Critical Multilingualism Studies
an interdisciplinary journal

Rasha Soliman
University of Leeds

Laila Familiar
New York University Abu Dhabi

CREATING A CEFR ARABIC VOCABULARY PROFILE: A FREQUENCY-BASED MULTI-DIALECTAL APPROACH

Abstract:
Measuring proficiency levels in second language (L2) teaching in Higher Education relies on certain international frameworks of proficiency levels such as the Common European Framework of Reference (CEFR) or the American Council on the Teaching of Foreign Languages (ACTFL) which, interchangeably, can easily be benchmarked. While these frameworks try to provide generic guidance, they encourage educators to create language-specific profiles for each language taught in their programmes. The CEFR guidance refers to these as the Reference Level Descriptors (RLDs) which list the linguistic items to be covered in each proficiency level including situational topics, grammatical rules and vocabulary lists. At the time of writing this paper, the CEFR website has published RLDs for 11 languages, not including Arabic.
In an effort to respond to this need for a profile for Arabic language, the authors of this paper have been working on a collaborative project that lists all vocabulary items to be taught in the first two CEFR levels (A1 and A2) as a starting point. These lists will aim to help teachers, curriculum designers, material writers and assessors to determine the language content for these levels. Although the CEFR is still based on a monolingual perception of L2 teaching and learning, its recent 2018 and 2020 volumes give attention to plurilingual competences, which, in the case of Arabic, would entail a variationist and a multidialectal view. The current project uses MSA as its base, but cross-checks among a range of Arabic dialects in order to provide a vocabulary profile that is more linguistically inclusive and that can help bridge between different registers. A number of selection and ranking criteria is used in creating this vocabulary profile; among them are multidialectal commonality, frequency of use, linguistic complexity and relevance to the CEFR descriptors.

Soliman, R. & Familiar, L. (2024). Creating a CEFR Arabic Vocabulary Profile: A frequency-based multi-dialectal approach. *Critical Multilingualism Studies,* 11(1) 266-286. ISSN 2325–2871.

This paper aims to discuss how a comparative multidialectal approach has an important role in the selection of vocabulary items to be learnt in each CEFR level in a way that would increase the learner's level of cross-dialectal comprehension, as well as their overall Arabic competency. The paper will present the methodology applied in this Arabic Vocabulary Profile (AVP) with examples drawn from multidialectal language use as a criterion for vocabulary benchmarking.

Keywords: Arabic ◆ Dialects ◆ CEFR ◆ Vocabulary ◆ Frequency

Introduction

Understanding more than one urban Arabic dialect is an essential skill for the learners of Arabic as a second language (L2) as it reflects the cross-dialectal comprehension abilities that the average educated Arabic speaker has (Abu-Melhim, 1992; Ezzat, 1974; Soliman, 2015) and which Arabic learners state as one of their main learning needs[1]. The skill of cross-dialectal comprehension equips learners with the confidence and the ability to successfully communicate with a range of Arabic speakers. Regardless of the variety/ies being focussed on in an Arabic programme, multidialectal awareness and comprehension are increasingly becoming crucial aspects of the students' learning journey as they encounter linguistic variation across the media and in their day-to-day interactions with Arabic speakers. In addition, learners come to the Arabic programmes with a diverse background knowledge with some of them being considered (semi)heritage learners who have already been exposed to a range of Arabic varieties. The last decade has witnessed a recognition of the importance of multidialectalism and a growing interest in research related to its integration in Arabic L2 teaching (Al-Batal, 2018; QFI, 2022; Soliman, 2015; 2023; Trentman & Shiri, 2020). When it comes to practical aspects of implementing a multi-dialectal approach in teaching, more research and guidance are needed to support teachers with the 'How to'. Without sufficient resources, guidance, and teacher education, the idea of a plurilingual approach to L2 teaching will remain abstract and face different levels of resistance to implement (Dooly & Vallejo, 2020).

The next section of this paper will briefly discuss the positions on linguistic variation in Arabic L2 teaching giving examples of some of the recent initiatives and resources that support the integration of multiple varieties in Arabic teaching. The paper will then discuss how linguistic

[1] On the topic of Arabic learning needs in relation to communication with different dialect speakers, please see: (Belnap, 2006; Husseinali, 2006; Khalil, 2011; Soliman, 2015)

variation has become an integral aspect of proficiency scales like the CEFR and how a multi-dialectal approach has been implemented in the current Arabic Vocabulary Profile project.

The Dilemma of the One and the Many

> "You cannot conceive the many without the one...The study of the unit is among those that lead the mind on and turn it to the vision of reality." Plato – a Greek Philosopher.

Research on diglossia has taken different perspectives over the decades, including theoretical work on definitions of diglossia (Francescato, 1986; Freeman, 1996), its position within sociolinguistics (Chelghoum, 2017; Schiffman, 1999) and its impact on L2 teaching (Ferguson, 1963; Giolfo & Sinatora, 2011). Arabic is one of the diglossic languages that has received particular interest from academics with some focussing on the dichotomous perspectives of MSA versus dialects while others grounding their research on the more variationist continuum of Badawi (1973) and the multi-dialectology of Arabic. In terms of how to deal with Arabic diglossia and variation in L2 teaching, one can say that the Integrated Approach (IA) developed by Munther Younes has been the most researched and applied approach in the last three decades (Younes, 1995; 2006; 2014). The IA is unique in how it broke with tradition as it changed teachers' and learners' monolingual perceptions of languages, how they are used and how they are taught and learnt. Such a change has not been easy and has required lots of efforts from academics who had to provide research-based evidence to promote the innovativeness of the IA. One recent example of seminal research into the effectiveness of the IA is Al-Batal's (2018) edited volume titled 'Arabic as One Language'. The volume highlighted the insufficiencies of the monolingual approach of teaching MSA only, referring to it as 'the Firewall Separation Vision of Arabic'. At the same time, the volume emphasised the importance of integrating dialects into Arabic programmes promoting the vision of 'Arabic as One' which Al-Batal (2018, p.7) describes saying:

> '[…] we propose here an alternative vision based on the belief that varieties of Arabic do not represent isolated entities but are part of one language system called "Arabic".'

