An Exploratory Analysis of a   
Small Corpus of Spoken Omani Arabic

Jonathan R. Schmid, Leipzig

Recent times have seen an exponential rise in the amount of recorded words produced per ever-decreasing unit of time. The Internet now poses only a few limitations to anyone who would have their thoughts heard (or read) and while the ease of production and proliferation of all these data might be considered a godsend by their senders and recipients, to the Humanities they present a challenge as much as an opportunity.

Quantitative methods are now finding their way into fields where work has traditionally been qualitative and the term of Digital Humanities has established itself[[1]](#footnote-1). The most visible advances at the boundaries of information technology and language have arguably been in statistics-based machine translation, speech synthesis and voice recognition. Other technologies that are perhaps more interesting to the scholar allow analyses of large amounts of text both within and beyond their meta data. Word frequency analysis, stylometric analysis, network analysis, topic modelling and more have advanced to a point that theoretical knowledge is now available in more comprehensible software form[[2]](#footnote-2), ready to be used by anyone.

Regrettably, many of the more advanced and involved systems still focus on languages written in Latin script and often on English only. If software is capable of handling other scripts, though, or if software authors or users have gone to great lengths transliterating data into Latin script, a reliance can still be seen on dictionaries or other catalogues which only contain data in the most common or most standardized form of a given language. In the case of Arabic, Modern Standard Arabic. Moving further away from the mainstream of corpus-based linguistics and related fields, dialectal or colloquial speech or writing are but tiny backwaters.

In a similar vain, for all the work that has been done in Arabic dialectology, only comparatively little has been done on the subject of the Arabic dialect of Oman.

This article shall:

* describe the creation of a small and manageable corpus of transcribed and, where of interest, transliterated contemporary speech
* give an overview of Omani Arabic (OA)
* try to provide a modest look into some computer linguistic techniques and their usefulness in exploring the obtained corpus.

# The Dialect of Oman

Of the Gulf dialects, it might be argued that the most interesting is found in Oman. Far from being homogeneous in appearance[[3]](#footnote-3), it can be further subdivided into northern and southern variants as the two city centers of Masqaṭ (or Muscat in lexical English spelling) and Ṣalāla are a straight line of more than 800km of desert apart from each other. Additionally there are differences between the eleven governorates[[4]](#footnote-4) (indeed, asking an Omani to judge where another might be from, judging by her or his language, usually yields a governorate’s name); and even nowadays distinctions between neighbouring settlements are common[[5]](#footnote-5). However, to avoid oversimplification, there are many other (and sometimes stronger) influences on how someone speaks. For example, women may speak differently from men[[6]](#footnote-6) and children differently from their parents[[7]](#footnote-7), not to mention their grandparents. Apart from cardinal directions, the most important scholarly divisor is between speakers who consider themselves Bedouin (badawī) and those who consider themselves sedentary (ḥaḍarī)[[8]](#footnote-8). Differences in dialect due to tribal affiliation have been declined by the informants used for this study, so far as to say that they are entirely unknown to them[[9]](#footnote-9). As none of the speakers examined could be questioned about their heritage (and it is available in print for only two of them), no such information was considered. A much-discussed factor in the above literature, and also mentioned by informants, is the geography of Oman[[10]](#footnote-10), of which the northern and southern mountain ranges are notable, which, together with the country’s western (desert) and eastern (sea) borders serve as natural boundaries not only for people but also for their ways of speaking[[11]](#footnote-11).

Socio-economical processes of modernisation, such as the increase from 38% urban population in 1975 to 74% in 2013[[12]](#footnote-12), all forms of education, including access to radio and television[[13]](#footnote-13), and lately also the internet, have led to a certain homogenisation of language centring around Muscat[[14]](#footnote-14), the country’s capital in the north.[[15]](#footnote-15) Counteracting such homogenisation only somewhat are sociocultural forces, which work to keep differentiations alive as many families still seek for their children to marry within the tribe. The relevancy of tribal matters today could only be answered through a detailed study on the subject. Generally speaking, and (as far as the author is aware) not yet described as such, it seems that an informal standardisation process has begun in OA. If that should indeed be the case, then in keeping with the previous paragraph, such an accent-free dialect (from a native OA speaker’s point of view) would be spoken by educated Omanis more than by the uneducated, by the youth more than by older generations, and by people closer to Muscat more than by those further away from it. This study seeks to look at contemporary OA through a selection of postmillennial videos, analyses of their contents, and comparison to the canon of research thus far in order to provide an overview of the field and determine possible topics for future work.

# The State of Research into Omani Arabic

Apart from OA, a number of other languages are spoken in Oman (most notably, the South Arabian Mehri and Jibbāli[[16]](#footnote-16)). They are well outside the scope of this work, but it needs to be noted that at this current time there are minority language communities[[17]](#footnote-17) other than the majority who are native speakers of OA). Shaaban further mentions Swahili and Persian[[18]](#footnote-18).

The first European-language documentation of Arabic being spoken in Oman is understood as an interest in communicating with the Imam of Muscat out of trade interests in the second half of the 17th century – allegedly the Dutch trading post at Bandar Abbas was unable to translate a letter sent to them, an “inconvenience [...] soon remedied, however, by the arrival of Herbert de Jager [...], who was an [...] Orientalist”[[19]](#footnote-19).

More than 200 years then passed by, before a first dialectological article (Ueber den arabischen Dialekt von Zanzibar) appears in the journal of the DMG. Despite its title it contains explicit mention of Oman. The author notes: “Es wird mir versichert, die Sprache von Zanzibar sei genau dieselbe wie in Mesket [sic]”[[20]](#footnote-20) and proceeds to list similarities and differences in vocabulary, phonology and grammar between his findings in Zanzibar and other Arabic dialects, quite evidently Farsi (see above), but also Amharic, Ethiopic, Maltese, and Tigrinya.

Nine years later, The O’mánee Dialect of Arabic[[21]](#footnote-21) is published in the Journal of the Royal Asiatic Society. Correspondingly it quite clearly lacks actuality but is otherwise still useful as a historical reference. The first part constitutes a short study of grammar while the second part is a glossary contrasting “Standard Arabic” meanings with “Omanee” ones while also giving “Origins, Remarks [and] Examples”, with the last page including examples of “Children’s Language”.

The next item that requires mention is Sammlung arabischer Schriftstuecke aus Zanzibar und Oman, mit einem Glossar[[22]](#footnote-22), which, being on written language only, bears little relevancy to the present study but would be of great interest to anyone looking to compare the way Arabic was written before modern media.

Not long thereafter, Carl Reinhardt finished his comprehensive book on what he describes as “a dialect spoken in the valley of the Beny Xharus [sic]”[[23]](#footnote-23). Reinhardt knew of Jayakar’s articles, but found them to be flawed: “[... dessen] relativer Werth [sic] leider in Folge der mangel[haften] Transcription [sic] eine beträchtliche Einbusse [sic] erlitten hat.”[[24]](#footnote-24). His own book benefits not only from a more consistent, although difficult to read, transcription, but also from stringent structuring as well as the provided texts and stories. It seems likely the author had studied Caspari and Müller’s grammar as he certainly tried to be as complete in describing a dialect as they were in their description of CA. Also, a concern of Reinhardt’s was to work off a clean slate: “Um dieses zu erreichen hatte ich vorurtheilsfrei [sic] d.h. ohne Rücksicht auf die Schrift- und die anderen arabischen Mundarten ab ovo vorzugehen”[[25]](#footnote-25). Unfortunately, his work is clouded by age as well, and the small spread of his informants’ places of birth and residence stands contrary to his claim of describing an entire country’s language.

