	0.00034	559	462476	0.00121	1239	736576	0.00168
girl (فتاة)	13	60082	0.00022	0	12388	0	39	119533	0.00033	22	65220	0.00034	90	252339	0.00036	46	151503	0.0003	37	172460	0.00021	26	152165	0.00017	55	198748	0.00028	73	186795	0.00039	513	462476	0.00111	914	736576	0.00124
boy (فتى)	98	60082	0.00163	18	12388	0.00145	273	119533	0.00228	123	65220	0.00189	1249	252339	0.00495	247	151503	0.00163	534	172460	0.0031	458	152165	0.00301	581	198748	0.00292	709	186795	0.0038	3116	462476	0.00674	7406	736576	0.01005
queen (ملكة)	1	60082	2.00E-05	0	12388	0	1	119533	1.00E-05	0	65220	0	1	252339	0	2	151503	1.00E-05	0	172460	0	0	152165	0	1	198748	1.00E-05	2	186795	1.00E-05	14	462476	3.00E-05	22	736576	3.00E-05
king (ملك)	57	60082	0.00095	4	12388	0.00032	116	119533	0.00097	86	65220	0.00132	865	252339	0.00343	704	151503	0.00465	679	172460	0.00394	831	152165	0.00546	945	198748	0.00475	586	186795	0.00314	2444	462476	0.00528	7317	736576	0.00993
wife (زوجة)	1	60082	2.00E-05	0	12388	0	4	119533	3.00E-05	0	65220	0	24	252339	0.0001	4	151503	3.00E-05	5	172460	3.00E-05	4	152165	3.00E-05	16	198748	8.00E-05	7	186795	4.00E-05	50	462476	0.00011	115	736576	0.00016
husband (زوج)	1	60082	2.00E-05	1	12388	8.00E-05	8	119533	7.00E-05	3	65220	5.00E-05	32	252339	0.00013	10	151503	7.00E-05	11	172460	6.00E-05	10	152165	7.00E-05	27	198748	0.00014	25	186795	0.00013	117	462476	0.00025	245	736576	0.00033
madam (سيدة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	13	252339	5.00E-05	1	151503	1.00E-05	0	172460	0	3	152165	2.00E-05	7	198748	4.00E-05	7	186795	4.00E-05	59	462476	0.00013	91	736576	0.00012
sir (سيد)	42	60082	0.0007	2	12388	0.00016	42	119533	0.00035	22	65220	0.00034	256	252339	0.00101	101	151503	0.00067	90	172460	0.00052	112	152165	0.00074	281	198748	0.00141	383	186795	0.00205	1508	462476	0.00326	2839	736576	0.00385

Table B2: Defining set words and their word count (WC), vocabulary size (VS), and word frequency (WF) for APCD's Arabic poems through the 12 eras starting from the Pre-Islamic era to the Modern era, including all eras.

Appendix C: Modern Profession Set Words

	Before Islam			Before 1900			After 1900			All Books		
English (Arabic) Word	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF
programmer_f (مبرمجة)	0	16460	0	0	2075505	0	3	1335027	0	3	2520372	0
programmer_m (مبرمج)	0	16460	0	1	2075505	0	11	1335027	1.00E-05	12	2520372	0
analyst_f (محللة)	1	16460	6.00E-05	384	2075505	0.00019	30	1335027	2.00E-05	415	2520372	0.00016
analyst_m (محلل)	0	16460	0	1272	2075505	0.00061	126	1335027	9.00E-05	1398	2520372	0.00055
trader_f (متداولة)	0	16460	0	290	2075505	0.00014	173	1335027	0.00013	463	2520372	0.00018
trader_m (متداول)	0	16460	0	485	2075505	0.00023	193	1335027	0.00014	678	2520372	0.00027
broker_n (سمسار)	0	16460	0	141	2075505	7.00E-05	38	1335027	3.00E-05	179	2520372	7.00E-05
operator_f (مشغلة)	0	16460	0	154	2075505	7.00E-05	78	1335027	6.00E-05	232	2520372	9.00E-05
operator_m (مشغل)	0	16460	0	58	2075505	3.00E-05	29	1335027	2.00E-05	87	2520372	3.00E-05
waitress_f (مضيفة)	0	16460	0	35	2075505	2.00E-05	22	1335027	2.00E-05	57	2520372	2.00E-05
waiter_m (مضيف)	0	16460	0	152	2075505	7.00E-05	39	1335027	3.00E-05	191	2520372	8.00E-05
chef_f (طباخة)	0	16460	0	49	2075505	2.00E-05	5	1335027	0	54	2520372	2.00E-05
chef_m (طباخ)	0	16460	0	473	2075505	0.00023	66	1335027	5.00E-05	539	2520372	0.00021
pilot_f (طيارة)	0	16460	0	158	2075505	8.00E-05	183	1335027	0.00014	341	2520372	0.00014
pilot_m (طيار)	0	16460	0	223	2075505	0.00011	85	1335027	6.00E-05	308	2520372	0.00012
captain_n (قبطان)	0	16460	0	61	2075505	3.00E-05	32	1335027	2.00E-05	93	2520372	4.00E-05
driver_f (سائقة)	0	16460	0	30	2075505	1.00E-05	8	1335027	1.00E-05	38	2520372	2.00E-05
driver_m (سائق)	0	16460	0	1396	2075505	0.00067	448	1335027	0.00034	1844	2520372	0.00073
mechanical_m (ميكانيكي)	0	16460	0	27	2075505	1.00E-05	54	1335027	4.00E-05	81	2520372	3.00E-05
painter_n (دهان)	0	16460	0	184	2075505	9.00E-05	47	1335027	4.00E-05	231	2520372	9.00E-05
electrician_n (كهربائي)	0	16460	0	6	2075505	0	210	1335027	0.00016	216	2520372	9.00E-05
plumber_n (سباك)	1	16460	6.00E-05	85	2075505	4.00E-05	14	1335027	1.00E-05	100	2520372	4.00E-05
carpenter_n (نجار)	0	16460	0	3277	2075505	0.00158	192	1335027	0.00014	3469	2520372	0.00138
policewoman_f (شرطية)	0	16460	0	3061	2075505	0.00147	1460	1335027	0.00109	4521	2520372	0.00179
policeman_m (شرطي)	0	16460	0	562	2075505	0.00027	171	1335027	0.00013	733	2520372	0.00029
player_f (لاعبة)	0	16460	0	46	2075505	2.00E-05	15	1335027	1.00E-05	61	2520372	2.00E-05
player_m (لاعب)	0	16460	0	533	2075505	0.00026	175	1335027	0.00013	708	2520372	0.00028
doctor_f (طبيبة)	0	16460	0	24	2075505	1.00E-05	67	1335027	5.00E-05	91	2520372	4.00E-05
doctor_m (طبيب)	0	16460	0	3004	2075505	0.00145	2254	1335027	0.00169	5258	2520372	0.00209
nurse_f (ممرضة)	0	16460	0	40	2075505	2.00E-05	74	1335027	6.00E-05	114	2520372	5.00E-05
nurse_m (ممرض)	0	16460	0	367	2075505	0.00018	90	1335027	7.00E-05	457	2520372	0.00018
inspector_f (مفتشة)	0	16460	0	1	2075505	0	13	1335027	1.00E-05	14	2520372	1.00E-05
inspector_m (مفتش)	0	16460	0	55	2075505	3.00E-05	262	1335027	0.0002	317	2520372	0.00013