While this holistic vision of how to teach a diglossic language like Arabic mirrors the reality of language use, and while it is true that the Arabic varieties are not isolated entities from each other, one cannot deny that each variety is described and prescribed in its own right. We need to name each of them, define them and make decisions related to which ones to teach in a language programme and which ones to prioritise. The quote above by Al-Batal may give an impression to educators that in teaching Arabic as L2, you randomly select the teaching content out of any varieties of Arabic without setting any boundaries between them. It is important to

understand that the quote is aimed at perceptions and ideologies rather than at the practicalities of Arabic L2 pedagogy. In the first chapter of the volume, Al-Batal presented six examples of American institutions that integrate dialectal Arabic in their programmes. All programmes teach MSA in a way or another and all ensure that learners are capable of using at least one dialect for various communicative tasks. Despite this dialectal integration, selection and prioritisation decisions are made by each of these institutions, as to whether focus on a particular dialect and then move onto MSA, or focus first on MSA and integrate a particular dialect from day one, in more or less equal proportions.

The successful models of dialect integration listed in the volume present a step in the right direction of integrating linguistic diversity in teaching Arabic. However, a further step ahead would be the integration and consideration of multiple dialects in Arabic L2 pedagogy. This would go beyond the duality of MSA versus one dialect to a more diverse view of Arabic as one and as many. This would relate to the ancient philosophical dilemma of 'The One and the Many' (Stokes, 1971) that looks at the unity and the multiplicity of everything around us trying to observe the links between the one entity and its variations recognising commonalities and differences. On a practical sense, this variationist approach in an Arabic programme means that in addition to equipping learners with sufficient knowledge of MSA and the confidence in using at least one dialect, familiarity with the linguistic similarities and differences across multiple Arabic varieties is to be embedded into the curriculum, the classroom practice, the learning materials and eventually proficiency assessments. As challenging as it may seem, the last decade has already witnessed initiatives that support the multi-dialectal vision such as 'We can learn Arabic' website[2] that incorporates aspects of MSA and dialectal elements which are common across some urban dialects, the Playaling website[3] providing listening materials in MSA and various dialects, the Arabic vs. Arabic publication comparing and drawing attention to shared linguistic features across 15 varieties (Aldrich, 2018), and the Khallina online modules that include different Arabic varieties in topical themes.[4] The project presented in this paper adds to these initiatives with the aim of providing vocabulary lists that are based on a variationist approach in word selection. Before describing the multi-dialectal approach of the project, the next section will present a brief introduction to CEFR proficiency model of content selection.

[2] https://www.wecanlearnarabic.com/
[3] https://playaling.com/
[4] https://khallina.org/

Word Lists and the CEFR Reference Level Descriptors

In the field of L2 education, vocabulary lists have been traditionally developed for different educational contexts (Matsuoka, 2012). They are considered an essential tool for practitioners when designing and developing curricula, as they help in deciding which lexis to include in the pedagogical materials and in what order (Kilgarriff et al., 2014; Leech et al., 2001; Nation, 2004). Word lists can be based on the intuition of educators or/and learners of what they see needed at a certain learning stage. They can also be purely thematic, or they can be based on frequency of use rather than themes. Some influential corpus-based frequency lists in English date back to Thorndike's work, The Teacher's Word Book (1921), and the subsequent list of 30,000 most frequent words by Thorndike & Lorge (1968). They were widely used in a variety of educational settings, and they paved the way for key projects that are specific to L2 teaching and learning, like the English Vocabulary Profile5.

For Arabic, efforts have been made ever since Moshe Brill and Jacob Landau elaborated their own lexical frequency lists, with the aim of helping learners of Arabic understand literary and media texts. In fact, Landau observed that a careful selection of the vocabulary used in the teaching of Arabic as L2 is necessary for the improvement of teaching materials, since it "may lead to greater efficiency in language teaching" (Landau, 1959). Multiple Arabic frequency lists were elaborated afterwards[6], but it was not until Buckwalter & Parkinson (2011) published their Frequency Dictionary that the field of TAFL finally realized the importance of employing modern corpus-based computational linguistics tools and methods in compiling world lists. One pioneering aspect in Buckwalter & Parkinson (2011 that we would like to highlight, is that, while previous Arabic frequency lists focused on MSA vocabulary, their dictionary included the most frequently used words in a variety of dialects. Lastly, we would like to mention Familiar's *Frequency Dictionary* (2021), which was developed with the specific purpose of creating a series of Graded Readers in Arabic. Because the resulting word list stems from a literary corpus that is exclusively composed of contemporary Arabic fiction, the nature of the lexicon included mirrors everyday life themes and topics.

Despite its importance and relevance to L2 teaching, frequency is only one criterion in creating a thorough vocabulary list that is benchmarked with proficiency scales like the CEFR. The CEFR has been one of the most commonly used proficiency scales in higher education globally

[5] https://www.englishprofile.org/
[6] For a thorough summary, please see the introduction in Buckwalter & Parkinson's dictionary (2011).

since its establishment in the early nineties[7]. University graduates who study a L2 as a major component of their degrees are expected to reach an advanced proficiency in that target language (MLA, 2009), and in some countries (such as in the UK), they are expected to achieve a C1/Advanced High level by the time they graduate (QAA, 2023). The CEFR's 'intentionally' generic description of proficiency skills is one of the reasons behind its popularity as it meant that the scale can be flexibly adopted to suit the needs of any language programme[8]. Nevertheless, the vagueness of the CEFR descriptors can be a double-edged sword that can give an impression of unsuitability for certain language learning contexts such as in teaching non-European languages, let alone a diglossic language. To combat the risk of possible limitations and misinterpretations of the CEFR, the Council of Europe encourages educators to develop what they refer to as Reference Level Descriptors (RLDs). They state:

> The […] CEFR is potentially applicable to all the languages taught in Europe and does not, therefore, relate to any specific one. However, […] language teachers have found its specifications to be insufficiently precise. *Reference Level Descriptions (RLDs) language by language* have therefore been drawn up to provide reference descriptions based on the CEFR for individual languages. These RLDs are made up of "words" of a language rather than general descriptors. Reference levels identify the forms of a given language (words, grammar and so on), mastery of which corresponds to the competences defined by the CEFR. […]. RLDs are not produced by the Council of Europe but by national teams using various approaches […]. Nevertheless, each one contains reference levels that are comparable to those in the CEFR.
>
> https://www.coe.int/web/common-european-framework-reference-languages/reference-level-descriptions

The need for language specific RLDs prompted the production of inventories for a number of languages. According to the Council of Europe website, and at the time of writing this paper,

7 If you are more familiar with the ACTFL proficiency scales, please see this link that compares the ACTFL and the CEFR levels: https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

8 See: https://rm.coe.int/1680459f97

RLDs have been produced for 11 languages, which do not include Arabic[9]. Some of these languages have RLDs that focus mainly on vocabulary such as the Oxford Learner's Word Lists which contains 5000 English words benchmarked with the CEFR[10] and the Portuguese Vocabulary Profile with approx. 2000 words[11] while other RLDs aimed at providing comprehensive descriptors such as the English Vocabulary and Grammar Profiles and the British Council – EAQUALS Core Inventory for General English which lists vocabulary, grammar, topics and functions[12]. Each project devised criteria for word and content selection and ranking. Oxford Word Lists are based on two criteria: frequency of use in multiple forms of English including British, American and World English; and the corpora of published learning materials which were then verified by two professors in applied linguistics who are experts in the field of English language teaching. The English Vocabulary Profile (EVP) utilised the Cambridge Learner Corpus and coursebook word lists in addition to seeking the validation from a wide group of teachers and users of the EVP online tool. The EAQUALS project for English language used the CEFR descriptors as a starting point supported by analysis of popular coursebooks and consensus from teachers through a survey. Most of the other available RLDs rely on a combination of learners' corpora, analysis of coursebooks, relevance to the CEFR generic descriptors, and the intuition of the teachers and the designers of the RLDs who tend to be known as experts in their perspective fields of L2 teaching and applied linguistics (Marello, 2012). It is difficult to identify the extent of how linguistic variation is (or is not) accommodated into the currently available RLDs. The EVP website provides vocabulary lists for British and American English separately while the Oxford Word Lists explicitly mention that their vocabulary inventory is based on several varieties of English. According to Marello (2012, p. 330), the German RLDs 'Profile Deutsch' includes Swiss and Austrian variants.

In the case of Arabic, the last two decades witnessed a growing interest in implementing the CEFR in Arabic L2 teaching; initially with the aim of developing standardised proficiency assessments, then with more focus on developing CEFR-based curricula. Most of these attempts

[9] See: https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions-rlds-developed-so-far

[10] https://www.oxfordlearnersdictionaries.com/wordlists/

[11] See: http://www.tufs.ac.jp/ts2/society/pvp/pvp.beta.html and (Torigoe, 2016)

[12] https://www.eaquals.org/resources/the-core-inventory-for-general-english/

relied solely on the intuition of small numbers of scholars in the field[13]. The authors of this paper are aware of two projects that aimed to benchmark Arabic vocabulary with the CEFR levels using more computational methods. The first is the KELLY project which aimed to benchmark vocabulary of nine languages including Arabic and relying on already available corpora of frequency lists for each language. Language comparisons and expert judgements were then utilised to give more validity to the created lists in terms of their relevance to L2 learning needs (Kilgarriff, et al., 2014). The second project is part of a current PhD study by Nouran Khallaf looking at computational simplification of Arabic texts by bringing the level of complexity of linguistic items from a certain proficiency level to a lower one within the CEFR scale. In order for the computer to carry out the simplification process, Khallaf designed a vocabulary list which relies purely on computational formulas that merged three corpora: the frequency list by Buckwalter and Parkinson (2011), the vocabulary lists in the popular textbook series '*Al-Kitaab*' and the list produced by the KELLY project (Khallaf & Sharoff, 2021). It is worth noting here that both projects relied heavily on computational tools in devising the lists and both had a monolingual approach as they eliminated lexical items that were considered dialectal. The limitations of these projects as well as the general aims of creating resources that encompass variation incited work on the Arabic Vocabulary Profile project described in the next sections.

The Arabic Vocabulary Profile (AVP) project

The idea of creating an Arabic vocabulary profile that is based on a combination of frequency, the CEFR descriptors, linguistic complexity and dialect commonality has been concurrently in the back of the minds of the project leaders: Laila Familiar with a vision of creating graded readers based on her literature-based frequency list (Familiar, 2021), Geri Atanassova with experience and an interest in aspects of extensive reading in the Arabic classroom, and Rasha Soliman as an academic with an interest in multi-dialectal variation and the application of the CEFR to Arabic L2 teaching (Soliman, 2018; 2023).

The project leaders' discussions, which started in 2020, led to envisioning an (AVP) that fills a gap in the current available teaching and learning resources by providing vocabulary lists that are thoroughly selected to match the CEFR descriptors. Similar to RLDs of other languages (as described above), the criteria of frequency and relevance to learning needs is integral to the vocabulary selection process. In addition, linguistic (phonological, morphological and

13 See (Soliman, 2018, pp. 122-123) for a brief review of recent work on the application of the CEFR to Arabic L2 teaching.

syntactic) complexity was another criterion that was applied in this project. Encompassing all these criteria is Arabic lexical variation considering MSA and several urban dialects that learners are likely to be exposed to. This is expected to be an ongoing project, that will need to continuously be reviewed by both Arabic education experts and learners, in order to keep abreast of developments in language use and needs. In the current stage of the project, vocabulary is benchmarked at the first two CEFR levels (A1 and A2) with the plan to expand to higher levels in the future.