After a long period of silence in OA dialect studies, Sultan Qaboos’s (almost unbloody)[[26]](#footnote-26) succession of the throne in 1970 eventually enabled further work, beginning with The Phonology of Omani Arabic (Shaaban 1983); The Spoken Arabic of Khābūra on the Bāṭina of Oman which, unlike its title suggests, is largely a glossary of agricultural terms; The Morphophonology of Muscat Arabic (Glover 1988); Towards a Dialect Geography of Oman (Holes 1989) and Notes on the Dialect and Way of Life of the Āl Wahība Bedouin of Oman (Webster 1991). Holes then seems to have focused on Gulf Arabic in general, but a few years later there is an article on OA again, Retention and Loss of the Passive Verb in the Arabic Dialects of Northern Oman and Eastern Arabia (Holes 1998). After the turn of the millennium, there are two articles by an Omani Scholar, Broken Plurals in the Muscat Dialect of Omani Arabic (Al Aghbari and Urbanczyk 2004) and Derogatory Forms of Personal Names in Omani Arabic (Al Aghbari 2010). There also exists two non-scholarly dictionaries, Muʿǧam al-Mufradāt al-ʿĀmmiyya al-ʿUmāniyya (Ḥabsī 2007) and Qāmūs al-Faṣāḥa al-ʿUmāniyya: Mā ḥafaẓathu-d-dāriǧa al-ʿumāniyya li-l-luġa al-ʿarabiyya (al-Ǧāmiʿī 2005).

Other noteworthy articles include A Glimpse of Oman. Date Palm Terminology. A Case Study (Ghobashi 2008); The Role of the b-Prefix in Gulf Arabic Dialects as a Marker of Future, Intent and/or Irrealis (Persson 2008); Retention of the passive verb in a Bedouin dialect of northern Oman (Eades 2009); A Transitional Arabic Dialect of the Northern Omani Interior (Eades 2011) and Phonological Variation in Modern Standard Arabic. The Case of the Affricate /ʤ/: Oman as a Sample (al-Sammer 2010).

A concise overview of OA dialects is provided by Lutz Edzard in his *E*ncyclopedia of Arabic Language and Linguistics article, which serves as the primary source for the observations on dialectal features later in the present article. Edzard bases the entry in part on unpublished material by Clive Holes. Good detail is provided on phonology, morphophonology and morphology but, as in all other discussed works from the second half of the 20th century, no discussion of syntax is given. Perhaps it was felt that Jayakar’s and Reinhardt’s works do not need revision or, perhaps more likely, that not enough data are available.

In the way of textbook-like works, there are *E*lementary Omani Colloquial Arabic for English Speakers (The Sultan’s Armed Forces 1974) with several copies remaining in various libraries; An Introduction to the Spoken Arabic of Oman (Galloway 1977) of which only two copies seem to exist now[[27]](#footnote-27); and Spoken Arabic of Oman (Glover 1980) and Spoken Omani Arabic for Medical Personnel (Deyoung 1982) which might both be lost completely. More recent, but not for OA specifically, there is Gulf Arabic (Feghali 2008). In the German language, a tourist language guide exists (Omanisch-Arabisch: Wort für Wort, Walther 2013). The simplified transliteration and small size make it unfit for scholarly use but an accompanying audio CD with dialog produced by native speakers of OA is forthcoming and may be of some value. The call for “textbooks on an appropriate level”[[28]](#footnote-28), referring to university-level studies of Arabic dialects, has as yet to be fulfilled in the case of OA.

# Acquisition of the Corpus

Production of the corpus was done by recruiting two OA native speakers as informants. They are referred to simply as informant *A*[[29]](#footnote-29) and informant *B*[[30]](#footnote-30), respectively. More information on them and the process of transcription is given further below. In the beginning, it was intuitive to listen to a video together, taking missteps as indicators for how to modify the approach. Later, adjustments were made after discovering and studying the modus operandiof articles such as (the in many ways instructive) George Bush, Bedouin Poet[[31]](#footnote-31).

Pointed out by Reinhardt back in 1894, the preferred source of such material to be researched are sufficiently uneducated native speakers of the dialect, whose social network should be centered if possible on other speakers of that dialect. As a result of the changes the world has undergone during the past century, this ideal goal has become more difficult to attain. Especially in Oman, where the opening of the country following Sultan Qaboos’s accession to the throne in 1970 has caused an all-embracing use of automobiles, television sets and most of all, cell phones. Nonetheless, every visitor to the country trained in Modern Standard Arabic will quickly begin to recognize a speaker of sedentary[[32]](#footnote-32) OA by their accent and choice of words. As lack of finances and time precluded an expedition to Oman for controlled collection of data for the present study, a more modern approach was adopted and the study was based on the exploration, selection, transcription, and analysis of videos available on the video sharing platform *YouTube*. Searches for terms like al-lahǧa alʿumāniyya were conducted, which at first did not yield satisfactory results[[33]](#footnote-33). Due to this, contacts in Oman were asked via email to provide links to videos that in their view contained dialectal speech. Only one of the videos obtained this way (an extract of which is transcribed as text I) was considered appropriate material. Other suggestions were discarded for containing too much MSA or for having low quality voice recording. But through watching all received suggestions led to YouTube became attuned to the desired subject matter, and in time, all other videos used in this work were found[[34]](#footnote-34). The process of listening to the videos confirmed the fact that it would not be possible to limit the study to a single geographical location and that indeed a limitation to any other single determinant would be difficult as well. Instead, a matrix[[35]](#footnote-35) of contrastive determinants was created based on what the material contained, how it was described (the YouTube fields title and description) and what informants thought of it.

Transcription was performed as follows: at first, the author would sit with informant *A* and listen to the video sentence by sentence. Informant *A* would then dictate what he had heard and the author would write it down in his own Arabic handwriting as instructed. An attempt was made to transliterate, but this approach was found to be impractical and thus abandoned. Additionally and most interestingly, informant *A* always had a clear idea about how every dialectal word or dialectal variant of an MSA word heard in the videos should be written. In order to get an idea of whether there might be consensus among Omanis as to the orthography of their dialect, all transcripts were now emailed to informant *B* who was asked to make sure that the transcripts reflected the videos. No mention of an interest in orthography was made at this point. By herself, informant *B* corrected any orthographic mistakes she found in the transcripts, most of which were typos in MSA words.

In all cases where informant *B* corrected the spelling of a dialectal word or dialectal variant of a MSA word, there had been a prior discussion with or remark by informant *A* about it[[36]](#footnote-36). After transcription, or sometimes at the time of transcription, each text was carried over to German, decoding verbum pro verbo where necessary[[37]](#footnote-37) and/or sensible and translating sensum pro senso where possible. Any assumptions about the signification of a sentence, part of a sentence or word were confirmed with informant *A* either directly or through email. Information obtained thereby was worked into the German version of the text. Next, proper English translations were produced by

* looking at the Arabic transcription first
* jumping to the sentence in the video second and listening to it a couple of times
* taking the formerly written-down German representation of the text into account as a last measure.