banker_f (مصرفية)	0	16460	0	3	2075505	0	30	1335027	2.00E-05	33	2520372	1.00E-05
banker_m (مصرفي)	0	16460	0	26	2075505	1.00E-05	27	1335027	2.00E-05	53	2520372	2.00E-05
producer_f (منتجة)	0	16460	0	76	2075505	4.00E-05	119	1335027	9.00E-05	195	2520372	8.00E-05
producer_m (منتج)	0	16460	0	186	2075505	9.00E-05	118	1335027	9.00E-05	304	2520372	0.00012
barber_f (حلاقة)	0	16460	0	20	2075505	1.00E-05	32	1335027	2.00E-05	52	2520372	2.00E-05
barber_m (حلاق)	0	16460	0	500	2075505	0.00024	94	1335027	7.00E-05	594	2520372	0.00024
lecturer_f (محاضرة)	0	16460	0	335	2075505	0.00016	1319	1335027	0.00099	1654	2520372	0.00066
lecturer_m (محاضر)	0	16460	0	1111	2075505	0.00054	196	1335027	0.00015	1307	2520372	0.00052
coach_f (مدربة)	0	16460	0	68	2075505	3.00E-05	49	1335027	4.00E-05	117	2520372	5.00E-05
coach_m (مدرب)	0	16460	0	97	2075505	5.00E-05	95	1335027	7.00E-05	192	2520372	8.00E-05
professor_f (بروفيسورة)	0	16460	0	0	2075505	0	0	1335027	0	0	2520372	0
professor_m (بروفيسور)	0	16460	0	0	2075505	0	3	1335027	0	3	2520372	0
engineer_f (مهندسة)	0	16460	0	8	2075505	0	9	1335027	1.00E-05	17	2520372	1.00E-05
engineer_m (مهندس)	0	16460	0	120	2075505	6.00E-05	336	1335027	0.00025	456	2520372	0.00018
photographer_f (مصورة)	0	16460	0	3223	2075505	0.00155	758	1335027	0.00057	3981	2520372	0.00158
photographer_m (مصور)	0	16460	0	1322	2075505	0.00064	569	1335027	0.00043	1891	2520372	0.00075
translator_f (مترجمة)	0	16460	0	178	2075505	9.00E-05	360	1335027	0.00027	538	2520372	0.00021
translator_m (مترجم)	0	16460	0	3852	2075505	0.00186	793	1335027	0.00059	4645	2520372	0.00184
designer_f (مصممة)	1	16460	6.00E-05	38	2075505	2.00E-05	68	1335027	5.00E-05	107	2520372	4.00E-05
designer_m (مصمم)	0	16460	0	207	2075505	0.0001	114	1335027	9.00E-05	321	2520372	0.00013
journalist_f (صحفية)	0	16460	0	14	2075505	1.00E-05	116	1335027	9.00E-05	130	2520372	5.00E-05
Journalist_m (صحفي)	0	16460	0	124	2075505	6.00E-05	316	1335027	0.00024	440	2520372	0.00017
paramedic_n (مسعف)	0	16460	0	52	2075505	3.00E-05	23	1335027	2.00E-05	75	2520372	3.00E-05
welder_m (لحام)	0	16460	0	298	2075505	0.00014	36	1335027	3.00E-05	334	2520372	0.00013
manager_f (مديرة)	0	16460	0	27	2075505	1.00E-05	65	1335027	5.00E-05	92	2520372	4.00E-05
manager_m (مدير)	0	16460	0	912	2075505	0.00044	1944	1335027	0.00146	2856	2520372	0.00113
accountant_f (محاسبة)	0	16460	0	673	2075505	0.00032	349	1335027	0.00026	1022	2520372	0.00041
accountant_m (محاسب)	0	16460	0	166	2075505	8.00E-05	130	1335027	0.0001	296	2520372	0.00012
technician_f (فنية)	0	16460	0	216	2075505	0.0001	1360	1335027	0.00102	1576	2520372	0.00063
technician_m (فني)	0	16460	0	1614	2075505	0.00078	1015	1335027	0.00076	2629	2520372	0.00104
marketer_f (مسوقة)	0	16460	0	415	2075505	0.0002	673	1335027	0.0005	1088	2520372	0.00043
marketer_m (مسوق)	0	16460	0	706	2075505	0.00034	914	1335027	0.00068	1620	2520372	0.00064
advertiser_f (معلنة)	0	16460	0	102	2075505	5.00E-05	160	1335027	0.00012	262	2520372	0.0001
advertiser_m (معلن)	0	16460	0	203	2075505	0.0001	81	1335027	6.00E-05	284	2520372	0.00011
nanny_f (مربية)	0	16460	0	35	2075505	2.00E-05	84	1335027	6.00E-05	119	2520372	5.00E-05
babysitter_f (حاضنة)	0	16460	0	474	2075505	0.00023	66	1335027	5.00E-05	540	2520372	0.00021
employee_f (موظفة)	0	16460	0	45	2075505	2.00E-05	55	1335027	4.00E-05	100	2520372	4.00E-05
employee_m (موظف)	0	16460	0	97	2075505	5.00E-05	674	1335027	0.0005	771	2520372	0.00031
observer_f (مراقبة)	0	16460	0	602	2075505	0.00029	776	1335027	0.00058	1378	2520372	0.00055
observer_m (مراقب)	0	16460	0	247	2075505	0.00012	216	1335027	0.00016	463	2520372	0.00018
astronomer_n (فلكي)	0	16460	0	96	2075505	5.00E-05	537	1335027	0.0004	633	2520372	0.00025
lawyer_f (محامية)	0	16460	0	7	2075505	0	19	1335027	1.00E-05	26	2520372	1.00E-05
lawyer_m (محامي)	0	16460	0	14	2075505	1.00E-05	49	1335027	4.00E-05	63	2520372	2.00E-05
developer_f (مطورة)	0	16460	0	2	2075505	0	3	1335027	0	5	2520372	0
developer_m (مطور)	0	16460	0	5	2075505	0	3	1335027	0	8	2520372	0
evaluator_f (مختبرة)	0	16460	0	3	2075505	0	3	1335027	0	6	2520372	0
evaluator_m (مختبر)	0	16460	0	130	2075505	6.00E-05	52	1335027	4.00E-05	182	2520372	7.00E-05
writer_f (كاتبة)	0	16460	0	165	2075505	8.00E-05	172	1335027	0.00013	337	2520372	0.00013
writer_m (كاتب)	0	16460	0	22015	2075505	0.01061	5916	1335027	0.00443	27931	2520372	0.01108
boxer_f (ملاكمة)	0	16460	0	1	2075505	0	14	1335027	1.00E-05	15	2520372	1.00E-05
boxer_m (ملاكم)	0	16460	0	5	2075505	0	14	1335027	1.00E-05	19	2520372	1.00E-05
commentator_f (معلقة)	2	16460	0.00012	4647	2075505	0.00224	1445	1335027	0.00108	6094	2520372	0.00242
commentator_m (معلق)	0	16460	0	5104	2075505	0.00246	976	1335027	0.00073	6080	2520372	0.00241
clown_n (مهرج)	0	16460	0	102	2075505	5.00E-05	47	1335027	4.00E-05	149	2520372	6.00E-05
musician_n (موسيقار)	0	16460	0	15	2075505	1.00E-05	40	1335027	3.00E-05	55	2520372	2.00E-05

Table C1: Modern profession set words and their word count (WC), vocabulary size (VS), and word frequency (WF) for Shamela's books through the four-time periods (before Islam, before 1900, after 1900, and all Shamela).