Which Arabic variety is the AVP for? Is it for one or for many?

As discussed earlier in this paper, efforts in integrating variation into Arabic classrooms and resources do not mean random selection of linguistic items that belong to any particular Arabic variety. Decisions on which variety/ies to focus on need to be made in any Arabic pedagogical initiative to help clarify to the users and learners how to utilise the resource effectively according to the contexts of learning. The ensuing goal of creating graded readers out of the AVP entailed an MSA focussed profile. However, and as discussed in ample literature, MSA solely does not qualify as the suitable variety for the early stages of Arabic learning when basic communicative needs are conducted in dialects (Giolfo & Salvaggio, 2017). At the same time, opting for a dialectal form meant choosing a specific dialect, which again would make the AVP limited and might discourage teachers and learners from utilising the AVP if they teach a different dialect.

Another possibility was to create parallel lists for multiple Arabic varieties giving the user the option to choose a specific variety[14]. Although this would serve a large community of users, it would create unnecessary duplications with certain lexis, such as most prepositions, being almost identical across the Arabic varieties. In order to respond to this dilemma of the One and the Many, we decided to base the AVP on MSA as a form mostly neutral to geographical regions[15] but with a multidialectal approach to vocab selection that is not limited to certain regions but rather relies on the linguistic commonalities among the different Arabic forms. Within the continuum of spoken dialects, we utilised the Manchester Dialect Database

14 Once an initial profile is benchmarked with the CEFR, different parallel dialectal versions can easily be created using comparative computational tools such as the ones designed by CAMeL Lab at: https://nyuad.nyu.edu/en/research/faculty-labs-and-projects/computational-approaches-to-modeling-language-lab.html

15 We recognise that MSA is also variable across regions, but within lexis, it mostly manifests at higher proficiency levels (Van Mol, 2003).

(hereinafter, MDD)[16] and MADAR Lexicon[17] as well as other available resources/dictionaries on a range of dialects[18] to compare lexical items and prioritise the commonly shared ones.

What is a multidialectal approach to lexis selection?

When considering multidialectal linguistic variation in the design of the AVP, one of the first issues that comes up is the lexical overlap that exists among multiple Arabic dialects, and the frequency of certain vocabulary items when compared to others that have the same meaning. For example, a word like 'window' is obviously a basic word to introduce in novice levels. While the most frequently used lexical item in written MSA is نافِذة [19], it is not common in the Arabic dialects. In fact, searching through MDD and MADAR Lexicon show that the alternative vocable شباك is a more common one across a large number of dialects. Table 1 below shows the equivalents of 'window' in the dialects.

Table 1: A screenshot of the search results of the word 'window' on the MDD

window

Run Search | Download Results

Sample	Location	Tags	English Phrase	Arabic Phrase
DZA001	Oum El Bouaghi, Algeria	Vocab: Misc N	window	tāqa tyəqq
EGY001	Fayoum Oasis, Egypt	Vocab: Misc N	window	šabbāk, šabābīk
IRN001	Ahwaz, Iran	Vocab: Misc N	window	panjara
IRQ001	Baghdad, Iraq	Vocab: Misc N	window	šubbāč, šabābīč
IRQ002	Basra, Iraq	Vocab: Misc N	window	šubbāk, šabābīk
JOR001	Karak, Jordan	Vocab: Misc N	window	šib-bāk, šababīk
KWT001	Kuwait City, Kuwait	Vocab: Misc N	window	dərīša
LBN001	Nabatiya, Lebanon	Vocab: Misc N	window	šəbbāk šababīk
LBY001	Sabha, Libya	Vocab: Misc N	window	nāfia, nwāfid
MAR001	Meknes, Morocco	Vocab: Misc N	window	seržəm, ṣrāžəm
OMN001	Salalah, Oman	Vocab: Misc N	window	xalfa, xalāf
YEM001	Marib, Yemen	Vocab: Misc N	window	šubbāk, šubbālīk
SYR003	Damascus, Syria	Vocab: Misc N	window	šibbāk šabābīk
SYR001	Aleppo, Syria	Vocab: Misc N	window	šubbāk šabābīk
SDN001	Khartoum, Sudan	Vocab: Misc N	window	šubbāk
SAU002	Jedda, Saudi Arabia	Vocab: Misc N	window	šubbāk šabābīk
SAU001	Riyadh, Saudi Arabia	Vocab: Misc N	window	šubbak, šebabīk
PSE001	Tul Karim, Palestine	Vocab: Misc N	window	šubbāk šababīk
SYR002	Daraa, Syria	Vocab: Misc N	window	šubbāk nawāfiḏ/šabābīk
EGY002	Cairo, Egypt	Vocab: Misc N	window	šibbāk šabābīk

[16] http://www.arabic.humanities.manchester.ac.uk/database-of-arabic-dialects/

[17] https://sites.google.com/nyu.edu/madar/

[18] Resources consulted include: (Ben Abdelkader, 1977; Chekayri, 2011; Clarity, 2003; Hinds & Badawi, 1986; Holes, 2010; Stowasser & Ani, 2004)

[19] As demonstrated by Buckwalter & Parkinson (2011) and Familiar (2021).

When faced with this kind of data, we decided to prioritize the lexis that would increase the learner's level of cross-dialectal comprehension, as well as their overall Arabic competency in the early stages of their learning.