These English translations were presented to informant *B* together with the Arabic transcripts she had already proofread at an earlier date. Sometimes informant *B* would make changes to the translations, in which case the signification of that section was discussed with informant *A* again, this time using Arabic instead of German as the language of conversation. If informant *B*’s changes were found to be improvements, they were worked into the translation. During these later discussions with informant *A*, when, although rarely, a third possible signification was discovered, a choice was made without talking to informant *B* again. After texts I-IV had been fully translated, work on text V started and proved difficult despite the short length of the text. Neither informant has a bedouin background and neither was able to understand text V with confidence and in sufficient detail. Therefore it was decided to skip texts V as well as VI (which is bedouin, too) and only translate one more text, number VII. Although the concept of text selection is partly undermined by this limitation, it also removes two out of six variables.

The remaining texts are:

1. Containing greetings and salutations: transcript[[38]](#footnote-38) of an excerpt from Al-Fīlm al-ʿUmānī – “Rāḥū aṭ-Ṭayyibīn” – Al-Ḫātim. The short film seems to have been produced by students of Sultan Qaboos University as part of their curriculum. In places very fast language is employed yet other passages are more readily understandable.
2. An acted situation between husband and wife: The transcript[[39]](#footnote-39) is an excerpt from Al-Musalsal al-Kūmidī – “Al-Ǧīrān” – ʿUnwān al-Ḥalaqa “Ḥubūbī”. Al-Ǧīrān was a TV series on Oman TV during the early 2000s. Even though the actors clearly speak dialect throughout, the language is mostly very clear.
3. A comical notion of strong dialectal expression: a transcript[[40]](#footnote-40) of an excerpt from Al-Amṯāl al-ʿUmāniyya. Informant *A* remarked that the speakers in this video obviously try to use colloquial language to attain a comical effect. Like the video transcribed as text I, this video was produced by students from Sultan Qaboos University. In general it is spoken more slowly, making it easier to understand.
4. An official speech: a transcript[[41]](#footnote-41) of Ḫiṭāb as-Sulṭān Qābūs fī Sīḥ al-Makārim bi-Ṣuḥḥār. Unlike the other four texts, the audio files linked to from this transcript required the removal of several parts of sentences which must have been erroneously edited in by either the uploader or the creator of the video, perhaps in an attempt to reconstruct lost or unclear parts of the original recording. They were, however, placed so unfortunately that they made it difficult to work with the text.
5. An example of colloquial storytelling: a transcript[[42]](#footnote-42) of an excerpt from Qālū Ḥāl as-Saʿūdiyyīn ʾiḏā Daḫaltū Bahlāʾ ʾaw Nizwā Iqraʾū at-Taʿwīḏa – Aš-Šayḫ Sālim an-Nuʿmānī. The speaker is originally from Birkāʾ in al-Bāṭina. Informant *B* said his language “is typical for a real Omani” of his age. An interesting feature of the text is the speaker’s quick code switching between casual and more formal registers as a device of storytelling.

It is recommended to listen[[43]](#footnote-43) to each line’s audio recording since the Arabic-letter transcript has been kept as close to MSA orthography as possible and is of course not able to give an accurate phonetic representation of the speech. In addition to the transcripts, a commented glossary[[44]](#footnote-44) is also available.

# Observations on Dialectal Features

Edzard (2009) lists six attributes common to all sedentary OA dialects, which discern them as a dialect group[[45]](#footnote-45). Of these, four could be observed in the corpus.

1. Retention of phonemes /ṯ/ (i.e. miṯl: speaker Ṭ), /ḏ/ (kiḏāk: speaker Ḫ), /ẓ/ (maẓharik: speaker Ḥ).
2. An infixation of in(n) between an active participle acting as a verb and a suffigated personal pronoun (i.e. nāqṣinnək: speaker Ṭ).
3. Absence of the gahawa syndrome[[46]](#footnote-46), where “a short vowel follows a morphemeinternal back spirant (X), whenever this X is preceded by *a*”[[47]](#footnote-47), i.e. qahwa>gahwa>gahawa (X being h in the eponymous example).
4. Frequent use of apophonic (meaning word-internal) passive forms (tuzhaq: speaker Q) instead of VIIth-stem “mediopassive”[[48]](#footnote-48) verbs. This feature is attested to by the complete lack of VIIth-stem verbs in the entire corpus.

The other two attributes (5: common use of feminine plural forms in all relevant word classes; 6: 2nd f. sg. suffigated pronoun -iš) can not be considered, as they are only present in videos V, VI, VIII, VI and X. The first two of these were however commonly heard during conversation between informants and members of their families[[49]](#footnote-49). Informant *B*, when asked to read (while hurriedly translating) an English text designed to evoke different conjugations in her dialect readily provided feminine plural forms of verbs galas and gāy[[50]](#footnote-50).

All phonemes of the *abǧad* are present in OA, which might be attributable to the formal nature of text VII. Transcription, where it is used, follows what was heard as closely as possible and is otherwise done by rules of the DMG. OA has some variation in short vowel sounds that can make it difficult to understand what vowel exactly has been said. When writing in Latin script, if for some reason required to do so, informants mostly used the proper MSA vowels (*a*, *i* and *u*). For more detailed information on phonetics of OA than given here, the reader is referred to Shaaban’s work mentioned earlier[[51]](#footnote-51). As for the present corpus, a list of phonemes[[52]](#footnote-52) with remarks on deviations from CA/MSA found within can be seen next to the online resources previously mentioned.

Only discrete personal pronouns for *he*, *you* (m.), *I* and *we* as well as suffigated personal pronouns for *he/his, you/your, I/mine, we/us/ours, they/them/theirs* are contained in the corpus. Those lists were rounded out by directly asking informants *A* and *B* about “*aḍ-ḍamāʾir al-munfaṣila wa-l-muttaṣila fi-l-ʿumāniyya*”. Feminine plural forms were produced without the interviewer mentioning them, while dual forms were not. The result (see table 1) is mostly congruent with Edzard’s[[53]](#footnote-53), with the difference of speaker Ḥ using f. poss. pron. -ik as he is not from Āl-Wahība[[54]](#footnote-54). Suffigation of a discrete personal pronoun at the end of a sentence (tarā, aqūl lək anā: speaker Š), perhaps as a particle of intensification, is a common feature[[55]](#footnote-55). Speaker N assimilates verb+li+p.p. (i.e. qultluh, “I said to him”), a feature called *indirect object suffix* by Edzard.

|  |  |  |
| --- | --- | --- |
| 3rd m. sg. | huwə, huwwəh | -uh |
| 3rd f. sg. | hiyə, hiyyəh | -hā |
| 3rd m. pl. | hum | -hum, -həm |
| 3rd f. pl. | hinn | -hinn |
| 2nd m. sg. | (ʾ)ənta, ntəh | -ak, -ək |
| 2nd f. sg. | (ʾ)ənti, ntī | -ik, -ək, -iš |
| 2nd m. pl. | (ʾ)əntū, ntuh | -kum |
| 2nd f. pl. | (ʾi)ntan | -kin |
| 1st sg. | ʾana | -nī (v.), -ī (n.) |
| 1st pl. | ḥnūh, ʾiḥnā | -nā |

Table 1: List of personal pronouns

Proximal demonstrative pronouns are ḏī*/*hāḏī (f.) and ḏā*/*hāḏā (m.). No distal demonstrative pronouns occur in the corpus. The only relative pronouns are mā and (ə)llī, except for one case of higher-register speech by speaker Q where it is MSA allaḏī. Additionally, yawm is used as a relative pronoun (mā saʿadtuh yawm ḥtāg l-nā, “I didn’t help him when he needed us”: speaker *L*). Various interrogative pronouns are used, mostly mū “what?”, but also ḥmū “why?” (and one occurrence of lēš which the informants insist is not OA), heyn “where?”, min heyn “from where/how?”, keyf “how?”

