Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3053–3062
Marseille, 11–16 May 2020

A Large-Scale Leveled Readability Lexicon for Standard Arabic

Muhamed Al Khalil, Nizar Habash, Zhengyang Jiang
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
{muhamed.alkhalil, nizar.habash, zj522}@nyu.edu

Abstract

We present a large-scale 26,000-lemma leveled readability lexicon for Modern Standard Arabic. The lexicon was manually annotated in triplicate by language professionals from three regions in the Arab world. The annotations show a high degree of agreement; and major differences were limited to regional variations. Comparing lemma readability levels with their frequencies provided good insights in the benefits and pitfalls of frequency-based readability approaches. The lexicon will be publicly available.

Keywords: Readability, Lexicon, Arabic

1. Introduction

Modeling readability levels is relevant to a range of natural language processing (NLP) tasks from developing language education applications to user profiling. Much work has been done on readability leveling and its assessment and specification in English leading to the development of many resources and tools. However, this is not the case for many other languages.

The work presented in this paper is part of a project on *Simplification of Arabic Masterpieces for Extensive Reading* (SAMER) (Al Khalil et al., 2017; Al Khalil et al., 2018). Specifically, we discuss the challenges of, and solutions to, the development of a large-scale leveled readability lexicon for Modern Standard Arabic (MSA). Although some aspects of the effort are Arabic specific, we situate it within the larger frameworks of approaches to readability lexicon development.

Our contributions include the following:

- We define a five-level readability scale for MSA targeting native speakers and create annotation guidelines for it.
- We manually annotate a 26,578-lemma lexicon in triplicate by language professionals from three different regions in the Arab world.
- We report on a detailed analysis of the major disagreements among our annotators.
- We carefully study the relationship between frequency-based and intuition-based readability classifications using our newly created lexicon.
- We make this new lexicon publicly available to support research and tool development for Arabic readability tasks.[1]

The rest of this paper is structured as follows. We start with three background sections on approaches to readability lexicon development (Section 2), previous related work (Section 3), and relevant linguistic facts about Arabic (Section 4). Section 5 details our approach including guideline creation, data extraction, and manual annotation. We discuss our results in Section 6. Section 7 concludes and maps some future work directions.

2. Approaches to Readability Lexicons

We present next two different approaches to the development of leveled readability lexicons: the frequency approach and the intuitive approach.

2.1. The Frequency Approach

Frequency and frequency-derived measures have allowed for the study of language usage on a massive scale, and produced many innovative advancements in the mapping and understanding of language, especially for teaching and learning purposes. Examples include corpus-based dictionaries (Sinclair and others, 2003), text-book series (Beech, 2011), and graded systems for language learners.[2] A great advantage of the frequency approach is the relative ease (in terms of time and cost) of creating, and updating, frequency lists, assuming a machine-readable corpus is readily available. Yet, despite its contributions, corpus-based research has been criticized by theoretical linguists who maintain that no amount of text could account for humans' full linguistic competence, nor capture or represent all language use across time and place (Wynne, 2005; Teubert and Cermáková, 2007). Aside of the above mentioned limitations, using the frequency list approach to develop leveled readability lexicons presents us with two problems: gradability and readability.

Gradability Having generated a frequency list that is fairly representative of current usage in a language, and wanting to divide this list into a clear set of levels for pedagogical purposes, there seems to be no natural or golden rule as to how many levels the list should be divided into without reference to an outside scheme. But even when a scheme is adopted, say the 6-level elementary-to-advanced system common in US universities, the question persists as to where to place level boundaries along the frequency continuum. Should the list be divided equally (among six levels in our example)? Or should the lower levels receive

[1] http://resources.camel-lab.com/

[2] The Lexile Framework for Reading: https://lexile.com/

less words? And how many less? These questions suggest there is no mathematically objective way to automatically create a leveled scheme without external subjective input into the process.

Readability Readability is usually defined in relation to a text and signifies how accessible the text is for a specific reader. It is affected by several factors besides the commonness of the words in the text, for example by sentence length, syntactical structures, abstractness, etc. In a frequency list a word's readability can only be automatically assessed based on its frequency rank: the more common the word the more readable it is assumed to be. First, this brings back the corpus representation problem: in Buckwalter and Parkinson (2014)'s *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*, the word رئيس *rŷys*[3] 'president' ranks 47th in frequency whereas كتاب *ktAb book* ranks 210th, قراءة *qrA'ħ* 'reading' ranks 923rd, and مكتبة *mktbħ* 'library' ranks 1830th. Should the learners be taught the word for *president*, based on this dictionary's frequency list, which is clearly skewed toward "newspaper Arabic, before the other words which seem more relevant to a student's environment? Secondly, it gets more complicated if the goal is to use a frequency list to create a leveled system for young learners. For even in a more balanced and representative corpus, it is unlikely that the word سلحفاة *slHfAħ* 'turtle', a dear character in children's books, would be frequent enough to be automatically placed in first levels (the word does not even make it to the Buckwalter and Parkinson's dictionary mentioned above). Compounding this is the fact that language corpora generally reflect the world of adults, and children's literature is practically composed by adults (Zipes, 2013). For all these reasons, frequency-based readability seems ill-suited to be the sole basis for language leveling.

