A Leveled Reading Corpus of Modern Standard Arabic

Muhamed Al Khalil,[*] Hind Saddiki,[*†] Nizar Habash,[*] Latifa Alfalasi[‡]
[*]New York University Abu Dhabi, UAE
[†]Mohammed V University in Rabat, Morocco
[‡]Ministry of Education, UAE
{muhamed.alkhalil, hind.saddiki, nizar.habash}@nyu.edu, latifa.alfalasi@moe.gov.ae

Abstract

We present a reading corpus in Modern Standard Arabic to enrich the sparse collection of resources that can be leveraged for educational applications. The corpus consists of textbook material from the curriculum of the United Arab Emirates, spanning all 12 grades (1.4 million tokens) and a collection of 129 unabridged works of fiction (5.6 million tokens) all annotated with reading levels from *Grade 1* to *Post-secondary*. We examine reading progression in terms of lexical coverage, and compare the two sub-corpora (curricular, fiction) to others from clearly established genres (news, legal/diplomatic) to measure representation of their respective genres.

Keywords: Arabic, Corpus, Leveled Reading, Curriculum, Fiction

1. Introduction

Corpora are built for a wide range of purposes such as modeling language use for linguistics research, instructional material for educators, or training data for natural language processing (NLP) applications. Continued efforts in creating such resources are instrumental in furthering research for all application domains of NLP, namely, parsing and part-of-speech (POS) tagging, speech recognition, machine translation, document classification, etc.

Work in NLP for Modern Standard Arabic (MSA) is gaining momentum as more resources and tools are developed (Habash, 2010). Corpus data for MSA has been mostly sourced from the news genre (Zaghouani, 2014), while there are far fewer specialized resources, such as corpora for educational applications (Zaghouani et al., 2014; Alfaifi et al., 2013). As a particular type of educational resource, a level-annotated reading corpus can be leveraged for a multitude of applications: text simplification, automatic readability assessment, computer-assisted language learning, data-driven pedagogy, text genre and register profiling, and so on. Building a corpus of this nature contributes to the variety of resources at our disposal, allowing for research in Arabic NLP to progress in new directions.

In this paper, we present a reading corpus in MSA collected from textbooks of the United Arab Emirates (UAE) curriculum and a collection of 129 unabridged works of fiction. The curriculum texts are labeled with levels from grade 1 to 12 and the fiction texts are at a *Post-secondary* level, i.e., adult-level reading that is accessible to someone after achieving 12th grade reading proficiency. This corpus was created in the context of a project on the Simplification of Arabic Masterpieces for Extensive Reading (SAMER) intended to simplify works of Arabic fiction to a level that is more accessible for school-aged readers (Al Khalil et al., 2017).

The paper is organized as follows. Section 2 presents related work in corpus creation; Section 3 describes the corpus collection and annotation; We analyze the data in Section 4 before stating our conclusions and future work.

2. Background and Related Work

The multi-faceted complexity of MSA makes it a challenging language to tackle in NLP. There is the issue of morphological complexity due to its wide inflectional range and rich composition of clitics (Habash, 2010). Then, there is the challenge of resolving ambiguity due to its writing system with optional diacritics. While it is common to see fully diacritized texts for children, older readers are expected to resolve ambiguity from experience and context in readings where diacritics are often partial or omitted.

Corpora in Arabic have predominantly been collected from news data to serve as general purpose text for NLP applications (Habash, 2010; Zaghouani, 2014). In recent years, the various dialects of Arabic began receiving more attention (Shoufan and Al-Ameri, 2015; Khalifa et al., 2016; Jarrar et al., 2016). Specialized corpora have also been released for various NLP applications such as machine translation (Ziemski et al., 2016), plagiarism detection (Bensalem et al., 2013), sentiment analysis (Abdul-Mageed and Diab, 2012), and error correction (Alfaifi et al., 2013; Zaghouani et al., 2014) to name a few. High-resource languages, on the other hand, have enjoyed a wider variety of specialized corpora, including data for pedagogical and educational applications (Pravec, 2002; Braun et al., 2006; Laufer and Ravenhorst-Kalovski, 2010). Also, recently reignited interest in text readability assessment as a computational task has encouraged more work in the creation of curricular and pedagogical corpora (Collins-Thompson, 2014; François, 2014; Volodina et al., 2014; Zalmout et al., 2016).