Three adverbs of time (taw “now”, marra wāḥida “one time/once”, marratan ʾuḫra “once more/another time”) were observed; one of place (hunā “here”); those of manner were more plentiful (bas “just/only”, kiḏāk “like (this)”, zeyn “super/okay/well” and ribāʿa “together”).

The negative particle throughout the corpus is usually mā. In higher registers, lā is used as well, and generally as a negative imperative. Together with the conjuctive particle wa it can be used as an inversion (wa-lā ənta tīs*?*, “or are you a goat”: speaker B); it can serve as a question marker at the end of a sentence as well (ḥalīb ət-tīs, lā?: speaker Š). Prepositions in the corpus are as in MSA: li “for, to”; *bi* “in, with, by way of”; fī “in”; ʾila “towards”; ʿalā “on, onto”; min “since, of, from”; ʿan “about”; ʿind “with”, but also to signify possession of material goods or features of character, etc.; maʿa “with”; beyn “between”. The most-used conjuction (indeed, the second most-used lemma) is wa “and”, which can also be realised as u. Further, fa “since, so, and”; aw “or”; bas “but”; inn “that”; kann “as though”; alšān (sic) “in order to”; miṯl mā “like”; iḏa “if” are present. The vocational particle is yā and is not only used when calling someone by name, but in contexts that are more abstract as well (“yā ahla wa-sahla*”*: speaker Ḫ).

Only a small number of nouns ending in -iyya were found, such as sulṭāniyya (speaker Q), which runs counter to the pattern being described as “highly productive”[[56]](#footnote-56). The texts are devoid of unusual formation patterns for unsound plurals. Sound plurals (taʿlīmāt: speaker Q, murtabkiyīn: speaker N) and unsound patterns of formation known from MSA (aḫbār: speaker M, agzā: speaker B, šarāyḥ: speaker Q) are used. No diminutives were present.

With regard to verbs[[57]](#footnote-57), it is notable that in OA, female plural forms are still in use according to Edzard. That none are found in the corpus is intrinsic to the texts’ topics and speakers. For both māḍī (perfect) and muḍāriʿu (imperfect), there is a tendency to replace the vowel following R1 with i or a vowel approaching its colour. This observation does not hold true for derived forms though, which mostly occur in more formal speech (natakātif “we stand together”: speaker Q) with the exception of form II verbs, which are plentiful in lower registers and are characterised by a CCaCCv(C) pattern in the māḍī as opposed to CuCaCCv(C) in MSA. Text I contains a noteworthy example of a quadriliteral verb (nsōlif “we chat”: speaker A) which falls into a group of common quadriliteral verbs in OA.[[58]](#footnote-58)

Morphophonological features including syllable types, consonant clusters, rules of stress, assimilation, or dissimilation are outside the scope of this article. Interested readers are pointed towards Shaaban (1983) and Glover (1988) who provide ample discussions of these matters.

# Word Frequency Analysis

An interesting first computer linguistical approach to a text is word frequency analysis. The simplest forms of it are sometimes supported even by common word processing applications. But especially where English texts are concerned, WFA can become quite sophisticated and is beyond the scope of this article. With Arabic-script corpora, the first consideration needs to be that any software used to analyse it should be capable of processing Unicode-encoded files. First approximations of a corpus’s word frequencies do not need to be algorithmically complex, and so custom scripts written in Ruby[[59]](#footnote-59) are presented instead of relying on pre-existing software. First (a) the human-readable form of the transcripts is extracted into a machine-readable corpus without extraneous information, then (b) that corpus is normalized[[60]](#footnote-60) until the point where it can be counted well enough to give meaningful results.

Towards this end, experiments were conducted with the original *Buckwalter Arabic Morphological Analyzer*[[61]](#footnote-61) written in *Perl* and then with *AraMorph.NET*[[62]](#footnote-62), an improved rewrite of the *Java* program *AraMorph*[[63]](#footnote-63), which in turn seems to be based on Buckwalter’s original work. Finally, efforts were made to utilize *MADAMIRA*[[64]](#footnote-64), one of the current self-described[[65]](#footnote-65) “state-of-the-art” projects in automated PoS tagging, tokenisation and lemmatisation of MSA and Egyptian Arabic. In the end it was decided that the small size of this study’s text corpus lent itself well to manual tokenisation and lemmatisation, so as a last step before word counting, hand-crafted replacement tables[[66]](#footnote-66) are read and applied by the Ruby script. The result is a word frequency list. Step (b), the count\_words.rb[[67]](#footnote-67) script, is explained in more detail by figure 1.

When processed as described above, the five texts comprise a corpus of ∑w = 1474 words. A look at the script and both replacement tables shows that each particle is counted as a word, but suffigated pronouns are dropped entirely. Normal pronouns are not treated in a special manner and count for one word each as well. The script returns a total of ∑dw = 459 discrete words, meaning that statistically, each word exists ∑w/∑dw = 3.21 times in the corpus. The mean word length is ∑chars/∑dw  = 3.09 ʾabǧad letters. When comparing the top 30 words[[68]](#footnote-68) to Buckwalter’s 1999, 2001 and 2002 counts of large MSA corpora as well as the New General Service List (Browne, Culligan, Phillips 2003) as an English-language control, there are quite a number of similarities despite differences in sample size as well as method and degree of tokenisation/lemmatisation.

flowchart.emf  
Figure 1: Flowchart for count\_words.rb

Figure 2 gives an impression of how word frequency counts are distributed over their range- As a final remark regarding validity of the achieved word counts, it is pointed out that various methods of dispersion calculation were considered. Given the small scope of this study, the limited time available, and the relatively high additional effort in required programming work, no such methods were implemented. Judging by work like Brysbaert and New’s[[69]](#footnote-69), word counting efforts of Arabic corpora will remain far from approaching the level of sophistication present in the discussion of the topic (let alone availability of algorithms or implementations thereof) for English and other Indo-European (or perhaps even Germanic only) languages. One important lesson seems to be that, for English, corpora sizes beyond 16 million words[[70]](#footnote-70) decrease the necessity for weighting, so it would in turn be interesting to see what effect a weighting and dispersion calculation according to the DP measure described and suggested by Gries[[71]](#footnote-71) would have on the present corpus given its small size.

drawing.emf  
Figure 2: Distribution of word frequencies

Of the verbs, which hold special significance in Arabic, there are ∑v = 111 distinct ones in the corpus. Together with the participles (there are surprisingly little given the spoken and dialectal nature of the corpus, only ∑pp = 5 passive and ∑ap = 7 active ones), they make up the data for figure 3. In alignment with a low number of participles, no form VII verbs were found. The absence of form IX is easily explained by the corpus’s topics, which required none of the speakers to mention either colours or ailments.

No distinct word frequency tables were produced per-text at this point, as both stylometric analysis as well as topic modelling are much better suited for the comparison of texts. Even so, taking another look at the top 30 list already reveals some interesting information, such as the importance of the single definitive article al*,* and, very much at the top together with it, particles wa, bi and mā, and preposition ʿala*,* even before the first verb qāla, there is allāh*,* which is in line with what one would expect of spoken language in the social and cultural environment of Oman. Furthermore, verbs feature prominently as they are inclined to do in Arabic, with six of them part of the top 30 words.