2.2. The Intuitive Approach

In their seminal work on guided reading, Fountas and Pinnell (2017) offer an alternative, and far more comprehensive, approach to assign text readability. They list ten characteristics to consider when determining a text's level: genre, text structure, content, themes and ideas, language and literacy features, sentence complexity, vocabulary, words, illustrations, and books and print features. They rely on the teacher's experience and intuition to study these characteristics in a text, and then place it on a reading gradient from A to Z. This approach is mainly used in schools by groups of teachers to sort their book collections and place them along a reading continuum. They could then match their students to appropriate texts and follow their progress on the continuum.

In our research we combined aspects from both the frequency and intuition approaches to extract a large MSA lexicon from a corpus, and manually annotate it in a five-level scale with the help of Arabic language professionals.

[3]Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

3. Related Work

For readability measurement design, one of the most crucial tasks is deciding on an appropriate formula. The formula is constructed based on selected features that are associated with text's lexical content and are considered important to determine the text's readability. However, as shown in (Al-Khalifa and Al-Ajlan, 2010), those formulas are highly language dependent. For example, a very early and widely used Laesbarhedsindex (LIX) formula (Anderson, 1983), originally designed for Swedish, taking only number of words, number of sentences and number of difficult words (defined as words with more than 6 letters) into consideration, does not work well for determining Arabic readability. Another popular approach is to define semantic difficulty by frequency or appearance in a list of familiar words (Fry, 2002).

Many previous efforts have attempted to use word frequency as one of the factors to determine readability for Arabic texts. Al-Dawsari (2004) describes an Arabic readability formula that includes five features: average word length, average sentence length, word frequency, percentage of nominal clauses, and percentage of definite nouns. Al-Khalifa and Al-Ajlan (2010) processed educational materials for elementary, intermediate and secondary schools in Saudi Arabia using features such as average word length in letters and syllables, term frequency (ratio of duplicated words), and a bigram language model with a machine learning classifier. Forsyth (2014) used a machine learning approach to process the online curriculum of the Defense Language Institute Foreign Language Center and concluded that most (19 out of 20) of the best features are from the POS-based frequency feature set. Saddiki et al. (2018) presented results on the use of a large number of morphological and syntactic features and n-gram models to automatically predict Arabic readability level for native and non-native speakers.

The KELLY project (Kilgarriff et al., 2014) is another notable research effort that aims to develop monolingual and bilingual word lists for language learning. This project aims to map the most common 9,000 words in nine languages (including Arabic) onto Common European Framework of Reference for Languages (CEFR) levels through corpus-based frequency analysis and comparisons between translated language pairs across the said nine languages.

Research on the design and use of word frequency lists is abundant in English and other world languages. Of particular relevance to this paper, however, are the novel ways and techniques that some researchers use to address the challenges and shortcomings faced in creating these lists. For example, Brooke et al. (2012) created lexicons where words are classified based on the readability of the web documents they appear in and the words they co-occur with. Another example is Ehara (2018) who used crowd sourcing to simulate testing on the vocabulary size test, a well-studied English vocabulary test, by one hundred test-takers – producing a reliable English vocabulary knowledge data set of Japanese learners of English.

In the field of Arabic education, intuition and experience form the basis in most projects involving readability lev-

Category	Example	Token Count	Token%	Unique Type Count
All Words		11,188,566	100.0%	31, 542
Punctuation	, . ! ?	1,510,501	13.5%	19
OOV	بو تجيج *bwtjyj* 'Buttigieg'	188,694	1.7%	1
888	2002	151,178	1.4%	1
Latin Script	IBM	13,470	0.1%	1
Proper Noun	فَرَنْسا *faransA* 'France'	500,031	4.5%	4,112
Capitalized Noun	يُوان *yuwAn* 'Yuan'	77,480	0.7%	404
Capitalized Adjective	هِنْدِيّ *hindiy∼* 'Indian'	117,854	1.1%	367
Abbreviation	كلم *klm* 'km/kilometer'	22,714	0.2%	59
Filtered		2,581,922	23.1%	4,964
Annotated		8,606,644	76.9%	26,578

Table 1: Corpus statistics detailing the total token and type counts and the percentages of different categories that were excluded from annotation.

eling. There are many such projects, but one key recent effort is the Hanada's Text leveling System (HTLS), a 19-level procedural framework, based on the Fountas & Pinnell gradient, which seeks to match Arabic learners with texts suitable to their reading and learning capabilities. The pure qualitativeness of the system has been cited as one of its biggest advantages (Harb, 2019), and which contrasts with our more hybrid approach.