A second issue that arises when selecting what vocabulary items need to be learned at each CEFR level is the grammatical side of lexis and its complexity when comparing MSA and dialectal items. For example, when should the dual or the feminine plural pronouns be introduced, considering that these are not used in most Arabic dialects? Should we prioritize them and present them in A1, or should they be postponed? What is the place of syntactical and morphological complexity when designing a vocabulary profile?

The third issue that had to be considered is how this linguistic variation and complexity connects to the CEFR descriptors of skills and abilities that learners need to develop at each proficiency level. For example, greetings and farewells are basic communicative functions that learners need to master from A1. Given that some of these are multi-word expressions, we do not always find them in corpus-based vocabulary lists. In the AVP, such basic social expressions (e.g. مع السلامة 'good-bye') were included.

Linguistic variation in the AVP selection criteria

The following criteria were used when classifying the lexicon into A1 and A2 levels, not in any particular order of priority:

Frequency of use as recorded by the dictionaries of Familiar (2021) and Buckwalter & Parkinson (2011) provided useful word lists which learners are likely to encounter in various contexts, especially in written language. Both dictionaries focus largely on MSA, but they also contain some dialectal words. Familiar's list provided a starting point in aligning lexis due to its context of contemporary fiction, which is closer to everyday language when compared to Buckwalter & Parkinson's. For example, the adjective مُتَأَكِّد 'sure' has a ranking of 1772 in Familiar's, while in Buckwalter & Parkinson's, it is beyond the top 3000 frequent words which ultimately reflects the nature of corpora used in the elaboration of each dictionary. Frequency lists in most urban dialects are still scarce which made it difficult for us to compare the frequency of words across different varieties. Nevertheless, resources such as the MDD and MADAR Lexicon enabled us to compare word alternatives across many urban dialects and, in several incidents, it helped us prioritise certain lexis when found to be more common in multiple dialects. For example, the word 'newspaper' is a basic word to be learnt at A1 according to the

CEFR descriptors[20]. However, in MSA, there are two vocables used to convey the same meaning: جريدة and صحيفة. According to Familiar's and Buckwalter & Parkinson's dictionaries, the word صحيفة is more frequent. However, when checking the different dialects on the MDD, the word صحيفة appears to be occasionally used in Oman and Yemen. Most other urban dialects use جريدة and its plural جرايد (in addition to the French word approximation of جرنال/جرنان in Egypt and Algeria). Therefore, it made sense to introduce جريدة in A1, and delay صحيفة to A2 or B1 despite its higher frequency in MSA.

In other instances, frequency in MSA was a helpful tool in prioritising a word when it had different equivalents across the urban dialects. For example, the word 'weather' has two equivalents in MSA and the dialects: جَوّ and طَقْس. A search on MADAR Lexicon shows that the two words are quite common in many dialects. However, the former is ranked much higher in frequency when compared to the latter[21]. Therefore, the word جَوّ was included in the A1 list while طَقْس was recommended to be included in higher levels or introduced earlier if it is a common vocable in the dialect(s) that students are learning. There were also instances when neither frequency nor dialectal variation influenced the prioritisation of certain words such as in the case of صَحْن and طَبَق which are both common in many dialects. Table 2 shows their frequencies, which are very similar, even when compared to English. For such words, it was decided that both should be introduced at the same level, which in this case is A2.

Table 2: Comparison of frequencies of صَحْن and طَبَق

CEFR level according to the EVP	English equivalents	Frequency ranking in English22	Frequency ranking in Buckwalter & Parkinson's	Frequency ranking in Familiar's	
A1	Plate	1214	4518	1423	صَحْن
A2	Dish/plate	1387	4110	1558	طَبَق

[20] Also according to the EVP.

[21] جَوّ is ranked as 695 in Familiar's and 667 in Buckwalter & Parkinson's; while طَقْس is ranked as 1806 in Familiar's and 2657 in Buckwalter & Parkinson's.

[22] According to: https://frequencylist.com/

Alignment with the CEFR descriptors is integral to word selection in the AVP in two major areas:

A. Linguistic competence descriptors and learners' needs. This meant prioritising the selection of lexis to be taught and learnt in accordance with the CEFR descriptors and the actual needs of Arabic learners, at each proficiency level. Although word frequency was an important factor in word selection, it was not sufficient in the elaboration of a comprehensive classification. For example, the word 'teacher' مُدَرِّس comes as number 1940 in Familiar's and 2196 in Buckwalter & Parkinson's which are both low in frequency. However, in the context of learning Arabic, this is one of the first nouns of occupations to be learnt. Even when learners are self-studying, 'teacher' is a basic occupation in society and one that students will probably need to use; therefore, it has to be introduced early on, even when it does not rank high in corpus-based frequency dictionaries. In other words, relying on the logic of basic needs was an important factor in deciding which lexis to select at each CEFR level. Along with the word مُدَرِّس, the word أُسْتاذ was also ranked at A1 level due to its relevance as an occupation and as a title. In aligning the words with A1 and A2, we relied on our expertise and intuition as well as on cross-referencing in resources such as the EVP and the Oxford Learners Word Lists[23]. There were several instances of lexis such as the verbs 'to eat' and 'to drink' which were not highly frequent in the frequency dictionaries but essential to include in the low CEFR levels. In these instances, the CEFR descriptors were prioritised over frequency.

B. The promotion of plurilingualism, as recommended by the CEFR Companion Volume (2020). In the case of Arabic learners, this means training students to recognize and understand register and dialectal variation at the lexical level, early on. We believe that raising learners' awareness in cross-dialectal and cross-register competences from the beginning, can enable them to adjust and regulate themselves linguistically in a variety of communicative contexts, even when their lexical repertoire is limited. For example, expressions of daily life, such as classroom language expressions (e.g. ماشي، يَلّا، كفاية)

[23] It should be noted that, while the initial word alignment in the AVP was done by us as the project leaders, all relevant documents and word lists were made available to over a hundred practitioners of Arabic as L2 teaching that volunteered to participate in a validation process that is currently under way. Therefore, the final AVP could result in a slightly different form of what is being described here.

were incorporated/distributed across A1 and A2, based on the CEFR descriptors, even when these do not rank high in the Arabic frequency resources used and even when they may not be considered MSA words. Furthermore, the AVP recommends that teachers find the equivalents of those expressions based on the dialectal variety they teach.