# Stylometric Analysis

drawing.emf  
Figure 3: Distribution of stem forms

The first stylometric analysis is said to be Lorenzo Valla’s 1439 work De falso credita et ementita Constantini donatione where the so-called Constantine Donation is proven to be a forgery[[72]](#footnote-72). Modern Stylometry is still mostly used for the purpose of identifying forgeries and is a logical follow-up to WFA, given that it may use the latter as one of its preliminary steps. This is the case for some of the methods employable by way of the *stylo* package available for the statistical programming language *R.* Unlike Ruby, R does not guarantee proper handling of strings encoded in Unicode, but newer versions will generally do so when running under Unix-based operating systems[[73]](#footnote-73). The online encyclopedia *Wikipedia's* article on stylometry[[74]](#footnote-74) lists one more software package[[75]](#footnote-75) available both free-as-in-speech and free-as-in-beverages, not tailored to a specific text language, and not solely intended for author attribution work. Unfortunately, the software completely ignores any Arabic letters.

Going by the *stylo* package’s documentation, we find out that it is possible to feed an arbitrary number of arbitrary pieces of text into the *stylo()* function. Then we get as a result, among other things, a table listing the *distance* (or dissimilarity) of each of these texts with each of the others (and with itself, which is of course zero). That way, it should be possible to compare the OA corpus’s texts to a sample of Moroccan Arabic[[76]](#footnote-76) as well as the 65th and 66th surahs, Sūrat-ṭ-ṬalāqandSūrat-t-Taḥrīm. At 293 and 258 words respectively they are close in size to the OA corpus text sizes (255 words per text on average) and, being Classical Arabic texts, should be sufficiently different to show up as clear outliers.

A distance matrix can be visualized as a network diagram with *nodes* (each text being a node) and *edges* (the lines connecting the nodes with one another). Luckily, another *R* package called *igraph* offers a collection of different algorithms for a calculation of both node placement and edge lengths for a given distance matrix. We decide for the Kamada-Kawai algorithm, which uses a physical model of springs[[77]](#footnote-77), works with undirected graphs such as ours, and reproducibly always gives the same image for the same input data. Running a script[[78]](#footnote-78) using both *stylo* and *igraph* results in figure 4, which should be taken with a healthy pinch of salt: computer-based Stylometry is a statistical method and the texts are so small that they do not provide sufficient data to work with nor produce truly meaningful numbers. In addition, if accu­racy/correctness were a real consideration, the texts would require as much as possible of the sort of clean-up[[79]](#footnote-79) discussed earlier, i.e. more thorough tokenisation and some degree of lemmatisation, followed by the removal of common function words. Even so, the script runs without error, showing that both *R* and *stylo* are capable of processing[[80]](#footnote-80) non-Latin-script texts. The fact that the diagram is interesting to look at can be considered as the proverbial icing on the cake. We first note an unexpectedly short and strong connection (both equalling small stylistic distance) between Text II (OA2) and the Moroccan Arabic control (MA1), for which reason the script was enhanced to make the size of the circles representing each node a function of the file sizes of the corresponding texts. Indeed, Text II is the smallest text, which may be a contributing factor to its closeness to the Moroccan text. Another factor could be that both texts are dialogues between two people and contain a relatively large number of personal pronouns.

  
Figure 4: Network diagram of stylometric distances between texts

On the upside, we further note the expectedly small distance and strong link between the two surahs of the Quran. The clustering of all OA texts towards the diagrams’ southern and western edges is welcome but might be coincidental, whereas the long distance and weak link between Texts I and IV could indeed be caused by Text I being a conversation between multiple persons producing relatively many and short sentences with a relatively small vocabulary and Text IV being a speech given by a single person producing relatively few and long sentences with a relatively large vocabulary. This would be consistent with Text IV being much closer to the CA texts – Sultan Qaboos tends to utilize a readily understandable Arabic close to MSA and with few OA dialectal words in his speeches, mostly being recognizable as a native speaker of OA (and to other native speakers of OA as being from the Dhofar region) by his accent.

# Topic Modelling

In contrast to stylometric analysis, topic modelling is a very new method, which was first described in 1998 as *(probabilistic) Latent Semantic Indexing*[[81]](#footnote-81)and had developed into *Latent Dirichlet Allocation,* a generalisation of *pLSI*, by 2003[[82]](#footnote-82).

LDA is founded on two facts: the first being that topics tend to be described using combinations of words that are not used to describe other topics.[[83]](#footnote-83) A medical text, say, on open heart surgery might contain the *word* “heart” quite often, but seldom will it occur in the proximity of words such as “flower”, “beloved” or “my”. Furthermore, within a text on open heart surgery, words like “intracardiac” or “bloodless” could appear in proximity to “heart”, as well as words like “bypass”, “stabilised” and “harvesting”. The second being that *documents* (which are defined as a “sequence of N words”[[84]](#footnote-84), and may thus be as small as one sentence) speak only about a small number of topics[[85]](#footnote-85). Second, sticking with the hypothetical text on open heart surgery, a sentence (or document) containing “heart”, “intracardiac” and “bloodless” might be less likely to also include “bypass”, “stabilised” and “harvesting”. Now running such a text (or *corpus* of documents) through LDA will of course not give the algorithm any sort of understanding of open heart surgery, but it has a good chance of allowing a user with knowledge of the history of medicine to conclude that one thing (or *topic*) might have to do with working on an open heart that has been stopped (one method[[86]](#footnote-86)) and that another topic might have to do with working on a beating heart (another method[[87]](#footnote-87)).

Going back to Arabic, the same problems that LDA has to deal with in English still hold true, such as the existence of polysemy (one word with multiple meanings) and synonymy (many words that all have the same meaning). Similarly with Word Frequency and Stylometric Analysis, the higher occurrence of agglutination in Arabic when compared to English, not to speak of phenomena such as inner plurals all pose a problem. Our OA corpus, having been tokenised and lemmatised, still does not stand a chance of producing truly meaningful results through topic modelling, again due to its small size. It may, however, yield understandable results, which were obtained by using R’s *stm* package, which, while building on LDA and others[[88]](#footnote-88), adds additional capabilities as well as ease-of-use. As it produces lists of words with weights attached to them, so-called word clouds were chosen as the method of visualisation, making use of R’s *wordcloud* package.

drawing.emf  
Figure 5: Threshold graphs showing topics vs. culling numbers

Following the *stm* manual[[89]](#footnote-89), we first obtain Figure 5 which is a collection of removal threshold graphs that give an idea of how the number of topics that are to be extracted influences the culling of documents, words and tokens. After some experimentation, we settle on 6 topics. Running the script[[90]](#footnote-90) now produces Figure 6, the Expected Topic Proportions graph, which consists of a triplet of words per topic, the first[[91]](#footnote-91) of which has the greatest weight and should therefore bear the greatest significance with respect to the corpus. This seems to hold true for topics 1, 3, 4 and 6. When working with bigger corpora and consequently extracting more topics, this graph can serve as a guide in deciding what topics to look into more closely.

