The SAMER Arabic Text Simplification Corpus

Bashar Alhafni, Reem Hazim, Juan Piñeros Liberato, Muhamed Al Khalil, Nizar Habash
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
{alhafni,rh3015,juanpl,muhamed.alkhalil,nizar.habash}@nyu.edu

Abstract

We present the **SAMER Corpus**, the first manually annotated Arabic parallel corpus for text simplification targeting school-aged learners. Our corpus comprises texts of 159K words selected from 15 publicly available Arabic fiction novels most of which were published between 1865 and 1955. Our corpus includes readability level annotations at both the document and word levels, as well as two simplified parallel versions for each text targeting learners at two different readability levels. We describe the corpus selection process, and outline the guidelines we followed to create the annotations and ensure their quality. Our corpus is publicly available to support and encourage research on Arabic text simplification, Arabic automatic readability assessment, and the development of Arabic pedagogical language technologies.

Keywords: Arabic, Text Simplification, Readability

1. Introduction

Text simplification aims to reduce the complexity of a text while maintaining the overall grammaticality and core content. This is achieved through a series of different rewriting transformations at both the lexical and syntactic levels. Having simplified versions of texts has many benefits to users with cognitive and reading disorders (Carroll et al., 1998; Rello et al., 2013; Evans et al., 2014), second language learners (Paetzold and Specia, 2016b), and native speakers with low literacy levels (Candido et al., 2009; Watanabe et al., 2009). Text simplification can also be used as a preprocessing step to improve performance on other downstream NLP tasks such as machine translation (Štajner and Popovic, 2016; Hasler et al., 2017) and summarization (Silveira and Branco, 2012). Simplifying text can be achieved in multiple ways, depending on the target audience: for example, second-language learners and school-aged learners might struggle with texts containing different vocabulary items that go beyond their *respective* language proficiency levels. Yet, research in text simplification has mostly focused on developing models that produce a single simplification for a given input without the possibility of adapting to different users' needs. Moreover, studies on text simplification are heavily focused on English due to the availability of large parallel simplification corpora (Alva-Manchego et al., 2020). For other languages, data is limited in terms of size and domain. And when it comes to morphologically rich languages, particularly Arabic, we are not aware of any manually annotated publicly available datasets for text simplification.

In this paper, we present the **SAMER Corpus**, the first manually annotated Arabic parallel corpus for text simplification targeting school-aged learners. Our corpus comprises texts of 159K words selected from 15 publicly available Arabic fiction novels, 14 of which were published between 1865 and 1955, and one famous philosophical novel written in the 12th century. We focus on lexical simplification, i.e., replacing complex words in a given text with simpler alternatives of equivalent meaning. We define the text simplification task as paraphrasing into a controlled language with a vocabulary that is anchored in a readability-leveled lexicon. Our corpus includes readability level annotations at both the document and word levels, as well as two simplified parallels for each text, targeting school-aged learners at two different readability levels. We describe the corpus selection process and outline the guidelines we followed to create the annotations and ensure their quality. Our corpus is publicly available to support and encourage research on Arabic text simplification and automatic readability assessment, as well as the development of pedagogical language technologies. Table 1 presents an example from our simplification corpus.

This work is one of the publicly available resources created by the ***Simplification of Arabic Masterpieces for Extensive Reading*** **(SAMER)** project (Al Khalil et al., 2017),[1] which includes a readability leveled lexicon (Al Khalil et al., 2018; Jiang et al., 2020), and a Google Doc add-on (Hazim et al., 2022).

Next, we discuss related work and basic Arabic linguistic facts. In Section 4, we introduce our corpus, and describe its selection and annotation process. Section 5 presents the corpus statistics and discusses its simplification patterns.

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

LREC-COLING 2024, pages 16079–16093
20-25 May, 2024.

Original Level 5	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب **مضارعة**$_{1}$ أولئك الشعوب الذين يبادرون إلى **شن**$_{2}$ الغارات وفتك بعضهم بعضا على أقل **أرب**$_{3}$ **لا يعتد**$_{4}$ به، أو أدنى خرافة لا بيت لها في رقعة التمدن؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one **emulate**$_{1}$ those people who take to **launching**$_{2}$ raids and slaughtering each other for the triflest most **dismissible**$_{4}$ of **wants**$_{3}$, or for the weakest of superstitions that has no refuge in civilization's domain.
Level 4	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب **أن نشبه**$_{1}$ أولئك الشعوب الذين **يبادرون**$_{5}$ إلى **بدء**$_{2}$ **الغارات**$_{6}$ **وفتك**$_{7}$ بعضهم بعضا على أقل **هدف**$_{3}$ **لا يؤخذ**$_{4}$ به، أو أدنى خرافة لا بيت لها في **رقعة**$_{8}$ **التمدن**$_{9}$؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one **imitate**$_{1}$ those people who **take to**$_{5}$ **starting**$_{2}$ **raids**$_{6}$ **and slaughtering**$_{7}$ each other for the triflest most **negligible**$_{4}$ of **aims**$_{3}$, or for the weakest of superstitions that has no refuge in **civilization's**$_{9}$ **domain**$_{8}$.
Level 3	وليس جميع الحوادث والأحوال تساوي الدم الإنساني الذي لا يوجد أثمن منه، ولا يجب أن نشبه أولئك الشعوب الذين **يسرعون**$_{5}$ إلى بدء **الهجمات**$_{6}$ **وقتل**$_{7}$ بعضهم بعضا على أقل هدف لا يؤخذ به، أو أدنى خرافة لا بيت لها في **منطقة**$_{8}$ **التطور**$_{9}$؛	Not all events and conditions are worth human blood, for there is nothing more valuable than that blood. Nor should one imitate those people who **rush to**$_{5}$ starting **attacks**$_{6}$ **and killing**$_{7}$ each other for the triflest most negligible of aims, or for the weakest of superstitions that has no refuge in the **range**$_{8}$ **of progress**$_{9}$.