In aligning with the CEFR in this way, we believe that the AVP is well positioned to empower learners as social agents capable of operating successfully in a variety of communicative contexts, both at the linguistic and intercultural levels.

Linguistic complexity and root prevalence: The basic principle followed was that, in the existence of more than one Arabic vocable to convey the same meaning, the simpler one would have the priority when introduced at the earlier levels. It is not an easy task to define linguistic complexity, but for our context, it refers to the length of words, root complexity (e.g. sound versus weak roots), or even pronunciation difficulties. These factors were considered, as long as dialectal commonality and frequency are also taken into account. For example, in the case of the verb 'to return', we opted to prioritise 'يَرْجِع' over 'يَعود' due to the fact that the latter is a hollow verb which entails a certain degree of complexity when conjugating. Both lexemes are high in frequency, and both are used across a range of dialects.

There are instances when both multi-dialectal commonality and linguistic complexity guided the benchmarking of words. An example for this is the word 'building', which is essential for the basic function of providing an address. In principle, this word would need to be listed in the AVP at A1. However, because Arabic has three vocables to express this concept (بِناية – عمارة – مَبْنى), we had to consider the three criteria of frequency, multi-dialectal use, and complexity when selecting the lexical item that is most suitable for A1. To start with, it is important to note that the three vocables are interchangeably used in both MSA and a large number of dialects in different contexts. Therefore, we checked their frequency of use in Familiar's and Buckwalter & Parkinson's dictionaries as well as their commonality in multiple dialects using the MDD.

Table 3: Frequency of use in MSA and commonality in urban dialects for the words بِناية – عمارة – مَبْنى

Average percentage of use in the dialects	Number of instances of use out of 17 dialects[25]	Number of instances of use out of 20 dialects[24]	Buckwalter & Parkinson's frequency	Familiar's frequency	
41%	8 (47%)	7 (35%)	4578	1180	بناية
25%	5 (29%)	4 (20%)	2892	Not listed	عمارة
35%	4 (24%)	9 (45%)	1524	764	مبنى

The list of frequencies in table 3 shows that in MSA, the word مبنى is more common than the other two, and that there are clear differences in the rank of بناية and عمارة (due to the nature of the corpora used in the elaboration of Familiar's and Buckwalter & Parkinson's dictionaries). When looking at their use in urban dialects using MDD, the search showed that both مبنى and بِناية are interchangeably common in most dialects. Because there are no big differences between مبنى and بناية in terms of multi-dialectal commonality, and since they are both derived from the same root, word complexity was the criteria used in deciding which vocable to introduce at A1. Knowing that the word مبنى is اسم منقوص 'a weak noun' (with a slightly complicated plural), the word بناية (which has a regular feminine plural) was a better choice for A1. The word مبنى can then follow in A2 or even in B1. In this example, linguistic complexity aided lexical classification when frequency and dialect commonality were not sufficient indicators.

In addition to complexity, prevalence of root derivatives was a helpful tool in prioritizing lexis. Going back to the example of the word 'teacher', and as mentioned above, we listed مُدَرِّس and أُسْتاذ at A1 level due to their important relevance to the CEFR descriptors. However, one can ask why مُدَرِّس and not its equivalent مُعَلِّم? This was another challenging case as both words are common in multiple dialects. In terms of frequency, both dictionaries rank مُعَلِّم much higher than مُدَرِّس. However, looking at some of the derivatives of the root د – ر – س, such as basic A1 words (e.g. 'school' مَدْرَسة, 'to study' يَدْرُس, and 'a study' دِراسة), prioritising مُدَرِّس over مُعَلِّم was a logical decision as it can help learners cognitively connect between these derivatives and familiarise themselves with the system of roots and derivations in Arabic, from an early stage. Another linguistic factor to consider here is the relatively easier pronunciation of مُدَرِّس when

[24] A search using MDD provided a translation of the sentence 'It was the most modern and biggest building in the town' into 20 dialects which this data is based on.
[25] A search using MDD provided a translation of the sentence 'This old building is still standing' into 17 dialects which this data is based on.

compared with مُعَلِّم as the latter contains the voiced pharyngeal fricative ع which typically challenges learners, both in terms of recognition and production. We would like to emphasise that, teachers are certainly free to introduce both words at A1 if they see it fit with their teaching context; but if they are to prioritise one over the other, then a selective process that considers factors of frequency, multi-dialectal commonality, CEFR relevance, and linguistic complexity and prevalence can provide a reasoning that contributes to fulfilling their learner's needs.

To illustrate further how issues of linguistic complexity intersected with a multi-dialectal approach in designing the AVP, we would like to offer some additional examples:

A. Dual pronouns, feminine plural verbal forms, and the negation morpheme لم were not incorporated in the A1-A2 levels. When forms and lexis are not used in dialects, or are very rare, and when morphological complexity is involved (such as in the case of feminine plural conjugations, which pose a considerable cognitive load on the learner), these were postponed.

B. Weak nouns and adjectives such as غالٍ are listed in their definite form (that is غالي) considering that this is the common multi-dialectal usage and that there is a considerable complexity involved in its original MSA usage.

C. The AVP comes with a list of grammatical notes among which it is stated that case endings should not be included in the A1-A2 levels, with the exception of adverbal lexis that take the indefinite accusative marker (e.g. دائماً، طبعاً). In these cases, the case ending is not considered a declination marker, as dialectal usage demonstrates.