Topic 1, shown in Figure 7, is the strongest and quite sensibly so: the corpus in its entirety consists of people talking either about themselves or about what other people have said or are (supposed to be) saying. The words grouped together with qāla are not easily explainable, especially kāna, which never appears together in a sentence with qāla. This reminds us that *stm,* and indeed other topic modelling packages , too, ignore whitespace[[92]](#footnote-92) and punctuation.

drawing-1.emf  
Figure 6: Expected Topic Proportions for stm topic model

|  |  |
| --- | --- |
| drawing.emfFigure 7: Topic 1 from the stm topic model | drawing.emfFigure 8: Topic 2 from the stm topic model |
| drawing.emfFigure 9: Topic 3 from the stm topic model | drawing.emfFigure 10: Topic 4 from the stm topic model |
| drawing.emfFigure 11: Topic 5 from the stm topic model | drawing.emf  Figure 12: Topic 6 from the stm topic model |

The second most-often used verb after qāla is kāna (figure 8), which makes a lot of sense because a oft-used syntactical structure in OA is kāna followed by an active participle. For it to be accompanied by ʾana seems logical, as the texts contain many sentences in the 1st person’s singular. Befitting Arabic’s byname of al-luġat-u-llāh, the Language of God, but more importantly befitting the corpus’s contents, where the single most-used phrase is al-ḥamdu li-llah (“Praise be to God”), there are not one but two topics (3 and 5, figures 9, 11) centered around allāh. There is not much to be said about the words appearing together with them; it seems that they just happen to be there which is something the algorithm can of course have no understanding of. The one exception are the occurances of ḥamada “to praise” in topic 3 and šāʾa “to will” in topic 5 respectively, which are parts of fixed phrases together with allāh and were correctly recognized as belonging together with it. It is also worthy to point out the appearance of yaʿnī (“it means”, “I mean”, or just a general interjection such as “urm”) in topic 5, as it had already been identified as an important word with regards to the corpus by WFA. It is found in close proximity to allāh in text III, thus perhaps placing it in topics 3 and 5 instead of one of the others. Topic 6 (figure 12) is quite self-explanatory as all texts are full of personal pronouns in singular form, therefore in a way the image constitutes a true statement. The most puzzling of the extracted topics is indeed number 4, depicted in figure 10: It could be seen as a reflection of the importance of verbs in spoken OA, as it shows a tendency of some of the speakers to agree rather than disagree (hēh being the common way to say “yes” in OA) or simply as an indicator of what happens when a statistical method is not fed a sufficient amount of data. Looking at the topics and their visual representations as a whole, we may feel content with mostly seeing forms of truth about what is found in the corpus as long as we remember that topic modelling is designed to be applied to much bigger corpora.

# Beyond Now

As regards to our knowledge of Omani Arabic, a number of topics need re-visiting to account for decades, and in some cases, more than a century, of change. There are also topics that have not been dealt with in depth at all. Shaaban wrote in 1983 that “[...] to have a better understanding of OA and its place in relation to [MSA] and other dialects of Arabic, and in relation to Omani society, more research is needed.”[[93]](#footnote-93) More specifically, he recommended studying the areas of syntax and semantics, child language and sociolinguistics. These recommendations are still valid today. Given the beauty of the country and the openness and hospitality of its people, further work, both in the field and through digital material already available, needs to be undertaken. Considering the current political climate in the Arabic world, it is possible that even quiet and peaceful Oman will undergo significant change in the near future or that its dialectal richness might further decline through processes of globalisation. There is evidence[[94]](#footnote-94) of the usefulness of using dialectological data with computer-aided methods as well as vice versa. In the past, projects such as the widely used *Modern Standard Arabic*[[95]](#footnote-95)textbook's project have made, at times pioneering, use of digital technologies not only during their creation but also as part of their operation. It can only be hoped that as our technical capabilities develop, so will our willingness and verve to use them for good.