Budding research in computational readability for MSA has led to the creation of leveled corpora from curriculum texts. For instance, a corpus of 150 texts from the Saudi Arabian (KSA) curriculum labeled with [easy, intermediate, difficult] (Al-Khalifa and Al-Ajlan, 2010), and a corpus of 1196 texts totaling 400K words from the Jordanian curriculum (Al Tamimi et al., 2014). To the best of our knowledge, a corpus at the scale of the curricular data collected in our work (1.4M tokens) has yet to be released.

Grade 2	الطّالِبَةُ النَّجيبَةُ مَيْثاءُ طالِبَةٌ نَشيطَةٌ، تُطيعُ والِدَيْها، وَتُحافِظُ عَلى صَلاتِها، تَصْحو مِنْ نَوْمِها مُبَكِّرَةً، تَتَناوَلُ فَطورَها، وَتُنَظِّفُ أَسْنانَها، وَتَرْتَدي مَلابِسَ الْمَدْرَسَةِ، تَلْتَقي مَعَ زَميلاتِها مُبْتَسِمَةً، تَجْلِسُ في صَفِّها بِهُدوءٍ وَانْتِباهٍ. *Maitha is a clever hard-working student. She listens to her parents, and keeps her prayers. She wakes up early, eats her breakfast, brushes her teeth, and puts on her school uniform. She greets her classmates with a smile, and sits quietly and attentively in her class.*
Grade 7	شاعَ شِعْرُ الحِكْمَةِ في الأَدَبِ العَرَبِيِّ، وَهُوَ يُوَضِّحُ القِيَمَ وَالمَبادِئَ وَالأَخْلاقَ وَالأَوامِرَ الإِلَهِيَّةَ، وَيُفْصِحُ عَنْ تَجارِبَ وَخِبْراتٍ سابِقَةٍ تُنْقَلُ عَبْرَ الأَجْيالِ، وَتَحْكي قِصَصًا نَتَعَلَّمُ مِنْها المَواعِظَ وَالْعِبَرَ، وَيَرِدُ هذا الشِّعْرُ في بَيتٍ أَوْ أَبْياتٍ أَوْ في قَصائِدَ كامِلَةٍ. *Poetry of wisdom became prevalent in Arabic literature. It is a kind of poetry that clarifies divine commandments, morals, principals, and values. It also discloses and transmits past experiences across generations, telling stories from which we learn lessons and wisdom. This poetry can come in the form of one line, a few lines, or a whole poem.*
Grade 10	وقد حدث يومًا وأنا مدرس في المدرسة الخديوية أن دخلت غرفة الصف فألفيت على مكتبي كل أدوات الرياضة مرصوصة على نحو لا شك أنه متعمد، وكان تلاميذي لا يجهلون كرهي للرياضة، وكنت أنا لا أكتمهم أني أعد نفسي جاهلاً بها، وكان غرضهم من رصّ هذه الأدوات أن يعابثوني عسى أن أثير الضجة التي يشتهونها ولا يفوزون مني بها، ولكني لم أفعل بل اكتفيت بأن دعوت الفراش فحمل هذه الأدوات ووضعها في مكانها ثم بدأت الدرس. *One day when I was teaching at the Khedive School I entered the classroom and found all the mathematics tools lined up purposefully in a pattern. My students were not ignorant of my hate of mathematics, and I never concealed to them that I considered myself ignorant in the field. Their goal was to jest with me so that I make the big fuss they desire but never attain. And I did not; I only called the janitor who carried the tools and put them back in their place; then I started the lesson.*
Novel	ثم أنا أمضي إلى هذه النافذة، فلا أكاد أفتحها حتى تمتلئ نفسي روعةً وجلالًا لهذه الأشجار النائمة، وهذه الأزهار المتأرجة، وهذه الأطيار التي تحلم في ثنايا الغصون، وكل هذا لي ملك خالص لا يشاركني فيه أحد، ولا يزاحمني عليه أحد، أستطيع أن أعبث به إن شئت، ومتى شئت، وكيف شئت، لا يسألني أحد عما أفعل! *Then I went to this window, and no sooner had I opened it than my soul filled up with majestic awe of these slumbering trees, these fragrant flowers, and these birds dreaming in the nooks of branches. This is all mine, I share it with no one, and no one crowds me for it. I can toy with it if I wish, whenever I wish, however I wish, and I answer to no one about it.*