	Pre-Islamic Era			Islamic Era			Umayyad Era			btw Umayyad & Abbasid			Abbasid Era			Andalusia Era			Fatimid Era			Ayyubid Era			Mamluk Era			Ottoman Era			Modern Era			All Eras		
English (Arabic) Word	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**	**WC**	**VS**	**WF**
programmer_f (مبرمجة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
programmer_m (مبرمج)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
analyst_f (محللة)	0	60082	0	0	12388	0	0	119533	0	1	65220	2.00E-05	8	252339	3.00E-05	0	151503	0	3	172460	2.00E-05	2	152165	1.00E-05	2	198748	1.00E-05	0	186795	0	1	462476	0	17	736576	2.00E-05
analyst_m (محلل)	1	60082	2.00E-05	0	12388	0	1	119533	1.00E-05	1	65220	2.00E-05	18	252339	7.00E-05	1	151503	1.00E-05	3	172460	2.00E-05	9	152165	6.00E-05	12	198748	6.00E-05	6	186795	3.00E-05	32	462476	7.00E-05	84	736576	0.00011
trader_f (متداولة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
trader_m (متداول)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	1	152165	1.00E-05	0	198748	0	0	186795	0	0	462476	0	2	736576	0
broker_n (سمسار)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	1	151503	1.00E-05	0	172460	0	0	152165	0	4	198748	2.00E-05	0	186795	0	3	462476	1.00E-05	9	736576	1.00E-05
operator_f (مشغلة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	5	252339	2.00E-05	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	2	186795	1.00E-05	2	462476	0	10	736576	1.00E-05
operator_m (مشغل)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	1	152165	1.00E-05	1	198748	1.00E-05	4	186795	2.00E-05	5	462476	1.00E-05	11	736576	1.00E-05
waitress_f (مضيفة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
waiter_m (مضيف)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	1	151503	1.00E-05	2	172460	1.00E-05	2	152165	1.00E-05	4	198748	2.00E-05	4	186795	2.00E-05	2	462476	0	16	736576	2.00E-05
chef_f (طباخة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	1	462476	0	1	736576	0
chef_m (طباخ)	1	60082	2.00E-05	0	12388	0	0	119533	0	0	65220	0	6	252339	2.00E-05	0	151503	0	0	172460	0	0	152165	0	5	198748	3.00E-05	1	186795	1.00E-05	0	462476	0	13	736576	2.00E-05
pilot_f (طيارة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	3	252339	1.00E-05	1	151503	1.00E-05	0	172460	0	1	152165	1.00E-05	4	198748	2.00E-05	0	186795	0	19	462476	4.00E-05	29	736576	4.00E-05
pilot_m (طيار)	0	60082	0	1	12388	8.00E-05	0	119533	0	2	65220	3.00E-05	6	252339	2.00E-05	13	151503	9.00E-05	8	172460	5.00E-05	5	152165	3.00E-05	6	198748	3.00E-05	1	186795	1.00E-05	18	462476	4.00E-05	60	736576	8.00E-05
captain_n (قبطان)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	1	462476	0	1	736576	0
driver_f (سائقة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	1	172460	1.00E-05	1	152165	1.00E-05	0	198748	0	0	186795	0	1	462476	0	3	736576	0
driver_m (سائق)	4	60082	7.00E-05	0	12388	0	3	119533	3.00E-05	2	65220	3.00E-05	32	252339	0.00013	14	151503	9.00E-05	12	172460	7.00E-05	17	152165	0.00011	41	198748	0.00021	30	186795	0.00016	103	462476	0.00022	258	736576	0.00035
mechanical_m (ميكانيكي)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
painter_n (دهان)	0	60082	0	0	12388	0	0	119533	0	1	65220	2.00E-05	3	252339	1.00E-05	1	151503	1.00E-05	2	172460	1.00E-05	1	152165	1.00E-05	1	198748	1.00E-05	1	186795	1.00E-05	13	462476	3.00E-05	23	736576	3.00E-05
electrician_n (كهربائي)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	4	462476	1.00E-05	4	736576	1.00E-05
plumber_n (سباك)	1	60082	2.00E-05	0	12388	0	0	119533	0	1	65220	2.00E-05	4	252339	2.00E-05	1	151503	1.00E-05	1	172460	1.00E-05	1	152165	1.00E-05	5	198748	3.00E-05	2	186795	1.00E-05	23	462476	5.00E-05	39	736576	5.00E-05
carpenter_n (نجار)	1	60082	2.00E-05	0	12388	0	10	119533	8.00E-05	3	65220	5.00E-05	16	252339	6.00E-05	8	151503	5.00E-05	7	172460	4.00E-05	11	152165	7.00E-05	6	198748	3.00E-05	9	186795	5.00E-05	29	462476	6.00E-05	100	736576	0.00014
policewoman_f (شرطية)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	4	186795	2.00E-05	1	462476	0	6	736576	1.00E-05
policeman_m (شرطي)	0	60082	0	0	12388	0	0	119533	0	1	65220	2.00E-05	13	252339	5.00E-05	1	151503	1.00E-05	4	172460	2.00E-05	1	152165	1.00E-05	10	198748	5.00E-05	0	186795	0	7	462476	2.00E-05	37	736576	5.00E-05
player_f (لاعبة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	2	252339	1.00E-05	1	151503	1.00E-05	4	172460	2.00E-05	1	152165	1.00E-05	2	198748	1.00E-05	0	186795	0	13	462476	3.00E-05	23	736576	3.00E-05
player_m (لاعب)	6	60082	0.0001	0	12388	0	2	119533	2.00E-05	2	65220	3.00E-05	21	252339	8.00E-05	8	151503	5.00E-05	11	172460	6.00E-05	8	152165	5.00E-05	12	198748	6.00E-05	8	186795	4.00E-05	52	462476	0.00011	130	736576	0.00018
doctor_f (طبيبة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	1	186795	1.00E-05	1	462476	0	3	736576	0
doctor_m (طبيب)	4	60082	7.00E-05	6	12388	0.00048	19	119533	0.00016	5	65220	8.00E-05	76	252339	0.0003	30	151503	0.0002	26	172460	0.00015	19	152165	0.00012	41	198748	0.00021	42	186795	0.00022	213	462476	0.00046	481	736576	0.00065