1. *Jones, S. E.*, The Emergence of the Digital Humanities.  
   New York, NY; London 2014, p. 3, 29ff. [↑](#footnote-ref-1)
2. Just one example is the Voyant Tools (<http://voyant-tools.org>). [↑](#footnote-ref-2)
3. *Holes, C.*, Towards a Dialect Geography of Oman.  
   In: Bulletin of the School of Oriental and African Studies, 3(1989)52. [↑](#footnote-ref-3)
4. By name, *ad-Dāḫiliyya*, *aẓ-Ẓāhira*, *al-Bāṭina aš-šamāliyya*, *al-Bāṭina al-Ganūbiyya*, *al-Buraymi, al-Wusṭa*, *aš-Šarqiyya aš-Šamāliyya*, *aš-Šarqiyya al-ganūbiyya*, *Ẓufār*, *Masqaṭ*, *Musandam*. [↑](#footnote-ref-4)
5. i.e. *Eades, D.*, A Transitional Arabic Dialect of the Northern Omani Interior.  
   In: Zeitschrift für arabische Linguistik, 54(2011). [↑](#footnote-ref-5)
6. *Shaaban, K. A.*, The Phonology of Omani Arabic. Ann Arbor; London 1983, p. 10f. [↑](#footnote-ref-6)
7. *Shaaban, K. A.*, ibid., p. 16f. [↑](#footnote-ref-7)
8. *Eades, D.*, A Transitional Arabic Dialect of the Northern Omani Interior. In: Zeitschrift für arabische Linguistik, 54(2011). *Fischer, W. and Behnstedt, P.*, Handbuch der arabischen Dialekte. Wiesbaden 1980. *Holes, C.*, Towards a Dialect Geography of Oman. In: Bulletin of the School of Oriental and African Studies, 3(1989)52. *Holes, C. et al.*, Ingham of Arabia. Leiden 2013. *Wer, E. al- et al.*, Arabic Dialectology. Leiden; Boston 2009. [↑](#footnote-ref-8)
9. Also cf. *Valeri, M.*, Domesticating Local Elites. Sheikhs, Walis and State-Building Under Sultan Qaboos. In: Wippel (ed.), Regionalizing Oman, Dordrecht 2013, p. 268. [↑](#footnote-ref-9)
10. <https://sixtyfive.github.io/spoken-arabic-of-oman/map.html>. [↑](#footnote-ref-10)
11. Art. Omani Arabic (*Edzard, L.*).  
    In: Encyclopedia of Arabic Language and Linguistics, 2009, p. 479. [↑](#footnote-ref-11)
12. Data taken from United Nations’ World Population Policies 2013 report. [↑](#footnote-ref-12)
13. *Shaaban, K. A.*, The Phonology of Omani Arabic. Ann Arbor; London 1983, p. 15f. [↑](#footnote-ref-13)
14. *Shaaban, K. A.*, ibid., p. 15f. [↑](#footnote-ref-14)
15. A similar effect might exist with regard to southern Oman and the second biggest city, Ṣalālah, but no sources could be found to support such a theory [↑](#footnote-ref-15)
16. Art. South Arabian, Modern. In: Encyclopedia of Arabic Language and Linguistics, 2009. [↑](#footnote-ref-16)
17. Edzard also names the Baluchis, the Šiḥūḥ tribes and the Ḥarsūsī as well as guest workers, most of whom are from India, Pakistan and Bangladesh Art (Omani Arabic (*Edzard, L.*).  
    In: Encyclopedia of Arabic Language and Linguistics, 2009). [↑](#footnote-ref-17)
18. *Shaaban, K. A.*, The Phonology of Omani Arabic. Ann Arbor; London 1983, p. 15f. [↑](#footnote-ref-18)
19. *Floor, W.*, First Contacts between the Netherlands and Masqat or A Report on the Discovery of the Coast of Oman in 1666. In: Zeitschrift der Deutschen Morgenländischen Gesellschaft, (1982)132, p. 295. [↑](#footnote-ref-19)
20. *Praetorius, F.*, “Ueber den arabischen Dialekt von Zanzibar.”  
    In: Zeitschrift der Deutschen Morgenländischen Gesellschaft, (1880)34. [↑](#footnote-ref-20)
21. The O’mánee [sic] Dialect of Arabic.  
    In: Journal of the Royal Asiatic Society of Great Britain and Ireland, 3(1889)21. [↑](#footnote-ref-21)
22. “Sammlung arabischer Schriftstuecke aus Zanzibar und Oman.  
    Mit einem Glossar.” Stuttgart 1892. [↑](#footnote-ref-22)
23. *Reinhardt, C.*, “Ein arabischer Dialekt, gesprochen in ʿOmān und Zanzibar. Nach praktischen Gesichtspunkten für das Seminar für orientalische Sprachen in Berlin.” Amsterdam 1894, p. VII. [↑](#footnote-ref-23)
24. *Reinhardt, C.*, ibid., p. VIII. [↑](#footnote-ref-24)
25. *Reinhardt, C.*, ibid., p. VIII. [↑](#footnote-ref-25)
26. *Allen, C. H. and Rigsbee, W. L.*, “Oman under Qaboos. From coup to constitution.  
    1970-1996. ” London; Portland 2002. [↑](#footnote-ref-26)
27. One in the possession of the author, the other available at Exeter University Library. [↑](#footnote-ref-27)
28. Art. Dialectology (*Behnstedt, P. and Woidich, M.*).  
    In: Encyclopedia of Arabic Language and Linguistics, 2013, p. 314f. [↑](#footnote-ref-28)
29. A trained linguist born and raised in Izki, male. [↑](#footnote-ref-29)
30. A physician from Muscat, female. [↑](#footnote-ref-30)
31. Holes and Abu Athera 2007. [↑](#footnote-ref-31)
32. It is not always clear why someone would describe him- or herself as sedentary or as bedouin. In many cases, there is clear family history, but the author has also experienced cases where everybody agrees to be sedentary while one or more sons have declared their branch of the family bedouin. [↑](#footnote-ref-32)
33. It should be noted that *YouTube* performs an analysis of a user’s search requests and views over time and adjusts the results based on what its algorithms determine as the user’s interests. A helpful function after a while but not when first looking at a new subject. [↑](#footnote-ref-33)
34. A list of videos partially watched and taken into consideration for inclusion into the corpus is available at [https://www.you‌tube.com/play‌list?list=PL01I9fA2‌nm7‌Ersy7‌DB67L7YD1‌ig‌Nf0X‌sx](https://www.youtube.com/playlist?list=PL01I9fA2nm7Ersy7DB67L7YD1igNf0Xsx). The final selection before beginning transcription of the texts is available at: [https://‌www.you‌tube.com/play‌list?list=PL01I9fA2‌nm7E4x13‌dDOQXlEPAZX‌uVr4X1](https://www.youtube.com/playlist?list=PL01I9fA2nm7E4x13dDOQXlEPAZXuVr4X1). All weblinks were last accessed in October 2017. [↑](#footnote-ref-34)
35. <https://sixtyfive.github.io/spoken-arabic-of-oman/properties.html>. [↑](#footnote-ref-35)
36. This suggests that there may indeed be unwritten rules about the orthography of dialectal OA, which seems to be supported by Moritz. [↑](#footnote-ref-36)
37. Wherever decoding seemed required as an intermediary step towards understanding a sentence, it was simply done ad hoc and without much methodology. The only goal was to be able to form a cognitive representation in the author’s mind of the sen­tence’s meaning. [↑](#footnote-ref-37)
38. <https://sixtyfive.github.io/spoken-arabic-of-oman/transcripts/text01.html>. Speakers identified as Ḫ (Ḫalfān, who might not be very bright himself, but certainly more down-to-earth than his friend Saʿīd), A (ʿAbdullah, the host) and S (Saʿīd, who always has something else on his mind). [↑](#footnote-ref-38)
39. <https://sixtyfive.github.io/spoken-arabic-of-oman/transcripts/text02.html>.  
    Only two speakers, Ḥ (Ḥamdān, the husband and *Ṭ*, Ṭība, the wife). [↑](#footnote-ref-39)
40. <https://sixtyfive.github.io/spoken-arabic-of-oman/transcripts/text03.html>. There are three speakers in this text. First, speaker D(*al-*Muqaddam, the presenter), of which informant A‘s opinion is that this could be a student trying to imitate the voice of an old man from *ad-*Dāḫiliyya or the inner Šarqiyya. Second and third, speakers Š (al-Muštarī*),* the actor portraying the potential buyer who might have grown up in *ad-*Dāḫiliyya and B (al-Bāʾiʿ*),* the salesman’s actor who might be from Smaylor Musanna)*.* [↑](#footnote-ref-40)
41. <https://sixtyfive.github.io/spoken-arabic-of-oman/transcripts/text04.html>. The sole speaker in this sample is speaker Q (Qābūs b. Saʿīd Āl Saʿīd, Oman‘s head of state at the time of this writing), who was born in Ṣalālah on November 18th, 1940 (Allen 2002) but is felt by both informants to not be speaking with an accent that would betray his place of birth. [↑](#footnote-ref-41)
42. [https://sixtyfive.github.io/spoken-arabic-of-oman/transcripts/text‌07.html](https://sixtyfive.github.io/spoken-arabic-of-oman/transcripts/text07.html). Another text with a single speaker, *N* (*Sālim an-Nuʿmānī*, according to informant B the current Sultan Qaboos University‘s imam, who, despite his educatedness, has clearly not forgotten colloquial speech). [↑](#footnote-ref-42)