Table 1: An example consisting of two sentences (in three punctuated fragments) and its simplified parallels from the Arabic novel "The Forest of Truth" (Marrash, 1865). Level 4 is the simplified version of the original text where all level 5 words (in **red**) are simplified to level 4 or lower according to Al Khalil et al. (2018)'s readability lexicon. Level 3 is the simplified version of the Level 4 text where all level 4 words (in **blue**) are simplified to level 3 or lower. The words that change during the simplification are co-indexed with subscript numbers and carry the same color marking of the level they changed from.

2. Related Work

We first provide an overview of existing work related to general text simplification approaches and datasets, before zooming in on Arabic text simplification specifically.

2.1. Text Simplification Approaches

When simplifying text, different rewriting transformations are performed. Such transformations range from *lexical simplification*, which is the process of replacing complex words or phrases with simpler synonyms, to *syntactic simplification*, which includes splitting or reordering sentences. Most research on text simplification has focused on simplifying individual sentences. This allows for easier curation of data and reduces the complexity of modeling. Several modeling approaches for text simplification have been explored. This includes syntactic simplification, lexical simplification, and end-to-end models that can learn and induce both syntactic and lexical transformations.

Efforts on lexical simplification often involve four subtasks: Complex Word Identification, Substitution Generation, Substitution Selection, and Substitution Ranking (Shardlow, 2014) with approaches ranging from lexicon-based lookups (Elhadad and Sutaria, 2007; Kajiwara et al., 2013) to statistical machine learning systems (Paetzold and Specia, 2016a; Gooding and Kochmar, 2018, 2019), and more recently, deep learning models (De Hertog and Tack, 2018; Maddela and Xu, 2018; Qiang et al., 2020, 2021; Sheang et al., 2022).

In contrast, efforts for syntactic simplification focused on rule-based systems (Chandrasekar et al., 1996; Gasperin et al., 2009) and statistical machine learning techniques by drawing inspirations from phrase- and tree-based statistical machine translation models (Specia, 2010; Zhu et al., 2010; Wubben et al., 2012).

Finally, end-to-end text simplification approaches are the dominant paradigm in the literature. End-to-end models can perform multiple simplification transformations simultaneously, while learning very specific and complex rewriting patterns. The majority of approaches treat text simplification as a monolingual machine translation task, where both statistical (Coster and Kauchak, 2011a; Wubben et al., 2012; Xu et al., 2016) and neural machine translation models (Nisioi et al., 2017; Zhang and Lapata, 2017; Štajner and Nisioi, 2018; Vu et al., 2018; Guo et al., 2018; Zhao et al., 2018; Surya et al., 2019; Martin et al., 2020; Maddela et al., 2021) were explored. These models require large amounts of parallel training data and provide little control or adaptability to different aspects of simplification, which inhibits interpretability and explainability. Moreover, these models are typically slow as they employ autoregressive decoders, i.e., output texts are generated in a sequential, non-parallel fashion. To address some of these limitations, sequence labeling and edit-based models were explored (Alva-Manchego et al., 2017; Omelianchuk et al., 2021).

2.2. Text Simplification Datasets

Most of the recent advancements in text simplification have focused on English, which is attributed to the availability of large parallel datasets. Most of the English datasets (Zhu et al., 2010; Coster and Kauchak, 2011b; Woodsend and Lapata, 2011; Kauchak, 2013; Hwang et al., 2015; Kajiwara and Komachi, 2016; Zhang and Lapata, 2017) were created by automatically aligning sentences from English Wikipedia and Simple English Wikipedia, a simplified version of English Wikipedia that is primarily aimed at English learners, but which can also be beneficial for students, children, and adults with learning difficulties.

Although the large scale and availability of Wikipedia-based corpora is a strong asset to build simplification models, studies have shown that Wikipedia-based text simplification corpora are limited in various ways, including the presence of noisy instances caused by misalignments and a lack of variety in simplification transformations (Yasseri et al., 2012). To address these limitations, several manually annotated datasets were introduced such as the Newsela Corpus (Xu et al., 2015), TurkCorpus (Xu et al., 2016), HSplit (Sulem et al., 2018), SimPA (Scarton et al., 2018), and OneStopEnglish (Vajjala and Lučić, 2018).