nurse_f (ممرضة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	3	252339	1.00E-05	0	151503	0	1	172460	1.00E-05	2	152165	1.00E-05	1	198748	1.00E-05	1	186795	1.00E-05	5	462476	1.00E-05	13	736576	2.00E-05
nurse_m (ممرض)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	6	252339	2.00E-05	2	151503	1.00E-05	6	172460	3.00E-05	7	152165	5.00E-05	3	198748	2.00E-05	1	186795	1.00E-05	4	462476	1.00E-05	29	736576	4.00E-05
inspector_f (مفتشة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	1	172460	1.00E-05	0	152165	0	1	198748	1.00E-05	0	186795	0	0	462476	0	2	736576	0
inspector_m (مفتش)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	4	462476	1.00E-05	5	736576	1.00E-05
banker_f (مصرفية)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
banker_m (مصرفي)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	2	152165	1.00E-05	0	198748	0	1	186795	1.00E-05	1	462476	0	5	736576	1.00E-05
producer_f (منتجة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	3	462476	1.00E-05	4	736576	1.00E-05
producer_m (منتج)	2	60082	3.00E-05	0	12388	0	0	119533	0	0	65220	0	1	252339	0	1	151503	1.00E-05	1	172460	1.00E-05	6	152165	4.00E-05	1	198748	1.00E-05	2	186795	1.00E-05	11	462476	2.00E-05	25	736576	3.00E-05
barber_f (حلاقة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	0	462476	0	1	736576	0
barber_m (حلاق)	1	60082	2.00E-05	0	12388	0	1	119533	1.00E-05	0	65220	0	2	252339	1.00E-05	0	151503	0	2	172460	1.00E-05	0	152165	0	0	198748	0	1	186795	1.00E-05	5	462476	1.00E-05	12	736576	2.00E-05
lecturer_f (محاضرة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	1	252339	0	0	151503	0	1	172460	1.00E-05	1	152165	1.00E-05	0	198748	0	0	186795	0	7	462476	2.00E-05	11	736576	1.00E-05
lecturer_m (محاضر)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	1	252339	0	0	151503	0	0	172460	0	1	152165	1.00E-05	3	198748	2.00E-05	1	186795	1.00E-05	7	462476	2.00E-05	14	736576	2.00E-05
coach_f (مدربة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	2	252339	1.00E-05	1	151503	1.00E-05	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	3	462476	1.00E-05	7	736576	1.00E-05
coach_m (مدرب)	5	60082	8.00E-05	0	12388	0	1	119533	1.00E-05	1	65220	2.00E-05	6	252339	2.00E-05	0	151503	0	1	172460	1.00E-05	2	152165	1.00E-05	3	198748	2.00E-05	3	186795	2.00E-05	24	462476	5.00E-05	46	736576	6.00E-05
professor_f (بروفيسورة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
professor_m (بروفيسور)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
engineer_f (مهندسة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	1	151503	1.00E-05	0	172460	0	0	152165	0	0	198748	0	0	186795	0	1	462476	0	2	736576	0
engineer_m (مهندس)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	2	151503	1.00E-05	0	172460	0	1	152165	1.00E-05	0	198748	0	2	186795	1.00E-05	17	462476	4.00E-05	23	736576	3.00E-05
photographer_f (مصورة)	0	60082	0	0	12388	0	4	119533	3.00E-05	3	65220	5.00E-05	4	252339	2.00E-05	1	151503	1.00E-05	3	172460	2.00E-05	2	152165	1.00E-05	2	198748	1.00E-05	0	186795	0	11	462476	2.00E-05	30	736576	4.00E-05
photographer_m (مصور)	0	60082	0	0	12388	0	2	119533	2.00E-05	1	65220	2.00E-05	20	252339	8.00E-05	5	151503	3.00E-05	8	172460	5.00E-05	3	152165	2.00E-05	11	198748	6.00E-05	21	186795	0.00011	61	462476	0.00013	132	736576	0.00018
translator_f (مترجمة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	2	172460	1.00E-05	0	152165	0	0	198748	0	0	186795	0	3	462476	1.00E-05	5	736576	1.00E-05
translator_m (مترجم)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	6	252339	2.00E-05	10	151503	7.00E-05	6	172460	3.00E-05	3	152165	2.00E-05	2	198748	1.00E-05	13	186795	7.00E-05	20	462476	4.00E-05	60	736576	8.00E-05
designer_f (مصممة)	1	60082	2.00E-05	0	12388	0	1	119533	1.00E-05	0	65220	0	3	252339	1.00E-05	0	151503	0	4	172460	2.00E-05	0	152165	0	0	198748	0	2	186795	1.00E-05	2	462476	0	13	736576	2.00E-05
designer_m (مصمم)	4	60082	7.00E-05	0	12388	0	5	119533	4.00E-05	1	65220	2.00E-05	18	252339	7.00E-05	22	151503	0.00015	12	172460	7.00E-05	12	152165	8.00E-05	7	198748	4.00E-05	5	186795	3.00E-05	25	462476	5.00E-05	111	736576	0.00015
journalist_f (صحفية)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
Journalist_m (صحفي)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	3	252339	1.00E-05	1	151503	1.00E-05	0	172460	0	0	152165	0	1	198748	1.00E-05	1	186795	1.00E-05	14	462476	3.00E-05	20	736576	3.00E-05
paramedic_n (مسعف)	0	60082	0	0	12388	0	2	119533	2.00E-05	0	65220	0	5	252339	2.00E-05	8	151503	5.00E-05	3	172460	2.00E-05	4	152165	3.00E-05	13	198748	7.00E-05	12	186795	6.00E-05	51	462476	0.00011	98	736576	0.00013
welder_m (لحام)	1	60082	2.00E-05	0	12388	0	1	119533	1.00E-05	0	65220	0	5	252339	2.00E-05	2	151503	1.00E-05	0	172460	0	2	152165	1.00E-05	0	198748	0	1	186795	1.00E-05	4	462476	1.00E-05	16	736576	2.00E-05
manager_f (مديرة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	4	198748	2.00E-05	0	186795	0	4	462476	1.00E-05	9	736576	1.00E-05
manager_m (مدير)	2	60082	3.00E-05	0	12388	0	0	119533	0	1	65220	2.00E-05	11	252339	4.00E-05	14	151503	9.00E-05	5	172460	3.00E-05	10	152165	7.00E-05	8	198748	4.00E-05	14	186795	7.00E-05	81	462476	0.00018	146	736576	0.0002