Figure 1: Samples of reading text from different levels of the corpus

3. Corpus Description

In this section, we discuss the variety observed in the corpus with illustrative examples. We then document the data collection and processing efforts, and present descriptive statistics and details of the text annotations.

3.1. Text Varieties in the Corpus

This corpus consists of two sub-corpora: a diverse body of texts combining the full UAE curriculum, and a body of fiction texts derived from the Hindawi collection. A curricular sub-corpus, especially one covering different subjects, includes almost all kinds of texts: expository, transactional, procedural, argumentative, informative, narrative, literary, scientific, etc. A fiction-based corpus provides a special register of the language, and has been used to study both general linguistic features and more specific stylistic features (Biber, 2011). The key difference between the two bodies of texts is that while the curricular sub-corpus is focused on information delivery and educational growth assessment, the second is occupied with the literary aesthetic and is thus pleasantly blasé about teaching and learning. Between the two, however, one can capture the full spectrum of written language phenomena that a school-educated Arabic-speaker would experience, allowing the corpus to qualify as a general corpus (McEnery et al., 2006).

Illustrative Examples To give samples of the texts included in each level, we chose four short pieces that best reflect the nature and variety of those texts. For the first three pieces, each piece comes from a grade that tends to be midrange in the grades of that level; with the fourth piece coming from, perhaps, the best well-known novel in that literary collection. The first textual piece comes from the 2nd grade and it describes a person and her daily habits. It is fully diacritized. The text is – as is expected in this introductory level – direct, concrete, and less complex. It is generally one-dimensional comprised mainly of short declarative sentences. The second piece comes from the 7th grade and it describes a genre of poetry in Arabic. It is also fully diacritized. It is expository, conceptual, and meta-lingual (using language about language). It is more complex in terms of both vocabulary and sentence structure and length. The third piece comes from the 10th grade and it is excerpted from a memoir. It is not diacritized. It is story-like told in the first person. Its style is narrative made of several complex sentences and expressions. The fourth piece comes from a well-known novel in the Hindawi collection, *The Call of the Curlew* by Taha Hussein.[1] It is not diacritized. It is an introspective musing by the omnipresent narrator. It is made of run-on complex sentences with more abstract vocabulary. It has a clear literary style, typically found in fiction: mixing the concrete with the poetic to produce a pleasant emotive sense.

3.2. Data Gathering and Extraction

Curriculum The curriculum textbooks were obtained as `InDesign`[2] files spanning 12 grades (Elementary Grade 1 to Secondary Grade 12) and three subjects (Arabic lan-

[1] Accessible at `http://www.hindawi.org/books/13052715/`

[2] Adobe InDesign desktop publishing software `http://www.adobe.com/products/indesign.html`

guage, social studies, Islamic studies). We converted each InDesign file into an intermediary HTML format then into raw UTF-8 text format. The curriculum files were obtained from the UAE Ministry of Education.[3]

Fiction We collected 129 works of fiction available in the public domain from the online catalog of the Hindawi Foundation.[4] We downloaded the individual e-book files in `.epub`[5] format and converted them to an intermediary HTML format then into raw UTF-8 text format.