While the most popular (and generally larger) resources available for simplification are in English, there are some resources that have been built for other languages such as Basque (Gonzalez-Dios et al., 2014), Brazilian-Portuguese (de Medeiros Caseli et al., 2009; Aluísio and Gasperin, 2010), Danish (Klerke and Søgaard, 2012), French (Grabar and Cardon, 2018; Gala et al., 2020; Cardon and Grabar, 2020), German (Klaper et al., 2013; Battisti et al., 2020; Aumiller and Gertz, 2022), Italian (Brunato et al., 2015; Tonelli et al., 2016; Miliani et al., 2022), Japanese (Goto et al., 2015; Hayakawa et al., 2022), Spanish (Bott et al., 2012; Xu et al., 2015; Saggion et al., 2015), Russian (Dmitrieva and Tiedemann, 2021; Sakhovskiy et al., 2021), and Urdu (Qasmi et al., 2020).

Level	Grade	Age	Examples
L1	1	6	بَيْت، شَجَرَة، صَنَعَ، لكِن house, tree, to make, but
L2	2-3	7-8	جَزيرة، داكِن، خَدَعَ، إذا island, dark, to cheat, if
L3	4-5	9-10	مُتْحَف، رِئة، لَدى، كَيْ museum, lung, with, for
L4	6-8	11-14	اِقْتِصاد، طُمَأنينة، راقِي، نَكَثَ economy, tranquility, sophisticated, to breach
L5	9+	15+	قَسْطَرة، هَيْضة، لَوْذَع، شُعَبيّ catheterization, cholera, witty, bronchial

Table 2: The five readability levels, their grade equivalencies, and lemma and English gloss examples, abridged from Al Khalil et al. (2020).

2.3. Arabic Text Simplification

While there are some limited efforts to publishing simplified and abridged texts in Arabic, the only simplification resource that was used in NLP, to our knowledge, is the simplified version of “Saaq al-Bambuu (The Bamboo Stalk)” (Sanusi, 2012), an internationally acclaimed Arabic novel, that has been rewritten for Arabic-as-a-second-language learners (Familiar and Assaf, 2017). Khallaf et al. (2022) automatically aligned sentences from “Saaq al-Bambuu” and sampled 2,980 parallel sentences from the original and simplified books at two different literacy levels. Unfortunately, due to copyright restrictions, the corpus is not publicly available.

More recently, there have been grassroots efforts to create Arabic text simplification resources (Al Khalil et al., 2017, 2018, 2020; Jiang et al., 2020; Hazim et al., 2022). As part of the SAMER Project, Al Khalil et al. (2020) developed a 26K-lemma lexicon with a five-level readability scale, later extended to 40K lemmas (Jiang et al., 2020). The levels range from **L1** (Low Difficulty/Easy Readability) to **L5** (High Difficulty/Hard Readability). See examples in Table 2. We use this lexicon as our main reference for readability leveling when creating our corpus to ensure that our simplified texts are appropriate for the audience we are targeting. Hazim et al. (2022) created a Google Docs add-on for automatic Arabic word-level readability visualization, which includes a lemmatization component that is connected to the five-level readability lexicon (Al Khalil et al., 2020) and Arabic WordNet-based substitution suggestions (Black et al., 2006). The add-on enables users to edit texts easily based on a specific target readability level. We use the add-on as our main annotation tool to enable human annotators to identify text readability levels and to simplify texts in a controlled setting.

3. Arabic Linguistic Facts

Arabic is a morphologically rich language that inflects for gender, number, person, case, state, aspect, mood and voice, in addition to numerous attachable clitics such as prepositions, particles, and pronouns (Habash, 2010). This results in a large number of forms for any particular word, with different morpho-syntactic restrictions. In addition to its morphological richness, Arabic is orthographically ambiguous and uses diacritics to specify short vowels and consonantal doubling. These diacritics are optional and often omitted, leaving readers to decipher words using contextual and templatic morphology clues. Orthographic ambiguity and morphological richness interact heavily with each other. For instance, the word درسها *drshA* has different readings with varying analyses including دَرَّسَها *dar∼asa+hA* 'he taught her', دَرَسَها *darasa+hA* 'he studied it', and دَرْسُها *darsu+hA* 'her lesson'. Moreover, these different readings have three unique lemmas (lexical entries) that abstract away from the various inflections: دَرَّس *dar∼as 'taught'*, دَرَس *daras 'studied'*, and دَرْس *dars 'lesson'*. This issue highlights the complexity of lexical simplification in Arabic, which cannot be accomplished through a simple word dictionary lookup.

The SAMER project lexicon (Al Khalil et al., 2020) discussed above anchors readability at the lemma representation of the words. There are publicly available Arabic morphological disambiguation and lemmatization tools that diacritize Arabic text and map each word to a predicted lemma (Pasha et al., 2014; Obeid et al., 2022).

In this paper we focus on Modern Standard Arabic (MSA) and do not discuss dialectal Arabic variants, which are not typically used in high literature.