accountant_f (محاسبة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	2	252339	1.00E-05	2	151503	1.00E-05	2	172460	1.00E-05	3	152165	2.00E-05	0	198748	0	0	186795	0	1	462476	0	11	736576	1.00E-05
accountant_m (محاسب)	1	60082	2.00E-05	0	12388	0	1	119533	1.00E-05	0	65220	0	0	252339	0	0	151503	0	0	172460	0	1	152165	1.00E-05	0	198748	0	1	186795	1.00E-05	9	462476	2.00E-05	13	736576	2.00E-05
technician_f (فنية)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	1	151503	1.00E-05	0	172460	0	0	152165	0	0	198748	0	1	186795	1.00E-05	9	462476	2.00E-05	11	736576	1.00E-05
technician_m (فني)	4	60082	7.00E-05	0	12388	0	2	119533	2.00E-05	1	65220	2.00E-05	19	252339	8.00E-05	14	151503	9.00E-05	10	172460	6.00E-05	9	152165	6.00E-05	32	198748	0.00016	43	186795	0.00023	97	462476	0.00021	231	736576	0.00031
marketer_f (مسوقة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	1	172460	1.00E-05	0	152165	0	0	198748	0	0	186795	0	5	462476	1.00E-05	6	736576	1.00E-05
marketer_m (مسوق)	0	60082	0	0	12388	0	3	119533	3.00E-05	0	65220	0	1	252339	0	1	151503	1.00E-05	2	172460	1.00E-05	0	152165	0	0	198748	0	2	186795	1.00E-05	14	462476	3.00E-05	23	736576	3.00E-05
advertiser_f (معلنة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	3	252339	1.00E-05	2	151503	1.00E-05	1	172460	1.00E-05	1	152165	1.00E-05	2	198748	1.00E-05	4	186795	2.00E-05	8	462476	2.00E-05	22	736576	3.00E-05
advertiser_m (معلن)	0	60082	0	0	12388	0	0	119533	0	1	65220	2.00E-05	4	252339	2.00E-05	3	151503	2.00E-05	3	172460	2.00E-05	7	152165	5.00E-05	3	198748	2.00E-05	5	186795	3.00E-05	20	462476	4.00E-05	46	736576	6.00E-05
nanny_f (مربية)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
babysitter_f (حاضنة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	1	252339	0	0	151503	0	1	172460	1.00E-05	0	152165	0	1	198748	1.00E-05	0	186795	0	2	462476	0	6	736576	1.00E-05
employee_f (موظفة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
employee_m (موظف)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	8	462476	2.00E-05	8	736576	1.00E-05
observer_f (مراقبة)	0	60082	0	0	12388	0	2	119533	2.00E-05	1	65220	2.00E-05	16	252339	6.00E-05	7	151503	5.00E-05	7	172460	4.00E-05	3	152165	2.00E-05	3	198748	2.00E-05	1	186795	1.00E-05	16	462476	3.00E-05	56	736576	8.00E-05
observer_m (مراقب)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	17	252339	7.00E-05	11	151503	7.00E-05	8	172460	5.00E-05	4	152165	3.00E-05	5	198748	3.00E-05	6	186795	3.00E-05	50	462476	0.00011	101	736576	0.00014
astronomer_n (فلكي)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	4	151503	3.00E-05	3	172460	2.00E-05	4	152165	3.00E-05	2	198748	1.00E-05	2	186795	1.00E-05	21	462476	5.00E-05	37	736576	5.00E-05
lawyer_f (محامية)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	0	462476	0	2	736576	0
lawyer_m (محامي)	0	60082	0	0	12388	0	2	119533	2.00E-05	0	65220	0	4	252339	2.00E-05	0	151503	0	1	172460	1.00E-05	1	152165	1.00E-05	2	198748	1.00E-05	1	186795	1.00E-05	5	462476	1.00E-05	16	736576	2.00E-05
developer_f (مطورة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	3	462476	1.00E-05	3	736576	0
developer_m (مطور)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
evaluator_f (مختبرة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
evaluator_m (مختبر)	1	60082	2.00E-05	0	12388	0	0	119533	0	1	65220	2.00E-05	9	252339	4.00E-05	3	151503	2.00E-05	5	172460	3.00E-05	2	152165	1.00E-05	5	198748	3.00E-05	7	186795	4.00E-05	19	462476	4.00E-05	52	736576	7.00E-05
writer_f (كاتبة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	2	198748	1.00E-05	0	186795	0	7	462476	2.00E-05	10	736576	1.00E-05
writer_m (كاتب)	4	60082	7.00E-05	0	12388	0	5	119533	4.00E-05	1	65220	2.00E-05	66	252339	0.00026	27	151503	0.00018	33	172460	0.00019	25	152165	0.00016	57	198748	0.00029	31	186795	0.00017	202	462476	0.00044	451	736576	0.00061
boxer_f (ملاكمة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	1	462476	0	1	736576	0
boxer_m (ملاكم)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	1	172460	1.00E-05	0	152165	0	0	198748	0	0	186795	0	0	462476	0	1	736576	0
commentator_f (معلقة)	1	60082	2.00E-05	0	12388	0	12	119533	0.0001	3	65220	5.00E-05	20	252339	8.00E-05	0	151503	0	2	172460	1.00E-05	2	152165	1.00E-05	3	198748	2.00E-05	2	186795	1.00E-05	23	462476	5.00E-05	68	736576	9.00E-05
commentator_m (معلق)	4	60082	7.00E-05	0	12388	0	25	119533	0.00021	7	65220	0.00011	29	252339	0.00011	5	151503	3.00E-05	5	172460	3.00E-05	8	152165	5.00E-05	14	198748	7.00E-05	8	186795	4.00E-05	38	462476	8.00E-05	143	736576	0.00019
clown_n (مهرج)	2	60082	3.00E-05	0	12388	0	0	119533	0	0	65220	0	3	252339	1.00E-05	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	4	462476	1.00E-05	9	736576	1.00E-05
musician_n (موسيقار)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	1	462476	0	2	736576	0

Table C2: Modern profession set words and their word count (WC), vocabulary size (VS), and word frequency (WF) for APCD's Arabic poems through the 12 eras starting from the Pre-Islamic era to the Modern era, including all eras.