4. Relevant Arabic Linguistic Facts

The Arabic language poses a number of challenges for natural language processing tasks (Habash, 2010). We focus here on three aspects that are most relevant to the question of modeling readability: morphological richness, orthographic ambiguity, and regional variations.

Morphological Richness Arabic is a morphologically rich and complex language. It employs a combination of templatic, affixational, and cliticization morphological operations to realize a large number of features such as gender, number, person, case, state, aspect, voice, and mood, in addition to a number of attachable pronominal, preposition and determiner clitics.

Orthographic Ambiguity Arabic is commonly written with optional diacritical marks – which are often omitted – leading to rampant ambiguity.

Orthographic ambiguity and morphological richness interact heavily with each other. For example, the undiacritized word وحدتها *wHdthA* has a number of readings with varying analyses including *waH∼adat+hA* 'she united it', *wi-Hdata+hA* 'her unity [accusative]', and *wa+Hid∼atu+hA* 'and her sharpness [nominative]'. The different readings have to be disambiguated in context. There are a number of tools for Arabic automatic disambiguation and lemmatization. In this work, we use the MADAMIRA tool (Pasha et al., 2014) which reports a 96.0% accuracy on Modern Standard Arabic (MSA) lemmatization. These three examples have three different lemmas (lexical entries) that abstract away from the various inflections: وَحَّد *waH∼ad* 'to unite', وِحدَة *wiHdaħ* ' unity', and حِدَّة *Hid∼aħ* 'sharpness'.

In this work, we focus on the lemmas as the main unit of readability annotation. We acknowledge that different morphological and morphosyntatic features will have different readability levels associated with them (Saddiki et al., 2018), but we leave this issue to future work.

Regional Variations The third challenge we highlight here is that of Arabic dialectness and diglossia (Ferguson, 1959; Holes, 2004). Arabic today is a collection of variants amongst which MSA is the official prestigious standard of the media, literature and education. The other variants are the so-called dialects of daily speech and social media. While our focus in this work is on MSA, we acknowledge a degree of lexical variety reflecting regions with possibly different dialects. As such, it is possible that some MSA words may be more similar to the dialect of a certain region and different from another, and thus have a lower readability level in the former than the latter. To account for this issue in this work, we worked with three different human annotators from three distinct countries and dialectal regions: Egypt (Egyptian Arabic), Syria (Levantine Arabic) and Saudi Arabia (Gulf Arabic).

5. Approach

We present next our approach to building the readability lexicon. We first discuss the process of automatic extraction and filtering; and then detail the readability annotation guidelines, and their application.

5.1. Data Extraction and Preparation

Data Selection Given the goals of the SAMER project targeting the identification of levels of readability in works of fiction and simplifying them, as well as creating a resource to support applications for readability among school children in the Arab world, we chose to work with two readily available authentic data sets from the news and literature genres intended for proficient adult users:[4] (a) the Arabic Gigaword corpus (Parker et al., 2011), which is a comprehensive archive of newswire text data that has been

[4]We did not use the Curriculum Corpus described in (Al Khalil et al., 2018) due to copyright restrictions.

Level	Grade	Age	Examples
Level I	Grade 1	6	بَيْت، شَجَرَة، أَرْنَب، أَزْرَق، كَبير، صَنَعَ، أَكَلَ، فَرِحَ، عَلى، لكِن house, tree, rabbit, blue, big, to make, to eat, to be happy, on, but
Level II	Grades 2-3	7-8	جَزيرة، ذَهَب، سَنة، داكِن، أُسْطُواني، صَعْب، خَدَعَ، كافأَ، قُرْبَ، إذا island, gold, year, dark, cylindrical, difficult, to cheat, to reward, near, if
Level III	Grade 4-5	9-10	رِئة، مُتْحَف، مُعادَلة، مُمْكِن، مُوَحَّد، أَغْرى، نَدُرَ، لَدى، كَيْ، ما إنْ...حَتّى lung, museum, equation, possible, united, to entice, to be rare, with, for, no sooner... than...
Level IV	Grades 6-8	11-14	اِقْتِصاد، نُسْغ، طُمَأنينة، راقٍ، مُثْبَت، نَكَثَ، أَغْضى، إبّانَ، إنَّما، لَئِنْ economy, sap, tranquility, sophisticated, proven, to breach, to overlook, during, whereas, if (were)
Level V	Specialist	15 -	أَدَمة، قَسْطَرة، هَيْضة، مِطْياف، لَوْذَعي، شُعَبيّ، لَحا، ظَعَنَ، لَدُنْ، أَنّى epidermis, catheterization, cholera, spectroscope, witty, bronchial, to denounce, to depart, with (≈ *chez* in French), wherever

Table 2: The five readability levels, their grade equivalencies, and lemma examples, according to the authors' scheme. The English translations of the Arabic do not always capture the exact comparable readability level.

acquired from Arabic news sources, and (b) the Hindawi Corpus, a corpus of Arabic literature built by collecting 129 works of fiction available in the public domain from the online catalog of the Hindawi Foundation.[5] We specifically took a sample of ∼11M tokens in balanced distribution from the Arabic Gigaword corpus (5,594,256 tokens) and the Hindawi corpus (5,594,310 tokens).