The examples above are aligned with the project leaders' belief that an MSA grammatical 'good' compromise, at the lower proficiency levels, can yield better communicative results for the learner. On the contrary, a zeal for a grammar-purity approach poses an excessive cognitive overload on learners, and such rigidity impregnates their speech with unnatural markers that can impede successful communication with Arabic speakers.

Even though there were some instances of vocables that required thorough analysis utilizing the different criteria presented above, there were many instances when all criteria of frequency, multi-dialectal commonality, linguistic complexity and CEFR relevance harmoniously supported the prioritization of a word over another. An example for this is the equivalents of the noun 'shop' مَحَلّ - دُكّان – حانوت. In this case, the vocable مَحَلّ was prioritised due to its high frequency in MSA and its commonality across a range of dialects, in addition to its simplicity

of use in both singular and plural forms. Table 4 below presents the data found regarding the frequencies and dialect commonality of the three vocables.

Table 4: Frequencies and dialect commonality of مَحَلّ - حانوت – دُكّان.

Number of instances of dialectal use in MDD	Buckwalter & Parkinson's frequency	Familiar's frequency	
20	701	451	مَحَلّ
10	Not listed	1330	دُكّان
2	Not listed	Not listed	حانوت

Conclusions

This paper discussed the importance of a variationist vision in Arabic L2 teaching that should encompass aspects of teaching and learning materials, classroom activities, language resources, and eventually assessments. The initiative of creating an Arabic vocabulary inventory that is aligned with the CEFR descriptors highlights the role of multidialectal consideration in prioritising vocabulary to be taught at novice levels. As discussed in this paper, a multidialectal approach to Arabic L2 teaching does not mean denying the identity of each Arabic variety, but rather it supports the integration of as many varieties as relevant to the learning context, with the aim of creating an inclusive and a comprehensive approach to teaching the language. Within the decisions of prioritising One out of the Many, the Many cannot be ignored.

The AVP project presented here, with the future aim of creating CEFR-based graded readers and other pedagogical materials, provides learners and teachers with CEFR-aligned lexis that factor frequency, multi-dialectal commonality, linguistic complexity and root prevalence. Frequency in Familiar's dictionary paved the way as a starting point to lexis selection and in some instances, frequency was the main factor in prioritising equivalent vocables as in the case of جَوّ and طَقْس. There were also instances when frequency was overlooked when relevance to the CEFR descriptors entailed the inclusion of words that are less frequently used in written Arabic such as in the case of basic verbs like يَأْكُل – يَدْرُس – يَشْرَب. In other instances, linguistic complexity or the prevalence of common root derivatives guided the prioritisation of lexis such as in the example of مُدَرِّس and مُعَلِّم. Encompassing all these criteria was dialectal commonality which promoted the ordering of words such as in the examples presented of the equivalents بِناية – عمارة – and مَبْنى جريدة versus صحيفة. Multidialectal considerations also entailed the addition of some everyday language words and phrases that can controversially be considered not

appropriate to MSA teaching such as يَلّا - ماشي – خَلاص – صافي but which deemed to be frequently encountered by learners of any varieties of Arabic.

The multidialectal approach to vocabulary selection also entailed what we refer to as the 'good compromises' when it comes to the grammatical forms of certain words, such as in the case of weak forms in MSA (e.g. كافٍ and غالٍ) which were listed in their more dialectal forms (their classical definite forms كافي and غالي). The goal was to provide novice learners with the forms that they are more likely to encounter in their basic communicative interactions. These and other lexis that are closer to their dialectal forms are not seen as 'wrong' forms, but rather as better options in equipping learners with the vocabulary they need at the beginner level, and ones on which they can build on later with more complex variations.

At the time of writing this paper, the AVP has a list of approximately 400 lexical items at A1 level and approximately 800 at A2 level. These are currently being verified by Arabic educators who are familiar with the CEFR. In selecting the verifiers, diversity was an important aspect to factor in. We made sure that we have verifiers from different linguistic backgrounds in terms of the Arabic dialects they speak, the regions in which they teach Arabic at Higher Education level, with some of them speaking Arabic as L1 and others as L2 speakers. This was also a multidialectal approach that we believe would ensure that the different perspectives of Arabic language use are taken into consideration. Once the verification process and the analysis of data are completed, the Arabic Vocabular Profile at A1 and A2 CEFR levels will be available for use by learners and teachers globally.

References

Abu-Melhim, A. H. (1992). Communication across Arabic dialects: Code-switching and Linguistic accommodation in informal conversational interactions. PhD Thesis, Texas A&M University.

Al-Batal, M. (2018). *Arabic as one language: Integrating dialect in the Arabic language curriculum*. Georgetown University Press.

Aldrich, M. (2018). Arabic vs. Arabic - A dialect sampler: Lingualism.com.

Badawi, E. S. (1973). *Mustawayaat 'al carabiyya 'al-mucaasira fi Misr*. Dar 'al-Maᶜaarif.

Belnap, R. K. (2006). A profile of students of Arabic in U.S. universities. In K. M. Wahba, Taha, Z. A., & England, L. (Ed.), *Handbook for Arabic language teaching professionals in the 21st century*. Lawrence Erlbaum.

Ben Abdelkader, R. (1977). *Peace Corps English-Tunisian Arabic Dictionary*.

Buckwalter, T., & Parkinson, D. B. (2011). *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge.

Chekayri, A. (2011). *An introduction to Moroccan Arabic and culture*. Georgetown University Press.

Chelghoum, A. (2017). Social network sites and Arabic diglossia between threatening Modern Standard Arabic and strengthening Colloquial Arabic. *International Journal of Language and Linguistics. Special Issue: New Trends in Arabic Sociolinguistics*, 5(3-1), 36-43.

Clarity, B. E. (2003). *A dictionary of Iraqi Arabic*. Georgetown University Press.