4. The SAMER Arabic Text Simplification Corpus

4.1. Corpus Selection

In making the specific selection of texts to annotate, we aimed to cover Arabic fiction novels from a large historical span with high readability levels (i.e., hard to read) targeted toward proficient Arabic readers. But most importantly, we wanted the texts to be publicly available (out of copyright or under open licenses). We identified and selected 15 Arabic fiction novels that match these requirements from the online catalog of the Hindawi Foundation.[2] Most of the novels were published between 1865 and 1955 and one philosophical novel was from the 12th century. From each novel, we extracted the first ∼10K words based on chapter boundaries, and we ended up with ∼159K words in total. To make the annotation task easier, we further segmented the chapters based on paragraph boundaries if they consisted of more than 1,500 words. This resulted in 4,289 paragraphs. We were restricted by an annotation budget that affected how many novels we could work with. There were many interesting options that we decided to leave to future annotation follow-up projects. Table 3 presents the list of the selected books.

4.2. Corpus Annotation

Our goal is to simplify the Arabic fiction novels we selected so that they can be targeted toward school-age learners. Given Arabic morphological richness, we consider the lexical and syntactic aspects of text simplification to be independent and complementary to each other. In our work, we focus on lexical simplification for a number of reasons. First, research has shown that lexical simplification improves text readability (Leroy et al., 2012), benefiting those with lower literacy levels and non-expert readers (Xu et al., 2015). Second, we want to create our dataset in a controlled way to avoid context inconsistencies that may result from changing the syntactic structure of Arabic text. This process ensures that the simplified text is indeed of a lower complexity while being semantically equivalent to the original, and grammatically correct.

To ensure that the simplified texts are indeed appropriate for our target audience throughout the annotation process, we use the five readability levels that were defined by Al Khalil et al. (2020) and exemplified in Table 2. For the purpose of the corpus annotation, we consider the document readability level to be equal to the highest readability level found among the words in the document. Our focus is on the needed competence level to easily read a document rather than the use of documents as language-learning artifacts. Based on this and given the nature of the documents we selected, all documents will have a readability of **Level 5** (although some sentences in them may be of lower/easier readability levels). We focus on producing two simplified versions for each document targeting **Level 4** (grades 6–8) and **Level 3** (grades 4-5), respectively.

4.3. Annotation Interface

Three professional female computational linguists, all of whom are native speakers of Arabic, were hired through a linguistic annotation firm to complete the task.[3] The annotators were provided with 146 Google Docs that included the chapters that

[2] http://www.hindawi.org/

[3] https://www.ramitechs.com/

Book Title	Author	Date	Para.	Frag.	Words		
					Original	L4	L3
حي بن يقظان *Hayy Ibn Yaqzan*	ابن طفيل *Ibn Tufail*	1150	213	770	9,962	9,968	10,018
غابة الحق *The Forest of Truth*	فرانسيس مراش *Francis Marrash*	1865	241	1,019	10,081	10,111	10,159
لادياس *Ladiyas*	أحمد شوقي *Ahmed Shawqi*	1899	300	1,481	9,846	9,914	9,951
المحالفة الثلاثية في المملكة الحيوانية *The Tripartite Alliance of the Animal Kingdom*	أمين ريحاني *Ameen Rihani*	1904	315	1,716	10,577	10,595	10,664
الملك كورش *Cyrus the Great*	زينب فواز *Zaynab Fawwaz*	1905	235	1,540	9,910	9,939	9,958
الأجنحة المتكسرة *Broken Wings*	جبران خليل جبران *Kahlil Gibran*	1912	170	1,245	11,482	11,518	11,552
زينب *Zaynab*	محمد حسين هيكل Mohammed Hussein Heikal	1913	138	1,094	9,861	9,855	9,848
شجرة الدر *The Pearl Tree*	جرجي زيدان *Jurji Zaydan*	1914	296	1,593	12,230	12,263	12,268
إبراهيم الكاتب *Ibrahim Al-Katib*	إبراهيم عبد القادر المازني *Ibrahim Abd Al-Qadir Al-Mazini*	1931	295	1,230	10,173	10,198	10,232
ثورة في جهنم *A Revolution in Hell*	نقولا حداد *Niqula Haddad*	1938	492	1,702	11,713	11,743	11,788
سارة *Sara*	عباس محمود العقاد *Abbas Mahmoud Al-Aqqad*	1938	300	1,057	10,079	10,111	10,135
فارس بني حمدان *The Knight of Beni Hamdan*	علي الجارم *Ali Al-Jarem*	1945	329	1,607	10,820	10,838	10,865
على باب زويلة *On Bab Ziwaila*	محمد سعيد العريان *Mohammed Saeed Al-Aryan*	1951	313	1,576	11,701	11,743	11,786
نماذج بشرية *Human Examples*	أحمد رضا حوحو *Ahmad Rida Huhu*	1955	443	1,406	10,080	10,111	10,133
هذا التاج *This Crown*	واصف البارودي *Wasef Al-Baroudi*	1955	209	1,567	10,750	10,770	10,792
			4,289	**20,603**	**159,265**	**159,677**	**160,149**