43. Each transcript also links to the original video, or the used portion thereof, respectively, if visual context is desired. [↑](#footnote-ref-43)
44. <https://sixtyfive.github.io/spoken-arabic-of-oman/glossary.html>. [↑](#footnote-ref-44)
45. According to Edzard the same holds true for bedouin dialects of OA, except that they generally exhibit *gahawah* syndrome. [↑](#footnote-ref-45)
46. Speaker *N*’s speech has signs of *gahawa* syndrome as he says *wafəd* instead of *wafd*, but he is the only speaker in the corpus who readily exhibits the feature. [↑](#footnote-ref-46)
47. Art. Gahawa (*de Jong, R.*)  
    In: Encyclopedia of Arabic Language and Linguistics, 2009. [↑](#footnote-ref-47)
48. Art. Passive (*de Jong, R. and Edzard, L.)*In: Encyclopedia of Arabic Language and Linguistics, 2009. [↑](#footnote-ref-48)
49. Also, *Glover, B. C.*, The Morphophonology of Muscat Arabic. Los Angeles 1988, p. 199f. [↑](#footnote-ref-49)
50. Cf. <https://sixtyfive.github.io/spoken-arabic-of-oman/verbs.html>. [↑](#footnote-ref-50)
51. *Shaaban, K. A.*, The Phonology of Omani Arabic. Ann Arbor; London 1983. [↑](#footnote-ref-51)
52. <https://sixtyfive.github.io/spoken-arabic-of-oman/phonemes.html>. [↑](#footnote-ref-52)
53. Art. Omani Arabic (*Edzard, L.*)  
    In: Encyclopedia of Arabic Language and Linguistics, 2009. [↑](#footnote-ref-53)
54. *Edzard, L.*, ibid., p. 483. [↑](#footnote-ref-54)
55. *Edzard, L.*, ibid., p. 483. [↑](#footnote-ref-55)
56. Art. Omani Arabic (*Edzard, L.*).  
    In: Encyclopedia of Arabic Language and Linguistics, 2009. [↑](#footnote-ref-56)
57. A table of all sound form I verbs, as well as those of the doubly weak ones where R3 used to be a hamza in CA, is found in the supplemental materials at [https://sixty‌five.git‌hub.io/spo‌ken-ara‌bic-of-o‌man/verbs.html](https://sixtyfive.github.io/spoken-arabic-of-oman/verbs.html). [↑](#footnote-ref-57)
58. Art. Omani Arabic (*Edzard, L.*).  
    In: Encyclopedia of Arabic Language and Linguistics, 2009. [↑](#footnote-ref-58)
59. A programming language originally developed in Japan with strengths in coping with Uni­code texts (<https://www.ruby-lang.org>). [↑](#footnote-ref-59)
60. For pre-normalisation cleanup, Spence Green’s *basic\_ortho\_norm.py* Python script was used. It is available at [http://www.spence‌green.com/2011/01/19/how‌to-ba‌sic-ara‌bic-prepro‌cessing-for-nlp](http://www.spencegreen.com/2011/01/19/howto-basic-arabic-preprocessing-for-nlp) along with further reading into the subject. [↑](#footnote-ref-60)
61. <https://catalog.ldc.upenn.edu/LDC2002L49>. [↑](#footnote-ref-61)
62. *Sadam al-Azani et al.* 2013, <https://sourceforge.net/projects/aramorphnet>. It should be no­ted that the hosting website, SourceForge, limits access from ITAR-restricted countries. In such cases, an error message worded "download forbidden" is said to be shown. [↑](#footnote-ref-62)
63. Pierrick Brihaye, <https://sourceforge.net/projects/aramorph>. [↑](#footnote-ref-63)
64. Even though the web-based version ([https://camel.abudhabi.nyu.edu/ma‌damira](https://camel.abudhabi.nyu.edu/madamira)) is available for trial with small amounts of text, it is not suitable for working with larger amounts. Licenses are theoretically available for academic institutions, but one component requires separate licensing, which the author of this article was unable to obtain. [↑](#footnote-ref-64)
65. <http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf>. [↑](#footnote-ref-65)
66. [https://github.com/sixtyfive/spoken-arabic-of-oman/‌blob/‌master/‌wfa/‌lemma‌tisation‌\_replac‌e‌‌ment\_table.txt](https://github.com/sixtyfive/spoken-arabic-of-oman/blob/master/wfa/lemmatisation_replacement_table.txt); [https://git‌hub.com/‌sixty‌five/spo‌ken-ara‌bic-of-o‌man/‌blob/‌master/‌wfa/‌to‌k‌e‌n‌‌isation\_replacement\_table.txt](https://github.com/sixtyfive/spoken-arabic-of-oman/blob/master/wfa/tokenisation_replacement_table.txt). [↑](#footnote-ref-66)
67. <https://github.com/sixtyfive/spoken-arabic-of-oman/blob/master/wfa/count_words.rb>. [↑](#footnote-ref-67)
68. <https://sixtyfive.github.io/spoken-arabic-of-oman/top30words.html>. [↑](#footnote-ref-68)
69. *Brysbaert, M. and New, B.*, Moving Beyond Kučera and Francis. A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. In: Behavior Research Methods, 4(2009)41. [↑](#footnote-ref-69)
70. *Brysbaert, M. and New, B.*, ibid., p. 979. [↑](#footnote-ref-70)
71. *Gries, S. T.*, Dispersions and Adjusted Frequencies in Corpora.  
    In: International Journal of Corpus Linguistics, 4(2008)13. [↑](#footnote-ref-71)
72. Art. Constantine, Donation of. In: The New Encyclopaedia Britannica, 2010, p. 564. [↑](#footnote-ref-72)
73. Including distributions of *Linux* or *BSD* such as the popular *Ubuntu* and *MacOS* operating systems. [↑](#footnote-ref-73)
74. <https://en.wikipedia.org/wiki/Stylometry>. [↑](#footnote-ref-74)
75. Dr. Peter Millican’s *Signature,* <http://www.philocomp.net/humanities/signature.htm>. [↑](#footnote-ref-75)
76. Produced in the same manner as described earlier regarding the OA transcripts. [↑](#footnote-ref-76)
77. *Kamada, T. and Kawai, S.*, An Algorithm for Drawing General Undirected Graphs.  
    In: Information Processing Letters, (1989)31. [↑](#footnote-ref-77)
78. <https://github.com/sixtyfive/spoken-arabic-of-oman/blob/master/stylometry/stylo.r>. [↑](#footnote-ref-78)
79. The *stylo* package possesses built-in clean-up devices, but not, at this time, for Arabic. Only the small cleaning measure of removing punctuation characters such as full stop, comma, question mark, etc. is being performed by the script. [↑](#footnote-ref-79)
80. On the author’s Linux computer and likely on a Mac computer. It might be possible on a Windows computer as well, but would likely require additional work. [↑](#footnote-ref-80)
81. *Papadimitriou, C. H. et al.*, Latent Semantic Indexing: A Probabilistic Analysis, 1997. [↑](#footnote-ref-81)
82. *Blei, D. M. et al.*, Latent Dirichlet Allocation.  
    In: Journal of Machine Learning Research, 3(2003). [↑](#footnote-ref-82)
83. *Blei, D. M. et al.*, ibid. [↑](#footnote-ref-83)
84. *Blei, D. M. et al.*, ibid. [↑](#footnote-ref-84)
85. *Blei, D. M. et al.*, ibid. [↑](#footnote-ref-85)
86. *Weißer, C.,* Herzchirurgie. In: Enzyklopädie Medizingeschichte, 2005. [↑](#footnote-ref-86)
87. *Weißer, C*., ibid. [↑](#footnote-ref-87)
88. *Roberts, M. E. et al.*, stm: R Package for Structural Topic Models.  
    In: Journal of Statistical Software, IIVV. [↑](#footnote-ref-88)
89. *Roberts, M. E. et al.*, ibid. [↑](#footnote-ref-89)
90. <https://github.com/sixtyfive/spoken-arabic-of-oman/blob/master/tm/stm.r>. [↑](#footnote-ref-90)
91. It is prudent to not assume that Arabic words will always be displayed right-to-left, even if their letters clearly are. Luckily, the graph functions correctly and shows both words’ letters as well as words themselves from right-to-left, the one caveat being the commas between them, which are shaped left-to-right. [↑](#footnote-ref-91)
92. Whitespace is the term used for all space (printable and non-printable spaces, tabulator) and line-break (line-feed, return-carriage, line-feed-and-return-carriage) characters in plain text documents. [↑](#footnote-ref-92)
93. *Shaaban, K. A.*, The Phonology of Omani Arabic. Ann Arbor; London 1983. [↑](#footnote-ref-93)
94. *Embleton, S. et al.*, An Exploration into the Management of High Volumes of Complex Knowledge in the Social Sciences and Humanities. In: Journal of Quantitative Linguistics, 3(2004)11. [↑](#footnote-ref-94)
95. *Schulz, E. and Bouraima, B.*, Modernes Hocharabisch. Leipzig 2013. [↑](#footnote-ref-95)