3.3. Building the Corpus

For the curricular sub-corpus, all data pertaining to a given grade is labeled with its corresponding grade level going from primary grade level 1 to secondary (high school) grade level 12. Additional annotation for subject (Arabic Language, Social Studies, Islamic Studies), term (1st, 2nd, sometimes 3rd) and unit number (each unit is marked in the textbook's table of contents as a set of lessons under a theme with specific learning objectives).

Books in the fiction sub-corpus are all labeled at the *Post-secondary* level indicating they are accessible to readers having achieved reading proficiency of the full 12-grade curriculum. Each book has a unique ID tied to its meta-information (author and title) as well as manually annotated year of copyright and publication.

We annotated each token in the corpus with morphological information including lemma, POS using the MADAMIRA tool for morphological disambiguation (Pasha et al., 2014). We expect a drop in accuracy on this genre of text given that MADAMIRA has been trained on news data. An in-house evaluation on an example of literary fiction text[6] shows a drop of 4% absolute in word analysis performance for choice of lemma and POS. While lower than on news text, the performance is still at a high 92%.

Table 1 presents summary statistics on all the collected text, differentiating the curricular and fiction sub-corpora. The *Sentences* represent complete lines of text. Words counts in the text are reported by whitespace-based *tokens* (including punctuation and numbers as separate words). To get a sense of lexical richness, we also compute unique tokens, i.e., *types*, and unique *lemmas* for the word forms occurring in the text.

The learner's vocabulary after completing *Grade 12* education reaches 22K distinct lemmas (closer to 18K when proper nouns, punctuation and digits are excluded). When compared to English, Nation (2013) estimates a learner to require a vocabulary of 15K to 20K words in order to optimally read and comprehend text with no obstruction from unknown vocabulary. However, we bear in mind that vocabulary is not the only indicator of level. One must take into account how common or specialized the vocabulary is, semantic fields, discourse, style, and so on to fully assess reading level beyond word frequency.

Grade Level	Sentences	Tokens	Types	Lemmas
1	10,860	57,409	9,193	4,391
2	8,580	65,014	10,142	4,390
3	10,966	87,460	13,692	5,531
4	11,597	108,946	18,291	7,059
5	8,833	86,096	15,727	6,453
6	9,710	108,557	19,862	7,937
7	12,112	116,176	21,489	8,466
8	11,619	118,288	21,092	8,175
9	13,176	172,175	25,547	9,850
10	11,518	171,340	27,003	10,196
11	12,253	157,453	27,827	10,364
12	10,812	165,791	31,323	11,732
Curriculum (All)	132,036	1,414,705	89,446	22,143
Fiction (avg. per book)	1,279	43,367	10,584	4,719
Fiction (All)	165,005	5,594,310	261,920	44,498

Table 1: Summary statistics for the leveled reading corpus

4. Quantitative Corpus Analysis

We describe a preliminary exploration of the corpus by conducting two studies: lexical coverage progression over the curriculum as a measure of the grade-leveling scheme's validity, and a similarity comparison with other well-known corpora in the news genre (Gigaword (Parker et al., 2011)) and the legal/diplomatic genre (UN Corpus (Ziemski et al., 2016)) to establish curricular and fiction texts as distinct genres.

All studies in Section 4 are performed on *content* tokens only. In other words, we exclude punctuation and digits (non-content tokens) from our calculations, which make up 18% and 15% of all tokens in the curricular and fiction sub-corpora, respectively. We also discount any content words not in the MADAMIRA vocabulary database, i.e., out-of-vocabulary tokens, which amount to 0.96% of all content tokens in the curricular sub-corpus and 2.2% of all content tokens in the fiction sub-corpus.

Level	Lexical Coverage
1	n/a
2	93.6%
3	95.3%
4	96.1%
5	97.2%
6	97.3%
7	97.6%
8	98.6%
9	98.1%
10	98.5%
11	98.5%
12	99.4%
Post-secondary	97.1%

Table 2: Lexical coverage in levels 1 to 12; Average lexical coverage per book in the post-secondary level

[3]The corpus obtained from the UAE Ministry of Education pertained to the curriculum applied between 2014 and 2016. The current curriculum was designed with a richer selection of literary and informational readings. We look forward to analyzing the current curriculum as part of ongoing collaboration with the UAE Ministry of Education.