Table 3: The 15 books we selected to create the SAMER Arabic Text Simplification Corpus. L4-Words and L3-Words refer to the number of words in the Level 4 and Level 3 simplified versions of the books, respectively. Para. is *Paragraphs*; and Frag. is *Punctuated Fragments*.

needed to be simplified. Every Google Doc was equipped with the add-on developed by Hazim et al. (2022) which served as the primary annotation interface. Before sharing the documents with the annotators, all the documents were automatically labeled with their word-level readability using a Python API version of the add-on developed by Hazim et al. (2022). The Python API performs the same functionality as the add-on, but it relies on CAMeL Tools (Obeid et al., 2020) to tokenize and disambiguate the words in context for each chapter. Specifically, the API leverages the BERT unfactored morphological disambiguator developed by Inoue et al. (2022) to retrieve the most probable lemma and part-of-speech (POS) tag for each word. After that, the lemmas and POS tags are looked up in the lemma-based readability lexicon developed by Al Khalil et al. (2020) to identify the readability levels of the words. The add-on employs two additional levels to classify proper nouns (Level 0) and unknown words (Level 6) that are not present in the lexicon. After the word-level readability labeling is done, each chapter is loaded into a Google Doc where the add-on highlights words with different colors according to their readability levels. Figure 1(a) presents a visualization of using the add-on to analyze the readability of a short segment of an Arabic novel. The interface provides a summary of the text's readability distribution levels in a bar chart colored consistently with the readability level word highlights. The interface also provides the option for explicit word-level readability markup by adding a prefix (*#<level>#*) in front of each word, where *<level>* is an Indo-Arabic digit indicating the word readability level (see Figure 1). Moreover, the add-on incorporates the Arabic WordNet (Black et al., 2006) and supports word substitution by displaying suggestions for related words and phrases, e.g., synonyms, antonyms, hypernyms, and hyponyms, with different readability levels. Fig-

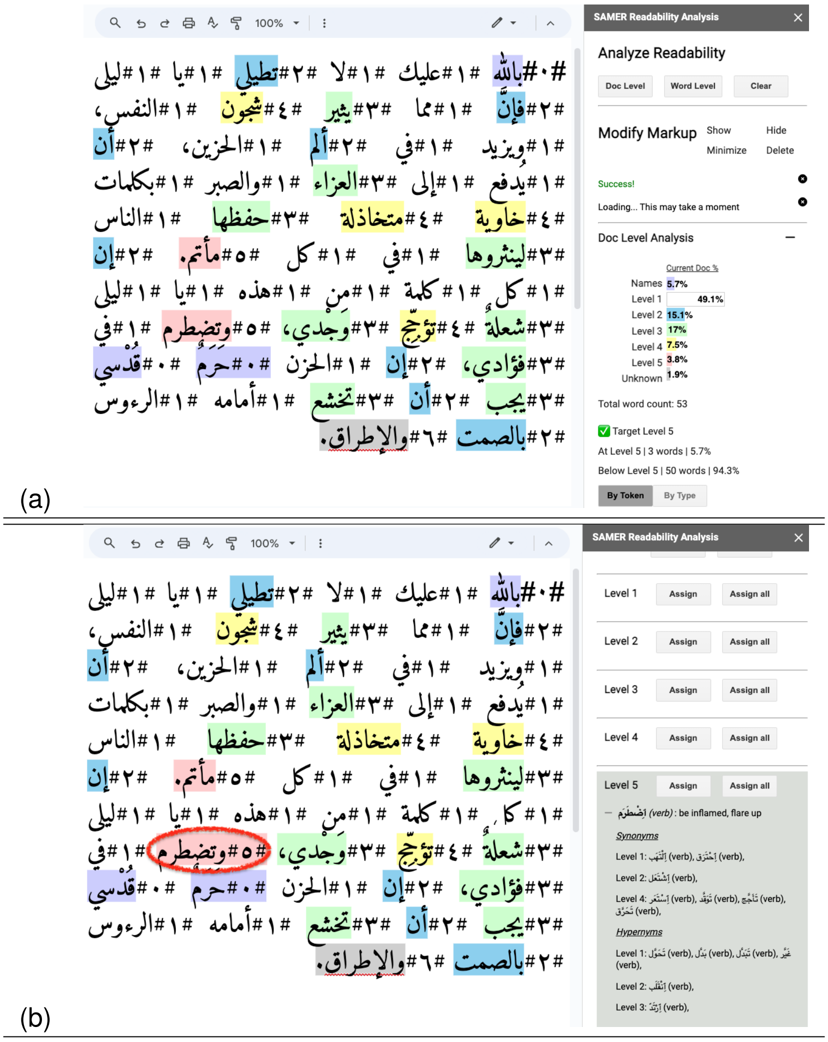

Figure 1: The Google Doc add-on annotation interface introduced by Hazim et al. (2022). Figure (a) is a visualization of word-level and document-level readability. Figure (b) is an example of selecting a specific word to identify all of its analyses and their readability levels. The add-on has multiple markup viewing modes. We show the maximally explicit view where each word is prefixed by an Indo-Arabic digit indicating its level.

ure 1(b) shows the result of selecting a specific word (وتضطرم *wtĎTrm* 'be inflamed'). A sidebar appears showing the different lemma analyses by their readability levels. If the annotators decide to change the automatically assigned readability level, they can either change it directly manually, or by clicking on the `Assign` button to change that specific word's readability level markup or the `Assign All` button to change all of its occurrences in the document.