Dooly, M., & Vallejo, C. (2020). Bringing plurilingualism into teaching practice: a quixotic quest? *International Journal of Bilingual Education and Bilingualism*, 23(1), 81-97.

Ezzat, A. G. E. (1974). *Intelligibility among Arabic dialects*. Beirut Arab University.

Familiar, L. (2021). *A frequency dictionary of contemporary Arabic fiction: Core vocabulary for learners and material developers*. Routledge.

Ferguson, C. A. (1963). Problems with teaching languages with Diglossia. Paper presented at the thirteenth Annual Round table Meeting on Linguistics and Language Studies, Georgetown University

Francescato, G. (1986). Bilingualism and diglossia in their mutual relationship. In C. A. Ferguson & J. A. Fishman (Eds.), The Fergusonian impact: in honor of Charles A. Ferguson (Vol. 2, pp. 395-402). Mouton de Gruyter.

Freeman, A. (1996). Perspectives on Arabic Diglossia. Retrieved 30/03/2007, from http://www-personal.umich.edu/~andyf/digl_96.htm

Giolfo, M., & Salvaggio, F. (2017). The role of VLE in enhancing authentic proficiency in Arabic in the light of CEFR. Paper presented at the The first biennial Arabic Language Teaching & Learning in UK Higher Education Conference.

Giolfo, M. E. B., & Sinatora, F. L. (2011). Rethinking Arabic diglossia: Language representations and ideological intents. In Valore, P (Ed.), *Multilingualis: Language, power, and knowledge* (pp. 103-128). Edistudio.

Hinds, M., & Badawi, E.-S. M. (1986). *A dictionary of Egyptian Arabic: Arabic-English*. Librairie du Liban.

Holes, C. (2010). *Colloquial Arabic of the Gulf : The complete course for beginners* ([New ed.] ed.). Routledge.

Husseinali, G. (2006). Who is studying Arabic and why? A survey of Arabic students' orientations at a major University. *Foreign Language Annals*, 39(3), 395-412.

Khalil, S. (2011). Talk like an Egyptian: Egyptian Arabic as an option for teaching communicative spoken Arabic. *Leeds Working Papers in Linguistics and Phonetics*, 16, 1-28.

Khallaf, N., & Sharoff, S. (2021). Automatic difficulty classification of Arabic sentences. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., et al. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1), 121-163.

Landau, J. M. (1959). *A word count of modern Arabic prose*. American Council of Learned Societies.

Leech, G. N., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. Longman.

Marello, C. (2012). Word lists in Reference Level Descriptions of CEFR (Common European Framework of Reference for Languages). Paper presented at the The 15th EURALEX International Congress, Oslo,Norway.

Matsuoka, W. (2012). Searching for the right words: Creating word lists to inform EFL learning. *Linguistic Insights*, 155, 151-177.

MLA. (2009). Report to the Teagle Foundation on the undergraduate major in language and literature. https://www.mla.org/content/download/3207/file/2008_mla_whitepaper.pdf

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3-13). John Benjamins Publishing Company.

QAA. (2023). Subject Benchmark Statement - Languages, Cultures and Societies. from https://www.qaa.ac.uk/docs/qaa/sbs/sbs-languages-cultures-and-societies-23.pdf?sfvrsn=3c71a881_8

QFI. (2022). Incorporating Arabic Varieties in the Communicative Classroom. from https://www.qfi.org/for-teachers/teacher-resources/arabic-communicative-classroom/

Schiffman, H. (1999). Diglossia as a sociolinguistic situation. Retrieved 12/12/2008, from http://ccat.sas.upenn.edu/~haroldfs/messeas/diglossia/

Soliman, R. (2015). Arabic cross-dialectal conversations with implications for the teaching of Arabic as a second language. University of Leeds, Leeds.

Soliman, R. (2018). The implementation of the Common European Framework of Reference for the teaching and learning of Arabic as a second language in higher education. In K. M. Wahba, L. England & Z. A. Taha (Eds.), *Handbook for Arabic Language Teaching Professionals in the 21st Century* (Vol. 2, pp. 118-137). Routledge.

Soliman, R. (2023). Arabic variation as a CEFR-based sociolinguistic competence: principles to inform Arabic L2 teaching. In M. Giolfo & F. Salvaggio (Eds.), *Relating Arabic L2 Teaching to CEFR Guidelines: Principles, Approaches, and Practices*. Aracne.

Stokes, M. C. (1971). *One and Many in Presocratic Philosophy* (Vol. 93): Center for Hellenic Studies. Harvard University Press.

Stowasser, K., & Ani, M. (2004). *A dictionary of Syrian Arabic : English-Arabic*. Georgetown University Press.

Thorndike, E., & Lorge, I. (1968). *The teacher's word book of 30000 word*s. Teachers College Press.

Thorndike, E. L. (1921). *The teacher's word book*. Teachers College, Columbia University.

Torigoe, S. (2016). Seeking the Portuguese vocabulary profile. Paper presented at the CILC2016. 8th International Conference on Corpus Linguistics.

Trentman, E., & Shiri, S. (2020). The mutual intelligibility of Arabic dialects: Implications for the language classroom. *Critical Multilingualism Studies*, 8, 104-134.

Van Mol, M. (2003). *Variation in Modern Standard Arabic in radio news broadcasts: A synchronic descriptive investigation into the use of complementary particles*. Peeters.

Younes, M. (1995). Elementary Arabic: An integrated approach. Yale University Press.

Younes, M. (2006). Integrating the Colloquial with Fusha in the Arabic as a Foreign Language Classroom. In K. M. Wahba, Z. A. Taha & L. England (Eds.), *Handbook for Arabic Language Teaching Professionals in the 21st Century* (pp. 157-168). Lawrence Erlbaum Associates.

Younes, M. (2014). *The integrated approach to Arabic instruction*. Taylor & Francis.