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

[5]`http://idpf.org/epub`

[6]Chapter 1 of Ibrahim Alkatib, by Ibrahim Al-Mazini (1889-1949).

4.1. Lexical Coverage

We examine whether the grade-leveling scheme is a valid indication of reading level by measuring lexical coverage. Lexical coverage is defined as follows: a word list is said to provide lexical coverage of 80% of a given text if 80% of all word tokens in said text occur in that word list. When reading a text, the amount of vocabulary familiar to the reader influences comprehension, which raises the question of lexical threshold, i.e., the minimum rate of lexical coverage for reading comprehension. Studies on lexical thresholds for reading set a lexical coverage of 95% as the minimum threshold for *adequate comprehension*[7] and lexical coverage of 98% as the threshold for optimal (unassisted) comprehension. See (Nation, 2006; Laufer and Ravenhorst-Kalovski, 2010) for further details.

Steps for the curricular sub-corpus lexical coverage:

- Selecting a target **Grade$_i$**
- Computing familiar vocabulary from all previous grades [1,i-1] as a list of unique lemmas
- Calculating the total count of tokens in **Grade$_i$** corresponding to lemmas that exist in the list of familiar vocabulary
- Reporting the lexical coverage as the ratio of tokens matching the list over total token count for the target **Grade$_i$**

Steps for the fiction sub-corpus lexical coverage:

- Selecting a target **Book$_i$**
- Computing familiar vocabulary from all curricular grades [1,12] as a list of unique lemmas
- Calculating the total count of tokens in **Book$_i$** corresponding to lemmas that exist in the list of familiar vocabulary
- Computing the lexical coverage as the ratio of tokens matching the list over total token count for the target **Book$_i$**
- Reporting the lexical coverage as the average of all lexical coverage ratios computed for the 129 books in the fiction sub-corpus[8]

Table 2 presents the results of the study carried out according to the steps described for both sub-corpora. We point out that no lexical coverage is reported for Grade 1. Although vocabulary acquisition does occur prior to Grade 1, our curricular sub-corpus lacks data for the Kindergarten level. We rely on the 95% minimum and 98% optimal thresholds for English as a ballpark estimate, being fully aware that these threshold numbers may vary for MSA and our target readership. We observe a clear progression across the curricular levels and a lexical coverage ratio indicating that the 95% minimum threshold is consistently met while the optimal threshold of 98% is reached starting the 8th Grade, at which time learners are expected to have acquired a much richer vocabulary. The post-secondary lexical coverage of 97.1% suggests that vocabulary acquired from readings in a 12-grade curriculum allows for adequate reading and understanding of a work of fiction.

[7]Usually measured by testing and scoring readers with comprehension questions (Nation, 2006).

[8]Averaging per book is more representative of the lexical coverage required for reading any work of fiction at a post-secondary level.

Gigaword	65.5%		
Curriculum	76.7%	71.0%	
UN	57.3%	68.5%	64.4%
	Fiction	**Gigaword**	**Curriculam**

Table 3: Dice Similarity (1) between corpora of different genres

4.2. Genre Similarity and Difference

A similarity comparison of our corpus with other established corpora in the news genre (Gigaword (Parker et al., 2011)) and the legal/diplomatic genre (UN Corpus (Ziemski et al., 2016)) can approximate difference in genre, which could potentially establish this corpus as representative of the curricular genre.
We use the Dice Coefficient (1) to compute similarity between pairs of corpora. Given that the curricular sub-corpus is the smallest in size with 1.4M tokens, for comparison we use randomly sampled subsets of nearly 1.4M tokens for each of Gigaword, UN and the Fiction sup-corpus. The similarity is calculated on unique lemma sets *A* and *B* for each comparison pair.