4.4. Annotation Guidelines

The annotation process starts with determining the original text readability level. To do so, the annotators are instructed to check the word-level readability levels assigned to each word using the add-on. The annotators would then determine the readability level of the document based on the highest readability level found among the words in the document. Since the readability levels are automatically generated, there might be cases where the predicted levels do not reflect the true readability of the words. This could happen either because the lemmas of some words are not in the lexicon or due to morphological tagging errors when the lemmas and POS tags are identified. Therefore, the annotators are asked to use the add-on to fix all of the anomalies related to the readability levels of the words. In cases where correct readability levels are not among the add-on suggestions, the annotators are asked to adjust the levels manually.

After making the corrections, the annotators need to re-run the add-on analysis to determine the document's readability level. If the document has a readability of **Level 5**, then the annotators have

	Original		L4		L3	
L0	2,631	1.7%	5,212	3.3%	5,246	3.3%
L1	83,772	52.6%	90,232	56.5%	95,898	59.9%
L2	23,103	14.5%	26,297	16.5%	30,015	18.7%
L3	22,517	14.1%	24,630	15.4%	28,990	18.1%
L4	14,965	9.4%	13,306	8.3%	0	0.0%
L5	9,463	5.9%	0	0.0%	0	0.0%
L6	2,814	1.8%	0	0.0%	0	0.0%
Total	159,265	100%	159,677	100%	160,149	100%

Table 4: Readability levels statistics of the words in our corpus over the original text (Original), Level 4 simplified text (L4), and Level 3 simplified text (L3).

to first simplify it to **Level 4**, and then to **Level 3** starting from **Level 4**. If the original document has a readability **Level 4** then the annotators need to simplify it to **Level 3**. However, if the document has a readability **Level 3** then no simplification is needed. When simplifying text, the annotators are allowed to perform minimal replacements, deletions, and insertions that are needed to reduce the readability level of the document. Each of these operations might involve more than one word at a time. However, the simplification should be done carefully so that the original meaning of the text is preserved, and as such no abridgement or summarization should take place. Once the simplification is done, the annotators need to re-run the add-on to analyze the text and verify that the document readability level has been adjusted as intended. At the end of the annotation process, each document will have **three** parallel versions: (1) The **Original** text; (2) **Level 4** simplified text; (3) and **Level 3** simplified text. Each of the parallel versions will also have document- and word-level readability annotations. Moreover, all three versions of each document will have the same number of paragraphs and sentences by design.

5. Corpus Overview and Statistics

5.1. Inter-Annotator Agreement

To validate the quality of the annotations, we selected ∼1300 words from each book (17 paragraphs or ∼20K words in total) to be double annotated. Quantitatively, the differences between the annotators' texts are on par with the differences from the original text. As such our inter-annotator agreement check is mostly qualitative. We iterate over the doubled-annotated files word by word to investigate all differences. The double-annotated data had ∼6.8% word-level mismatches when simplifying the original text to Level 4. The number of mismatches increased to 13.2% for Level 4 to Level 3 simplification. Most mismatches came from the fact that the annotators made different lexical simplification choices throughout the annotation process. Annotation mistakes were very infrequent and constituted ∼10% of all mismatches.

5.2. Corpus Readability Statistics

Table 4 presents the readability levels of the words in our corpus over the Original, Level 4, and Level 3 texts. After the annotation, the 15 original texts (159,265 words) resulted in 159,677 Level 4 words (0.3% increase from the original) and 160,149 Level 3 (0.3% increase from Level 4). Although the three versions of the texts are almost identical in size in terms of the number of words, the distribution of the readability levels of the words varies significantly. Comparing the readability levels of the words in the original text against the ones in Level 4 (L4), there is an overall shift from higher to lower levels in 8.8% of all original words. Specifically, all of the unknown words in the original text that are not present in the lemma-based readability lexicon (L6) were manually assigned a readability level and further simplified in case they were of Level 5 (L5). Furthermore, **all** of the Level 5 (L5) words and **some** of the Level 4 (L4) words in the original text were simplified to lower levels. This highlights that the simplification process of the original text to Level 4 involved a mix of manual corrections to the automatically assigned readability labels and lexical simplification of words that have a readability level higher than Level 4 (L4). When it comes to simplifying Level 4 (L4) text to Level 3 (L3), there is a comparable 8.3% change coming from simplifying Level 4 (L4) words to lower levels. Lastly, when comparing the readability levels of the words in the original text to the ones in Level 3 (L3), there is a 17.1% decrease in the overall readability levels.