Automatic Lemmatization We automatically annotated each token in our corpus in context with morphological information including lemma and part-of-speech (POS) using the MADAMIRA tool for morphological disambiguation (Pasha et al., 2014). MADAMIRA out-of-vocabulary (OOV) tokens are all collapsed into the lemma "OOV". Digits are also collapsed into the lemma "888". The total number of unique lemmas is 31,542.

Categorical and Frequency Filtering We decided that certain categories should be excluded from annotation for obvious reasons, namely, punctuation marks, digits, Latin script words, and OOV. We also excluded abbreviations, proper nouns, and nouns and adjectives that refer to nationalities (identified through English gloss capitalization) because those lemmas are more closely associated with a person's general knowledge than a grade/readability level. Table 1 summarizes the lemma token and type (unique token) counts in the corpus, what we filtered out, and what was finally annotated. The filtering excluded 15.7% of all unique lemmas corresponding to 23.1% of all token occurrences. All remaining 26,578 lemmas were annotated.

5.2. Guidelines for Readability Annotation

Five Readability Levels for Standard Arabic In this research we opted to examine our frequency list against five levels of readability for MSA. Needless to say, readability could be represented in many gradations as is indeed seen in various graded reader schemes in the publishing and educational domains. The schemes differ on account of various factors such as target age groups, genres covered, readability criteria, etc. In fact, one leading publisher, Pearson English, a division of Pearson plc., alone employs four graded reader schemes:[6]

- *Kids Readers* for children 6 to 12 years: 6 levels
- *Story Readers* for children 5 to 11 years: 4 levels
- *Readers* for English students 12 years & up: 7 levels
- *Active Readers* for English students 12 years & up: 5 levels.

Fountas & Pinnel, a leading authority on reading development in K12 education in the United States, assert five broad stages of reading development for the young reader up to the eighth grade: Emergent, Early, Transitional, Self-extending, and Advanced (Fountas and Pinnell, 2006). We adopted the five levels of this gradient (and its ceiling at the eighth grade), but we did not follow exactly how they were mapped to school grades because grades generally straddle two or more levels in this scheme. While such overlapping mapping is pedagogically sound as it reflects classroom students with different reading proficiencies, it is unnecessary in a scheme like ours that assumes mid-range developing readers in each grade. For our practical purposes, we developed our own clear-cut mapping between levels and school grades as shown in Table 2.

Guidelines for Annotators A succinct but clear set of guidelines were developed and shared with the annotators as to how our levels generally correspond to reading expectations in the first eight grades in Arab public schools

[5] On 06/29/2017 from `http://www.hindawi.org/`.

[6] `https://readers.english.com/choose-readers`

Level	Types	Token %	Up to Token %
Level I	253	first 50%	50%
Level II	1,089	next 25%	75%
Level III	1,999	next 12.5 %	87.50%
Level IV	2,731	next 6.25%	93.75%
Level V	20,506	last 6.25%	100.00%

Table 3: Frequency-based pre-annotations provided to as starting point for the manual level annotation.

	A2		A3		Average	
	Accuracy	Correlation	Accuracy	Correlation	Accuracy	Correlation
A1	58.1%	86.6%	51.8%	83.9%	77.5%	92.9%
A2			53.1%	85.2%	79.2%	94.0%
A3					72.5%	92.2%

Table 4: Comparing Accuracy and Correlation across different human annotators (A1, A2, A3) and their rounded average level (Average). A1 is Egyptian, A2 is Syrian, and A3 is Saudi Arabian.

Level	Type Count		Token Count	
Level I	2,884	11%	4,448,580	52%
Level II	2,379	9%	1,468,063	17%
Level III	4,302	16%	1,496,034	17%
Level IV	7,515	28%	881,585	10%
Level V	9,498	36%	312,382	4%
Total	26,578	100%	8,606,644	100%

Table 5: Distributions of reading levels in the rounded average annotation

and thus what kind of words are expected to be placed in each level. The annotators were then requested to examine each word and to identify its readability level based on the following understandings (see examples in Table 2).

Level I: Generally corresponding to Grade 1, this level focuses on tangibles and thus includes short sensory words for objects, states, or actions common in the reader's environment, in addition to very common simple connecting words.