Appendix D: Historical Profession Set Words

English (Arabic) Word	Before Islam			Before 1900			After 1900			All Books		
	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF
slavor_f (نخاسة)	0	16460	0	6	2075505	0	6	1335027	0	12	2520372	0
slavor_m (نخاس)	0	16460	0	118	2075505	6.00E-05	12	1335027	1.00E-05	130	2520372	5.00E-05
sculptor_f (نحاتة)	0	16460	0	31	2075505	1.00E-05	12	1335027	1.00E-05	43	2520372	2.00E-05
sculptor_m (نحات)	0	16460	0	25	2075505	1.00E-05	32	1335027	2.00E-05	57	2520372	2.00E-05
shoemaker_m (إسكافي)	0	16460	0	19	2075505	1.00E-05	4	1335027	0	23	2520372	1.00E-05
gardener_n (بستاني)	0	16460	0	264	2075505	0.00013	51	1335027	4.00E-05	315	2520372	0.00012
merchant_f (تاجرة)	0	16460	0	92	2075505	4.00E-05	25	1335027	2.00E-05	117	2520372	5.00E-05
merchant_m (تاجر)	1	16460	6.00E-05	2264	2075505	0.00109	855	1335027	0.00064	3120	2520372	0.00124
translator_n (ترجمان)	0	16460	0	1604	2075505	0.00077	596	1335027	0.00045	2200	2520372	0.00087
jeweler_m (جواهري)	0	16460	0	3	2075505	0	5	1335027	0	8	2520372	0
lumberjack_f (حطابة)	0	16460	0	9	2075505	0	2	1335027	0	11	2520372	0
lumberjack_m (حطاب)	0	16460	0	155	2075505	7.00E-05	83	1335027	6.00E-05	238	2520372	9.00E-05
tanner_m (دباغ)	0	16460	0	546	2075505	0.00026	70	1335027	5.00E-05	616	2520372	0.00024
auctioneer_f (دلالة)	0	16460	0	34420	2075505	0.01658	7392	1335027	0.00554	41812	2520372	0.01659
auctioneer_m (دلال)	0	16460	0	700	2075505	0.00034	213	1335027	0.00016	913	2520372	0.00036
shepherd_f (راعية)	0	16460	0	687	2075505	0.00033	230	1335027	0.00017	917	2520372	0.00036
shepherd_m (راعي)	2	16460	0.00012	2703	2075505	0.0013	486	1335027	0.00036	3191	2520372	0.00127
glazier_m (زجاج)	0	16460	0	1185	2075505	0.00057	412	1335027	0.00031	1597	2520372	0.00063
oiler_m (زيات)	0	16460	0	151	2075505	7.00E-05	49	1335027	4.00E-05	200	2520372	8.00E-05
poet_f (شاعرة)	0	16460	0	481	2075505	0.00023	467	1335027	0.00035	948	2520372	0.00038
poet_m (شاعر)	1	16460	6.00E-05	17436	2075505	0.0084	10164	1335027	0.00761	27601	2520372	0.01095
goldsmith_m (صائغ)	0	16460	0	366	2075505	0.00018	122	1335027	9.00E-05	488	2520372	0.00019
moneychanger_m (صيرفي)	0	16460	0	198	2075505	0.0001	23	1335027	2.00E-05	221	2520372	9.00E-05
miller_f (طحانة)	0	16460	0	10	2075505	0	2	1335027	0	12	2520372	0
miller_m (طحان)	0	16460	0	126	2075505	6.00E-05	18	1335027	1.00E-05	144	2520372	6.00E-05
porter_m (عتال)	0	16460	0	12	2075505	1.00E-05	2	1335027	0	14	2520372	1.00E-05
fortune_teller_f (عرافة)	0	16460	0	210	2075505	0.0001	56	1335027	4.00E-05	266	2520372	0.00011
fortune_teller_m (عراف)	0	16460	0	150	2075505	7.00E-05	82	1335027	6.00E-05	232	2520372	9.00E-05
spice_dealer_f (عطارة)	0	16460	0	101	2075505	5.00E-05	8	1335027	1.00E-05	109	2520372	4.00E-05
spice_dealer_m (عطار)	0	16460	0	597	2075505	0.00029	142	1335027	0.00011	739	2520372	0.00029
peasant_m (فلاح)	0	16460	0	1797	2075505	0.00087	671	1335027	0.0005	2468	2520372	0.00098
midwife_f (قابلة)	0	16460	0	1643	2075505	0.00079	815	1335027	0.00061	2458	2520372	0.00098
butcher_m (قصاب)	0	16460	0	219	2075505	0.00011	52	1335027	4.00E-05	271	2520372	0.00011
muezzin_m (مؤذن)	0	16460	0	4641	2075505	0.00224	952	1335027	0.00071	5593	2520372	0.00222
discipliner_f (مؤدبة)	0	16460	0	46	2075505	2.00E-05	39	1335027	3.00E-05	85	2520372	3.00E-05
discipliner_m (مؤدب)	0	16460	0	1125	2075505	0.00054	218	1335027	0.00016	1343	2520372	0.00053
hairdresser_f (ماشطة)	0	16460	0	324	2075505	0.00016	32	1335027	2.00E-05	356	2520372	0.00014
usurer_f (مرابية)	0	16460	0	1	2075505	0	1	1335027	0	2	2520372	0
usurer_m (مرابي)	0	16460	0	19	2075505	1.00E-05	10	1335027	1.00E-05	29	2520372	1.00E-05
wet_nurse_f (مرضعة)	1	16460	6.00E-05	1192	2075505	0.00057	350	1335027	0.00026	1543	2520372	0.00061
coppersmith_m (نحاس)	2	16460	0.00012	2968	2075505	0.00143	482	1335027	0.00036	3452	2520372	0.00137
tattooist_f (واشمة)	0	16460	0	60	2075505	3.00E-05	5	1335027	0	65	2520372	3.00E-05
tattooist_m (واشم)	0	16460	0	24	2075505	1.00E-05	1	1335027	0	25	2520372	1.00E-05
weaver_f (حائكة)	0	16460	0	5	2075505	0	3	1335027	0	8	2520372	0
weaver_m (حائك)	0	16460	0	545	2075505	0.00026	50	1335027	4.00E-05	595	2520372	0.00024
stableman_m (سائس)	1	16460	6.00E-05	310	2075505	0.00015	55	1335027	4.00E-05	366	2520372	0.00015
jailer_f (سجانة)	0	16460	0	4	2075505	0	1	1335027	0	5	2520372	0
jailer_m (سجان)	0	16460	0	34	2075505	2.00E-05	14	1335027	1.00E-05	48	2520372	2.00E-05
author_f (مؤلفة)	0	16460	0	790	2075505	0.00038	931	1335027	0.0007	1721	2520372	0.00068
author_m (مؤلف)	1	16460	6.00E-05	4720	2075505	0.00227	3597	1335027	0.00269	8318	2520372	0.0033
historian_f (مؤرخة)	0	16460	0	235	2075505	0.00011	229	1335027	0.00017	464	2520372	0.00018
historian_m (مؤرخ)	0	16460	0	1352	2075505	0.00065	3685	1335027	0.00276	5037	2520372	0.002
builder_m (بناء)	2	16460	0.00012	52863	2075505	0.02547	14133	1335027	0.01059	66998	2520372	0.02658
veterinarian_m (بيطار)	0	16460	0	259	2075505	0.00012	43	1335027	3.00E-05	302	2520372	0.00012
upholsterer_f (منجدة)	0	16460	0	27	2075505	1.00E-05	8	1335027	1.00E-05	35	2520372	1.00E-05
upholsterer_m (منجد)	0	16460	0	384	2075505	0.00019	88	1335027	7.00E-05	472	2520372	0.00019
chamberlain_m (حاجب)	0	16460	0	9518	2075505	0.00459	914	1335027	0.00068	10432	2520372	0.00414
bartender_f (ساقية)	0	16460	0	668	2075505	0.00032	151	1335027	0.00011	819	2520372	0.00032
bartender_m (ساقي)	0	16460	0	1264	2075505	0.00061	287	1335027	0.00021	1551	2520372	0.00062

Table D1: Historical profession set words and their word count (WC), vocabulary size (VS), and word frequency (WF) for Shamela's books through the four-time periods (before Islam, before 1900, after 1900, and all Shamela).