	Original-L4		L4-L3	
No Change	152,214	95.5%	145,090	90.8%
1-1	6,430	4.0%	13,508	8.5%
1-m	337	0.2%	476	0.3%
m-1	120	0.1%	235	0.1%
Insert	209	0.1%	354	0.2%
Delete	40	0.0%	122	0.1%

Table 5: Lexical simplification transformation statistics in terms of replacements, insertions (Insert), and deletions (Delete) when simplifying the original text to Level 4 (Original-L4) and Level 4 text to Level 3 (L4-L3). No Change indicates no transformations. 1-1 (one-to-one), 1-m (one-to-many), and m-1 (many-to-one) indicate the different types of replacements. The percentages are calculated against the source, i.e., Original for Original-L4, and L4 for L4-L3.

5.3. Transformations Statistics

To obtain the different types of lexical transformations that were applied throughout the annotation process, we use and extend the character edit distance (CED) word alignment tool developed by Khalifa et al. (2021). We obtain the word-level alignments between the original text and its Level 4 simplified version (Original-L4) and between the Level 4 text and its Level 3 simplified version (L4-L3). Each of these two alignments produces a sequence of word-level edit operations representing the lexical transformations in terms of insertions, deletions, and replacements that were a result of the manual simplification process. Each of these operations could involve more than one word at a time. Table 5 presents the statistics of the different lexical transformations in our corpus. The majority of the words did not involve any transformations (No Change) when simplifying the original text to Level 4 and the Level 4 text to Level 3. We quantify three types of replacements: one-to-one (1-1), one-to-many (1-m), and many-to-one (m-1) based on the number of words involved in each transformation. Out of the three replacement types, 1-1 replacements are the majority affecting 4% of the words in the original text when it is simplified to Level 4 and 8.5% of the Level 4 text when it is simplified to Level 3. We note that many-to-many (m-m) replacements do not occur in the corpus. When it comes to insertions and deletions, they represent small percentages of all transformations.

We note that the difference in the number of changes between lexical transformations (Table 5) and readability level shifts (Table 4) is due to corrections in leveling without any word change, e.g., incorrectly labelled proper nouns are mapped to Level 0 without changing them.

5.4. Fragments Statistics

After the annotation, we segmented the original paragraphs and their annotated parallels into smaller fragments based on punctuation. Since punctuation did not change throughout the annotation process, all of the parallel texts in our corpus will have the same number of punctuated fragments, and all of the fragments will be perfectly aligned. The segmentation was done carefully to guarantee that the fragments included a combination of both words and punctuation, rather than consisting solely of punctuation. We refer to the output of this segmentation process as fragments rather than sentences. This is because Arabic sentence segmentation is challenging due to the dearth of punctuation marks and the dual use of the Arabic comma (،) for phrase and clause boundaries (Habash et al., 2022). As such these fragments may be full sentences, subordinated clauses or phrases. Table 6 presents the statistics of the fragments in our corpus, with associated examples. In total, there are 20,603 fragments across all texts (Original, Level 4, and Level 3). On average, each fragment consisted of ∼7.5 words. Overall, 43.3% of all fragments did not have any changes, whereas 12.7% of the fragments included changes only when the original text was simplified to Level 4 (Change in L4 only). 30.9% of fragments included changes only when the Level 4 text was simplified to Level 3 (Change in L3 only). Lastly, 13.1% of all fragments included changes in both simplified Level 4 and Level 3 texts.

5.5. Corpus Splits

To aid reproducibility when using our corpus for various research experiments, we provide train (Train), development (Dev), and test (Test) splits. We split each novel based on the number of words into

	n	n%		Example
No Change	8,920	43.3%	**Original L4 & L3**	ولكنها تعلمت في المدارس الفرنسية أيضا، But she studied in French school as well.
Change in L4 only	2,610	12.7%	**Original**	ونسمعها **تندب وتنوح كالثكلى**. And we hear her **lamenting and wailing like a bereaved woman**.
			L4 & L3	ونسمعها **تبكي وتصرخ كفاقدة ابنها**. And we hear her **crying and screaming like one who lost her son**.
Change in L3 only	6,369	30.9%	**Original L4**	يجب أن **يترأس** الجلسة، He must **preside over** the session,
			L3	يجب أن **يقود** الجلسة، He must **lead** the session,
Change in L4 & L3	2,704	13.1%	**Original**	أحدهما مسرج **ملجم**، One of them is saddled and **bridled**$_1$,
			L4	أحدهما **مسرج مربوط**، One of them is **saddled**$_2$ and **tied**$_1$,
			L3	أحدهما **معد** مربوط، One of them is **readied**$_2$ and tied,
Total	20,603	100.0%		

Table 6: Statistics of fragments in our corpus based on the changes that are made to the text when it is simplified to Level 4 (L4) and then to Level 3 (L3). **n** is the number of fragments. The words in **red** are all of readability level 5 and the words in **blue** are all of readability level 4. L4 is the simplified version of the original sentence where all level 5 words (in **red**) are simplified to level 4 or lower. L3 is the simplified version of the L4 sentence where all level 4 words (in **blue**) are simplified to level 3 or lower.