Level II: Generally corresponding to Grades 2-3, this level begins to add an imaginative dimension to the sensory which is still dominant in this level. It includes sensory words further removed from the reader's environment, semi-abstractions, and fairly common connecting words that connect clauses and sentences.

Level III: Generally corresponding to Grades 4-5, this level includes simple abstractions and more detailed or complex sensory words in addition to connecting words that build more complex in-sentence and intra-sentence relationships.

Level IV: Generally corresponding to Grades 6-8, this level includes complex but somewhat common abstractions that reflect logical development. It includes connecting words that are suitable to express such logical relationships between sentences.

Level V: This level reflects specialist language use beyond the eighth grade whether in educational or non-educational contexts. It includes words specific to a certain specialization, field, or profession. It also includes words of archaic or very rare usage.

5.3. Readability Annotation

The readability level annotation was conducted with the help of three language professionals from Egypt, Syria and Saudi Arabia. The effort was done in collaboration with a professional linguistic annotation firm, Ramitechs.[7] Ahead of the annotation itself, we ran an initial pilot study with three Arabic native speakers at New York University Abu Dhabi, which allowed us to test drive the annotation interface and estimate the annotation speed. To facilitate the annotation process, we provided a frequency-based pre-annotation to help speed up the process. We discuss the details of this step next and compare the real annotations against it in the next section.

Automatic Frequency-based Pre-annotation We pre-annotated the vocabulary list with a simple frequency-based assumption: the first half of all token occurrences (253 unique types) is assigned to Level I; and the first half of what remains (next 25% of token count mass, or 1,089 unique types) is assigned to Level II; and so on. Table 3 details the token mass and numbers of unique types in all the levels. This automatic level assignment was done on all the data that was kept after filtering.

Manual Intuition-based Annotation The annotators were given the detailed instructions provided above. We used a simple Google Sheets based interface for the annotation. It provided, row-by-row, the lemma, its POS, its English gloss and a drop-down menu with five level labels. The automatic pre-annotations were provided as the default answer.

[7] http://www.ramitechs.com/

Level Distance	Type Count		Token Count	
0	8,520	32.1%	3,405,390	39.6%
1	17,563	66.1%	5,182,894	60.2%
2	118	0.4%	3,918	0.0%
3	233	0.9%	10,325	0.1%
4	144	0.5%	4,117	0.0%
Total	26,578	100.0%	8,606,644	100.0%

Table 6: Statistics on disagreements among the three annotators in terms of level distance between the highest and lowest levels annotated levels per lemma types. The token-based values are computed by weighing in lemma frequencies.

Lemma		POS	Gloss	Annotator Region			Diff	Class
				Egypt A1	Levant A2	Gulf A3		
فَرُّوج	*far∼uwj*	noun	chicken	5	1	5	4	food
كُبَّة	*kub∼aħ*	noun	kibbeh	5	1	3	4	food
مِشّ	*miš∼*	noun	fermented salty cheese	1	5	5	4	food
تِياترُو	*tiyAtruw*	noun	theater	1	5	5	4	location
ضَيْعَة	*Dayçaħ*	noun	village	5	1	4	4	location
رِيال	*riyAl*	noun	Riyal	4	4	1	3	money
لِيرَة	*liyraħ*	noun	Lira	4	1	3	3	money
شاوِيش	*šAwiyš*	noun	police sergeant	1	4	4	3	social role
مُطَوَّع	*muTaw∼aç*	noun	mutawwa (religious police)	5	5	1	4	social role
حَنْطُور	*HanTuwr*	noun	covered horse carriage	2	3	4	2	tools
أَجَنْدَة	*Âajandaħ*	noun	agenda/schedule	1	1	4	3	tools
حَلَّة	*Hal∼aħ*	noun	cooking pot	1	4	4	3	tools
مازُوت	*mAzuwt*	noun	diesel oil	4	2	5	3	tools
دَلَّة	*dal∼aħ*	noun	coffee pot	5	4	1	4	tools
عِقال	*çiqAl*	noun	headband	5	3	1	4	tools
بُكْلَة	*buklaħ*	noun	clasp	5	1	1	4	tools
بَغَى	*baγaý*	verb	want/desire	4	4	1	3	verb
ساب	*sAb*	verb	flow/neglect	1	5	2	4	verb
كَشَّر	*kaš∼ar*	verb	scowl	1	3	5	4	verb

Table 7: Examples of disagreements among the three annotators organized in classes reflecting local dialectal influences.

6. Results and Discussion

6.1. Manual Annotation Patterns

Table 4 presents two measures of annotation agreement (accuracy and correlation) among our three annotators, labeled A1 (Egypt), A2 (Syria/Levant), and A3 (Saudi Arabia/Gulf). The average accuracy (ratio of exact match in level annotation) over all pairs of annotators is 54.4%. The average correlation is 85.2%. The large difference between the two values suggest that most of the differences are minor, e.g., Level I for Level II. Moving forward, we use the rounded average, henceforth *average*, annotation in discussing the level annotations unless otherwise noted. Table 5 presents the distributions of the various average levels in our lexicon. As expected, levels I, II and III are smaller than levels IV and V in terms of unique type count, but larger in terms of token count.