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|} \quad (1)$$

We report the results of pairwise Dice similarity comparisons for the four corpora in Table 3. The UN corpus using specialized legal/diplomatic language behaves as expected, being the least similar to other genres. It presents with the lowest similarity score of 57.3% in the UN-Fiction comparison, given that legal or administrative language is quite different from literary writing. We note with interest the Gigaword-Fiction 65.5% similarity. This comparison of two corpora from clearly distinct genres (news and literary texts) gives us a better sense of what 65% similarity or rather 35% difference means between two clearly established genres. The 23%, 29% and 36% respective difference in Curriculum (-Fiction, -Gigaword, -UN) comparisons could indicate sufficient distance between the curricular corpus and the others for it to be representative of its own curricular/educational genre.

5. Conclusion and Future Work

We presented a corpus for reading in MSA that was collected from curricular texts (1.4M tokens) and works of fiction (5.6M tokens). The corpus was annotated with reading levels per grade for the curricular sub-corpus and a post-secondary level for the collection of novels in the fiction sub-corpus. We assessed the validity of a grade-leveling scheme using progression of lexical coverage over the curriculum. A similarity comparison with other established corpora in the news genre, and the legal/diplomatic genre

could potentially establish this corpus as representative of the curricular or educational genre.

In the future, we plan to use this corpus in modeling levels of reading proficiency to simplify works of fiction in the context of the SAMER project. We also plan on annotating portions of the corpus with morphological and syntactic information. It is our intent to work on releasing this data in full-text format and/or as an n-gram frequency dataset to be exploited in training NLP tools for any number of educational applications.

6. Acknowledgements

The work on this project is funded by a New York University Abu Dhabi Research Enhancement Fund grant. We would like to express our thanks to the UAE Ministry of Education for providing us with the curriculum materials, which are essential for this research project. We also thank Bassel Musfi for annotating the fiction sub-corpus with copyright and publication year information.

7. References

Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3907–3914.

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 Tamimi, A. K., Jaradat, M., Al-Jarrah, N., and Ghanem, S. (2014). Aari: automatic Arabic readability index. *Int. Arab J. Inf. Technol.*, 11(4):370–378.

Alfaifi, A., Atwell, E., and Abuhakema, G. (2013). Error annotation of the Arabic learner corpus. In *Language Processing and Knowledge in the Web*, pages 14–22. Springer.

Bensalem, I., Rosso, P., and Chikhi, S. (2013). A new corpus for the evaluation of Arabic intrinsic plagiarism detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 53–58. Springer.

Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature*, 1(1):15–23.

Braun, S., Kohn, K., et al. (2006). Corpus technology and language pedagogy: New resources, new tools, new methods. English corpus linguistics vol 3.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

François, T. (2014). An analysis of a French as a foreign language corpus for readability assessment. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107. Linköping University Electronic Press.

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

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2016). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A Large Scale Corpus of Gulf Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Laufer, B. and Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language*, 22(1):15.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1):59–82.

Nation, I. S. (2013). *Learning Vocabulary in Another Language Google eBook*. Cambridge University Press.

Parker, R., Graff, D., Chen, K., Kong, J., and Maeda, K. (2011). Arabic Gigaword Fifth Edition. LDC catalog number No. LDC2011T11, ISBN 1-58563-595-2.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. M. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *LREC 2014, Reykjavik, Iceland*.

Pravec, N. A. (2002). Survey of learner corpora. *ICAME journal*, 26(1):8–14.

Shoufan, A. and Al-Ameri, S. (2015). Natural language processing for dialectical Arabic: A survey. In *ANLP Workshop 2015*, page 36.

Volodina, E., Pilán, I., Borin, L., and Tiedemann, T. L. (2014). A flexible language learning platform based on language resources and web services. In *LREC*, pages 3973–3978.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large scale Arabic error annotation: Guidelines and framework. In *LREC 2014, Reykjavik, Iceland*.

Zaghouani, W. (2014). Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC*, pages 1–8.

Zalmout, N., Saddiki, H., and Habash, N. (2016). Analysis of foreign language teaching methods: An automatic readability approach. *NLPTEA 2016*, page 122.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1. 0. In *LREC*.