	Pre-Islamic Era			Islamic Era			Umayyad Era			btw Umayyad & Abbasid			Abbasid Era			Andalusia Era			Fatimid Era			Ayyubid Era			Mamluk Era			Ottoman Era			Modern Era			All Eras	
English (Arabic) Word	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF	WC	VS	WF
slavor_f (خادمة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	0	462476	0	1	736576	0
slavor_m (عامر)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	1	252339	0	0	151503	0	3	172460	2.00E-05	1	152165	1.00E-05	0	198748	0	2	186795	1.00E-05	3	462476	1.00E-05	11	736576	1.00E-05
sculptor_f (نحاتة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
sculptor_m (نحات)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	2	462476	0	2	736576	0
shoemaker_m (إسكافي)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
gardener_n (بستاني)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	4	252339	2.00E-05	4	151503	3.00E-05	0	172460	0	2	152165	1.00E-05	3	198748	2.00E-05	1	186795	1.00E-05	13	462476	3.00E-05	27	736576	4.00E-05
merchant_f (تاجرة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	1	151503	1.00E-05	1	172460	1.00E-05	0	152165	0	0	198748	0	0	186795	0	1	462476	0	4	736576	1.00E-05
merchant_m (تاجر)	2	60082	3.00E-05	0	12388	0	9	119533	8.00E-05	2	65220	3.00E-05	27	252339	0.00011	9	151503	6.00E-05	20	172460	0.00012	11	152165	7.00E-05	11	198748	6.00E-05	8	186795	4.00E-05	51	462476	0.00011	150	736576	0.0002
translator_n (ترجمان)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	3	252339	1.00E-05	4	151503	3.00E-05	9	172460	5.00E-05	6	152165	4.00E-05	14	198748	7.00E-05	13	186795	7.00E-05	40	462476	9.00E-05	89	736576	0.00012
jeweler_m (جواهري)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	3	152165	2.00E-05	2	198748	1.00E-05	0	186795	0	4	462476	1.00E-05	9	736576	1.00E-05
lumberjack_f (حطابة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
lumberjack_m (حطاب)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	4	462476	1.00E-05	4	736576	1.00E-05
tanner_m (دباغ)	0	60082	0	0	12388	0	0	119533	0	1	65220	2.00E-05	0	252339	0	1	151503	1.00E-05	0	172460	0	1	152165	1.00E-05	1	198748	1.00E-05	1	186795	1.00E-05	2	462476	0	7	736576	1.00E-05
auctioneer_f (دلالة)	1	60082	2.00E-05	0	12388	0	0	119533	0	2	65220	3.00E-05	13	252339	5.00E-05	14	151503	9.00E-05	4	172460	2.00E-05	5	152165	3.00E-05	13	198748	7.00E-05	6	186795	3.00E-05	35	462476	8.00E-05	93	736576	0.00013
auctioneer_m (دلال)	5	60082	8.00E-05	0	12388	0	7	119533	6.00E-05	3	65220	5.00E-05	43	252339	0.00017	13	151503	9.00E-05	14	172460	8.00E-05	10	152165	7.00E-05	28	198748	0.00014	28	186795	0.00015	127	462476	0.00027	278	736576	0.00038
shepherd_f (راعية)	1	60082	2.00E-05	0	12388	0	1	119533	1.00E-05	1	65220	2.00E-05	2	252339	1.00E-05	0	151503	0	4	172460	2.00E-05	1	152165	1.00E-05	0	198748	0	0	186795	0	13	462476	3.00E-05	23	736576	3.00E-05
shepherd_m (راعي)	4	60082	7.00E-05	0	12388	0	24	119533	0.0002	6	65220	9.00E-05	42	252339	0.00017	9	151503	6.00E-05	13	172460	8.00E-05	9	152165	6.00E-05	11	198748	6.00E-05	25	186795	0.00013	116	462476	0.00025	259	736576	0.00035
glazier_m (زجاج)	1	60082	2.00E-05	1	12388	8.00E-05	8	119533	7.00E-05	2	65220	3.00E-05	30	252339	0.00012	12	151503	8.00E-05	18	172460	0.0001	6	152165	4.00E-05	16	198748	8.00E-05	4	186795	2.00E-05	44	462476	0.0001	142	736576	0.00019
oiler_m (زيات)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	3	252339	1.00E-05	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	2	462476	0	5	736576	1.00E-05
poet_f (شاعرة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	11	462476	2.00E-05	13	736576	2.00E-05
poet_m (شاعر)	2	60082	3.00E-05	0	12388	0	20	119533	0.00017	8	65220	0.00012	118	252339	0.00047	47	151503	0.00031	51	172460	0.0003	51	152165	0.00034	61	198748	0.00031	67	186795	0.00036	884	462476	0.00191	1309	736576	0.00178
goldsmith_m (صائغ)	0	60082	0	1	12388	8.00E-05	0	119533	0	1	65220	2.00E-05	8	252339	3.00E-05	2	151503	1.00E-05	5	172460	3.00E-05	0	152165	0	7	198748	4.00E-05	7	186795	4.00E-05	20	462476	4.00E-05	51	736576	7.00E-05
moneychanger_m (صيرفي)	0	60082	0	0	12388	0	0	119533	0	1	65220	2.00E-05	1	252339	0	1	151503	1.00E-05	3	172460	2.00E-05	4	152165	3.00E-05	1	198748	1.00E-05	3	186795	2.00E-05	7	462476	2.00E-05	21	736576	3.00E-05
miller_f (طحانة)	0	60082	0	0	12388	0	0	119533	0	1	65220	2.00E-05	0	252339	0	0	151503	0	1	172460	1.00E-05	0	152165	0	0	198748	0	0	186795	0	2	462476	0	2	736576	0
miller_m (طحان)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	5	462476	1.00E-05	7	736576	1.00E-05
porter_m (عتال)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
fortune_teller_f (عرافة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	1	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	2	462476	0	4	736576	1.00E-05
fortune_teller_m (عراف)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	4	252339	2.00E-05	0	151503	0	1	172460	1.00E-05	2	152165	1.00E-05	2	198748	1.00E-05	2	186795	1.00E-05	5	462476	1.00E-05	16	736576	2.00E-05
spice_dealer_f (عطارة)	0	60082	0	0	12388	0	5	119533	4.00E-05	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	1	462476	0	7	736576	1.00E-05
spice_dealer_m (عطار)	0	60082	0	0	12388	0	3	119533	3.00E-05	2	65220	3.00E-05	17	252339	7.00E-05	3	151503	2.00E-05	4	172460	2.00E-05	11	152165	7.00E-05	10	198748	5.00E-05	2	186795	1.00E-05	9	462476	2.00E-05	61	736576	8.00E-05
peasant_m (فلاح)	5	60082	8.00E-05	0	12388	0	2	119533	2.00E-05	1	65220	2.00E-05	18	252339	7.00E-05	21	151503	0.00014	20	172460	0.00012	25	152165	0.00016	31	198748	0.00016	80	186795	0.00043	160	462476	0.00035	363	736576	0.00049