The average accuracy and correlation between the different annotators and the average annotation are 76.4% and 93.0%, respectively. This again suggests that the differences are minor. For reference, the accuracy and correlation of the automatic frequency based pre-annotation against the average annotation are 40.5%, and 44.1%, respectively. This indicates that the annotators did not rely on or always agree with the automatic frequency based annotation.

Next, we discuss the disagreement patterns in detail and then move on to discussing the relationship between human readability annotation and frequency.

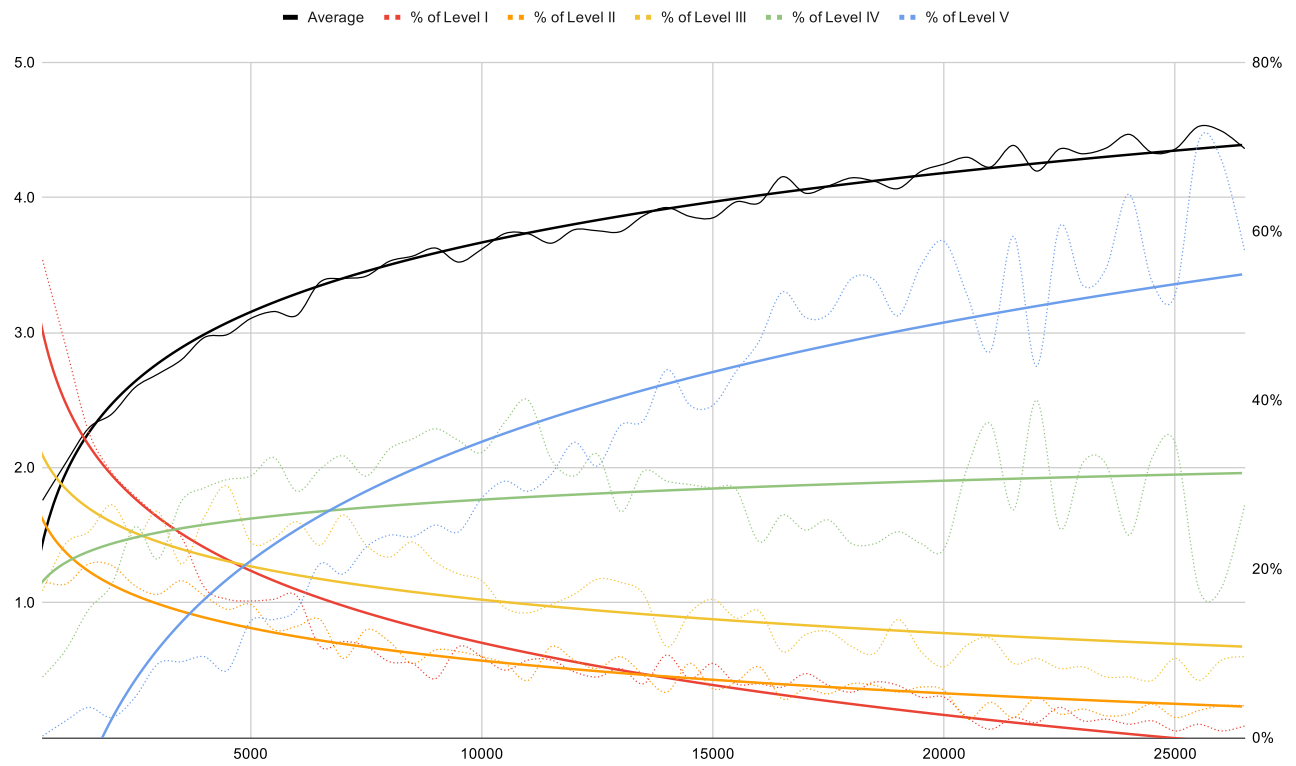

Figure 1: Comparison of annotation distributions over 53 500-lemma adjacent chunks. The thin black line presents the average readability level (left axis 1.0 to 5.0 corresponding to Level I to Level V). The thick black line is its logarithmic trend line. The dotted colored lines represent the percentages of each of the five levels in the 500-lemma chunks (against the right axis); and the colored thick lines are their corresponding logarithmic trend lines.

6.2. Annotation Disagreements

Table 6 presents a high level view of the overall degree and weight of disagreements among our annotators. The table presents disagreements in terms of *Level Distance*, defined as the distance between the highest and lowest levels assigned to any lemma. A Level Distance of value 0 means the three annotators are in full agreement; while a Level Distance of value 2 can be the result of a (Level V, Level IV, Level III) annotation triplet from the three annotators. The first example in Table 7, فَرُّوج *far∼uwj* 'chicken', has a Level Distance of 4. The table shows that in type space, 98.1% of the annotations are either exactly matching or within a Level Distance of 1. When accounting for the frequency of the lemmas (i.e., token space), that number goes to 99.8%. Exact agreement among all three annotators happens almost one-third of the types and almost two-fifths of the tokens. This high agreement gives us great confidence that the annotation task is reasonable and replicable. Table 7 presents 20 examples of the major disagreements (less than 1% of types). All of them are connected to regional variations, where a commonly used term in one region's dialect happens to be the same as the MSA form thus giving it a lower readability level in that region.

6.3. On Frequency and Readability

Figure 1 presents the average readability level (left axis 1.0 to 5.0 corresponding to Level I to Level V) for independent adjacent chunks of 500 lemmas ordered by frequency (thin black line) with their logarithmic trend line (thick black). The intuition that frequency correlates with readability is substantiated to a degree: the trend starts closer to Level I and rising monotonically towards Level V.

Figure 1 also includes five colored data points representing the percentages of each of the five levels in the 500-lemma chunks (against the right axis). At any vertical slice of the figure, we can see the average level for the 500-lemma chunk (in black) and the percentages of the five levels within it. The actual values are in dotted thin lines, while the logarithmic trend is in thick lines. Here again, the frequency-readability correlation is partially substantiated: Level I's percentage is highest among high frequency chunks, and it teeters near the low frequency chunks. Level V has a mirror image distribution. Levels III and IV show a similar pattern to Levels I and V, respectively, although less pronounced. Level II shadows Level III and neither is ever the dominant level in any of the 500 lemma chunks.

It can be concluded from this analysis that while generally

Lemma		POS	Gloss	Frequency	Rank	Level
فِي	*fiy*	prep	in	335409	1	I
أَنّ	*Âan∼*	conj sub	that	181283	3	II
الَّذِي	*Al∼aðiy*	pron rel	which/who/whom	94906	9	III
جَدِيد	*jadiyd*	adj	new/modern	10729	64	I
باب	*bAb*	noun	door/gate	9718	81	I
عامّ	*ςAm∼*	adj	general/common/public	8021	111	III
بَلَغ	*balag*	verb	reach/attain	6324	159	III
إِجْراء	*ĂijrA'*	noun	measures/steps	3306	375	IV
نَوَوِيّ	*nawawiy∼*	adj	nuclear/atomic/nucleic	3201	391	IV
اِقْتِراب	*AiqtirAb*	noun	approach/approximation	264	3,706	II
فِيدْيُو	*fiydyuw*	noun	video	262	3,727	I
سُخْرِيَّة	*suxriy∼aħ*	noun	sarcasm/ridicule	261	3,739	IV
مُمْتاز	*mumtAz*	adj	excellent	260	3,750	I
كادِر	*kAdir*	noun	cadres/staff	259	3,761	V
خاطِئ	*xATiŷ*	adj	mistaken/at fault	255	3,804	I
أَوْجَد	*Âawjad*	verb	find/obtain	126	6,025	III
تاه	*tAh*	verb	go astray/get lost	126	6,025	II
نُكْتَة	*nuktaħ*	noun	joke/wisecrack	126	6,025	I
مُرِيح	*muriyH*	adj	soothing/restful/comfortable	125	6,052	I
تَغَذَىْ	*taγaðaý*	verb	be fed/be nourished	51	9,789	I
غَزال	*γazAl*	noun	gazelle	51	9,789	I
مَبْعَث	*mabςaθ*	noun	cause/factor	51	9,789	V
مُضْطَجَع	*muĎTajaς*	noun	couch	51	9,789	IV
حَلَف	*Halaf*	verb	swear/take an oath	21	14,136	I
زُجّ	*zuj∼*	noun	ferrule/arrowhead	21	14,023	V
فَجَر	*fajar*	verb	live immorally	21	14,023	V
مُؤْذِي	*muŵðiy*	noun	harmful/offensive	21	14,023	I
ثَكَل	*θakal*	noun	bereavement	20	14,252	V
زُكام	*zukAm*	noun	common cold	20	14,252	I
شِيشَة	*šiyšaħ*	noun	hookah/sheesha	20	14,252	II
مَهْدُوم	*mahduwm*	adj	razed/demolished	6	19,321	II
نانُوتِكْنُولُوجِيّ	*nAnuwtiknuwluwjiy∼*	adj	nanotechnology	6	19,321	V

Table 8: A listing of examples demonstrating the high degree of variation between frequency-based rank and readability levels.

speaking there is some correlation between frequency and readability level, there is a limit to how this information can be used to determine the readability level as judged by a human. The high degree of fluctuation in the level percentages from one chunk to an adjacent chunk further support this analysis. Table 8 presents specific examples from different frequency/rank slices showing the wide variety in average levels for lemmas in very similar frequency ranks.