midwife_f (قابلة)	1	60082	2.00E-05	0	12388	0	0	119533	0	0	65220	0	0	252339	0	1	151503	1.00E-05	0	172460	0	3	152165	2.00E-05	2	198748	1.00E-05	0	186795	0	9	462476	2.00E-05	16	736576	2.00E-05
butcher_m (قصاب)	1	60082	2.00E-05	0	12388	0	0	119533	0	1	65220	2.00E-05	0	252339	0	1	151503	1.00E-05	0	172460	0	0	152165	0	1	198748	1.00E-05	1	186795	1.00E-05	7	462476	2.00E-05	12	736576	2.00E-05
muezzin_m (مؤذن)	1	60082	2.00E-05	0	12388	0	0	119533	0	1	65220	2.00E-05	20	252339	8.00E-05	7	151503	5.00E-05	2	172460	1.00E-05	8	152165	5.00E-05	15	198748	8.00E-05	12	186795	6.00E-05	30	462476	6.00E-05	96	736576	0.00013
discipliner_f (مؤدبة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	2	252339	1.00E-05	0	151503	0	0	172460	0	1	152165	1.00E-05	0	198748	0	1	186795	1.00E-05	0	462476	0	4	736576	1.00E-05
discipliner_m (مؤدب)	0	60082	0	0	12388	0	1	119533	1.00E-05	1	65220	2.00E-05	10	252339	4.00E-05	1	151503	1.00E-05	6	172460	3.00E-05	4	152165	3.00E-05	4	198748	2.00E-05	4	186795	2.00E-05	21	462476	5.00E-05	52	736576	7.00E-05
hairdresser_f (ماشطة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	3	198748	2.00E-05	0	186795	0	4	462476	1.00E-05	7	736576	1.00E-05
usurer_f (مرابية)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
usurer_m (مرابي)	1	60082	2.00E-05	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	1	198748	1.00E-05	0	186795	0	1	462476	0	3	736576	0
wet_nurse_f (مرضعة)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	4	252339	2.00E-05	2	151503	1.00E-05	4	172460	2.00E-05	1	152165	1.00E-05	1	198748	1.00E-05	2	186795	1.00E-05	11	462476	2.00E-05	26	736576	4.00E-05
coppersmith_m (صفار)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	5	252339	2.00E-05	3	151503	2.00E-05	4	172460	2.00E-05	1	152165	1.00E-05	2	198748	1.00E-05	4	186795	2.00E-05	34	462476	7.00E-05	53	736576	7.00E-05
tattooist_f (وشامة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
tattooist_m (وشام)	0	60082	0	0	12388	0	1	119533	1.00E-05	0	65220	0	1	252339	0	1	151503	1.00E-05	0	172460	0	1	152165	1.00E-05	0	198748	0	1	186795	1.00E-05	11	462476	2.00E-05	16	736576	2.00E-05
weaver_f (حائكة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	1	462476	0	1	736576	0
weaver_m (حائك)	1	60082	2.00E-05	0	12388	0	0	119533	0	0	65220	0	6	252339	2.00E-05	4	151503	3.00E-05	2	172460	1.00E-05	0	152165	0	2	198748	1.00E-05	4	186795	2.00E-05	1	462476	0	20	736576	3.00E-05
stableman_m (سائس)	2	60082	3.00E-05	0	12388	0	0	119533	0	2	65220	3.00E-05	8	252339	3.00E-05	5	151503	3.00E-05	5	172460	3.00E-05	4	152165	3.00E-05	2	198748	1.00E-05	9	186795	5.00E-05	15	462476	3.00E-05	52	736576	7.00E-05
jailer_f (حجانة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	0	198748	0	0	186795	0	0	462476	0	0	736576	0
jailer_m (حجان)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	0	172460	0	1	152165	1.00E-05	0	198748	0	1	186795	1.00E-05	12	462476	3.00E-05	15	736576	2.00E-05
author_f (مؤلفة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	6	252339	2.00E-05	1	151503	1.00E-05	1	172460	1.00E-05	0	152165	0	4	198748	2.00E-05	2	186795	1.00E-05	11	462476	2.00E-05	25	736576	3.00E-05
author_m (مؤلف)	0	60082	0	0	12388	0	2	119533	2.00E-05	0	65220	0	12	252339	5.00E-05	2	151503	1.00E-05	2	172460	1.00E-05	1	152165	1.00E-05	2	198748	1.00E-05	18	186795	0.0001	30	462476	6.00E-05	69	736576	9.00E-05
historian_f (مؤرخة)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	0	252339	0	0	151503	0	0	172460	0	0	152165	0	2	198748	1.00E-05	6	186795	3.00E-05	177	462476	0.00038	185	736576	0.00025
historian_m (مؤرخ)	0	60082	0	0	12388	0	0	119533	0	0	65220	0	2	252339	1.00E-05	1	151503	1.00E-05	0	172460	0	0	152165	0	3	198748	2.00E-05	3	186795	2.00E-05	90	462476	0.00019	99	736576	0.00013
builder_m (بناء)	4	60082	7.00E-05	1	12388	8.00E-05	21	119533	0.00018	3	65220	5.00E-05	75	252339	0.0003	45	151503	0.0003	35	172460	0.0002	30	152165	0.0002	31	198748	0.00016	53	186795	0.00028	368	462476	0.0008	666	736576	0.0009
veterinarian_m (بطار)	0	60082	0	0	12388	0	4	119533	3.00E-05	0	65220	0	4	252339	2.00E-05	2	151503	1.00E-05	0	172460	0	2	152165	1.00E-05	1	198748	1.00E-05	1	186795	1.00E-05	3	462476	1.00E-05	17	736576	2.00E-05
upholsterer_f (منجدة)	1	60082	2.00E-05	0	12388	0	0	119533	0	0	65220	0	1	252339	0	0	151503	0	1	172460	1.00E-05	0	152165	0	0	198748	0	2	186795	1.00E-05	3	462476	1.00E-05	8	736576	1.00E-05
upholsterer_m (منجد)	0	60082	0	0	12388	0	2	119533	2.00E-05	4	65220	6.00E-05	30	252339	0.00012	28	151503	0.00018	13	172460	8.00E-05	21	152165	0.00014	21	198748	0.00011	30	186795	0.00016	125	462476	0.00027	274	736576	0.00037
chamberlain_m (حاجب)	12	60082	0.0002	3	12388	0.00024	24	119533	0.0002	6	65220	9.00E-05	64	252339	0.00025	40	151503	0.00026	34	172460	0.0002	30	152165	0.0002	65	198748	0.00033	53	186795	0.00028	128	462476	0.00028	459	736576	0.00062
bartender_f (ساقية)	2	60082	3.00E-05	0	12388	0	0	119533	0	0	65220	0	11	252339	4.00E-05	2	151503	1.00E-05	3	172460	2.00E-05	0	152165	0	8	198748	4.00E-05	1	186795	1.00E-05	28	462476	6.00E-05	55	736576	7.00E-05
bartender_m (ساقي)	6	60082	0.0001	0	12388	0	10	119533	8.00E-05	10	65220	0.00015	37	252339	0.00015	36	151503	0.00024	11	172460	6.00E-05	23	152165	0.00015	37	198748	0.00019	67	186795	0.00036	197	462476	0.00043	434	736576	0.00059

Table D2: Historical profession set words and their word count (WC), vocabulary size (VS), and word frequency (WF) for APCD's Arabic poems through the 12 eras starting from the Pre-Islamic era to the Modern era, including all eras.