7. Conclusions and Future Work

We presented our effort to create a large-scale 26,000-lemma leveled readability lexicon for Standard Arabic. The lexicon was manually annotated in triplicate by language professionals from three different regions in the Arab World. Comparing human judgements on lemmas with their frequencies provided good insights in the benefits and pitfalls of frequency-based readability approaches. The lexicon will be publicly available from the website of the Com-

putational Approaches to Modeling Language (CAMeL) Lab (http://resources.camel-lab.com/).

Our next steps will be to use the lexicon as part of the SAMER project to help with automatic readability identification and to guide manual simplification. In the future, we plan to expand the coverage of the lexicon by annotating more lemmas extracted from other genres of text, or identified directly in the SAMA morphological analysis database (Maamouri et al., 2010) used in the MADAMIRA disambiguation system (Pasha et al., 2014). We also plan to expand this effort to cover Arabic dialects following the effort by Bouamor et al. (2018) on the MADAR project; and target non-native speakers of Arabic (Saddiki et al., 2015; Saddiki et al., 2018).

Acknowledgements

The work on this project is funded by a New York University Abu Dhabi Research Enhancement Fund grant. We would like to thank Ramy Eskander and his team of annotators at Ramitechs.com for their valuable help, productive discussions and feedback throughout the annotation process.

8. Bibliographical References

Al-Dawsari, M. (2004). The assessment of readability books content (boys-girls) of the first grade of intermediate school according to readability standards. *Sultan Qaboos University, Muscat*.

Al-Khalifa, H. S. and Al-Ajlan, A. A. (2010). Automatic readability measurements of the Arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35(2 C):103–124.

Al Khalil, M., Habash, N., and Saddiki, H. (2017). Simplification of Arabic masterpieces for extensive reading: A project overview. *Procedia Computer Science*, 117:192–198.

Al Khalil, M., Saddiki, H., Habash, N., and Alfalasi, L. (2018). A Leveled Reading Corpus of Modern Standard Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Beech, L. W. (2011). *240 Vocabulary Words Kids Need to Know (Series)*. Scholastic.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Brooke, J., Tsang, V., Jacob, D., Shein, F., and Hirst, G. (2012). Building readability lexicons with unannotated corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 33–39. Association for Computational Linguistics.

Buckwalter, T. and Parkinson, D. (2014). *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge.

Ehara, Y. (2018). Building an english vocabulary knowledge dataset of japanese english-as-a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ferguson, C. F. (1959). Diglossia. *Word*, 15(2):325–340.

Forsyth, J. (2014). Automatic readability prediction for modern standard Arabic. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*.

Fountas, I. C. and Pinnell, G. S. (2006). *Leveled books (k-8): Matching texts to readers for effective teaching*. Heinemann Educational Books.

Fountas, I. C. and Pinnell, G. S. (2017). *Guided reading: Responsive teaching across the grades*. Heinemann.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher*, 56(3):286–291, 11. Name - International Reading Association; Random House Inc; Copyright - Copyright International Reading Association Nov 2002; Last updated - 2019-11-23; CODEN - REDTAH; SubjectsTermNotLitGenreText - United States–US; New York.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Harb, M. (2019). Hanada's text leveling system (htls) from text engagement to text engagingness. *Academic Journal of Interdisciplinary Studies*, 8(2):272–276.

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163, Mar.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., and Kulick, S. (2010). Standard Arabic morphological analyzer (sama) version 3.1. *Linguistic Data Consortium, Catalog No.: LDC2010L01*.

Parker, R., Graff, D., Chen, K., Kong, J., and Maeda, K. (2011). Arabic gigaword fifth edition ldc2011t11. *Philadelphia: Linguistic Data Consortium*.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Saddiki, H., Bouzoubaa, K., and Cavalli-Sforza, V. (2015). Text readability for Arabic as a foreign language. In *Proceedings of the International Conference of Computer*

Systems and Applications (AICCSA), pages 1–8, Marrakech, Morocco.

Saddiki, H., Habash, N., Cavalli-Sforza, V., and Al Khalil, M. (2018). Feature optimization for predicting readability of arabic l1 and l2. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–29.

Sinclair, J. et al. (2003). *Collins COBUILD advanced learner's English dictionary*. Harper Collins Publishers.

Teubert, W. and Cermáková, A. (2007). *Corpus linguistics: A short introduction*. Continuum.

Wynne, M. (2005). *Developing linguistic corpora: A guide to good practice*, volume 92. Oxbow Books Oxford.

Zipes, J. (2013). *Sticks and stones: The troublesome success of children's literature from Slovenly Peter to Harry Potter*. Routledge.