Train (70%), Dev (15%), and Test (15%), while respecting the full chapter boundaries. We follow the recommendations of Diab et al. (2013) and select full chapters from each novel such that the chapters that are in the Dev and Test sets are taken from well-separated regions of the novel. This ensures that results derived from the Dev and Test sets are not due to mere proximity of subject matter. Altogether, we end up with 113,476 words (71%) for Train, 22,280 words (14%) for Dev, and 23,509 words (15%) for Test.

6. Conclusions and Future Work

We presented the first manually annotated Arabic parallel corpus for text simplification targeting school-aged learners. Our corpus comprises texts of 159K words selected from 15 publicly available Arabic fiction novels. Our corpus includes readability level annotations at both the document and word levels, as well as two simplified parallel versions for each text targeting learners at two different readability levels. We described the corpus selection process and outlined the guidelines we followed to create the annotations and ensure their quality. The corpus, its parallel versions and splits, as well as, the annotation guidelines are publicly available on the SAMER Project website: http://samer.camel-lab.com/.

In future work, we plan to extend our corpus to include text from other genres and domains. We also plan to use it in developing models for readability assessment and automatic simplification for Arabic. By building our corpus and making it publicly available, we hope to encourage research on Arabic text simplification and automatic readability assessment, as well as development of personalized Arabic pedagogical applications.

Ethics Statement

All of the text we selected are in the public domain. All of the annotators were paid fair wages for the tasks of simplification. This effort did not require the use of extensive and compute-heavy learning models. We acknowledge that our resource, like many others in NLP, can be used to guide controlled text generation for unethical purposes it was not intended for such as plagiaristic rewriting.

Acknowledgements

This project was funded by a New York University Abu Dhabi Research Enhancement Fund grant.

7. Bibliographical References

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for Standard Arabic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3053–3062, Marseille, France. European Language Resources Association.

Muhamed Al Khalil, Nizar Habash, and Hind Saddiki. 2017. Simplification of Arabic masterpieces for extensive reading: A project overview. *Procedia Computer Science*, 117:192–198.

Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A Leveled Reading Corpus of Modern Standard Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the Arabic wordnet project. In *Proceedings of the Global WordNet Conference (GWC)*, pages 295–300.

Arnaldo Candido, Erick Maziero, Lucia Specia, Caroline Gasperin, Thiago Pardo, and Sandra Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado. Association for Computational Linguistics.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Will Coster and David Kauchak. 2011a. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon. Association for Computational Linguistics.

William Coster and David Kauchak. 2011b. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Dirk De Hertog and Anaïs Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 328–334, New Orleans, Louisiana. Association for Computational Linguistics.

Mona Diab, Nizar Habash, Owen Rambow, and Ryan Roth. 2013. LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

Richard Evans, Constantin Orăsan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 131–140, Gothenburg, Sweden. Association for Computational Linguistics.

Caroline Varaschin Gasperin, Erick Galani Maziero, Lucia Specia, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. In *Congresso da Sociedade Brasileira de Computação - CSBC*. SBC.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applica-*

tions, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019. Recursive context-aware lexical simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4853–4863, Hong Kong, China. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli, and Omar Kallas. 2022. Camel treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2672–2681, Marseille, France. European Language Resources Association.

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

Eva Hasler, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. 2017. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45:221–235.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado. Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Zhengyang Jiang, Nizar Habash, and Muhamed Al Khalil. 2020. An online readability leveled Arabic thesaurus. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 59–63, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan. The COLING 2016 Organizing Committee.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Salam Khalifa, Ossama Obeid, and Nizar Habash. 2021. Character Edit Distance Based Word Alignment. `https://github.com/CAMeL-Lab/ced_word_alignment`.

Nouran Khallaf, Serge Sharoff, and Rasha Soliman. 2022. Towards Arabic sentence simplification via classification and generative approaches. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 43–52, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gondy Leroy, James E. Endicott, Obay Mouradi, David Kauchak, and Melissa Just. 2012. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:522–31.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3536–3553, Online. Association for Computational Linguistics.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Francis Marrash. 1865. *Ghabat Al-Haqq (The Forest of Truth)*. Hindawi Foundation.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 319–326, Abu Dhabi, UAE. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016a. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016b. Understanding the lexical simplification needs of non-native speakers of English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 717–727, Osaka, Japan. The COLING 2016 Organizing Committee.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021. Lsbert: Lexical simplification based on bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8649–8656.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Human-Computer Interaction – INTERACT 2013*, pages 203–219, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sa'ud Sanusi. 2012. *Saaq al-bambuu*. Arab Scientific Publishers, Inc., Beirut.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. SimPA: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kim Cheng Sheang, Daniel Ferrés, and Horacio Saggion. 2022. Controllable lexical simplification for English. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 199–206, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Sara Botelho Silveira and António Branco. 2012. Using a double clustering approach to build extractive multi-document summaries. In *Text, Speech and Dialogue*, pages 298–305, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language*, pages 30–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sanja Štajner and Sergiu Nisioi. 2018. A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.

Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Facilita: Reading assistance for low-literacy readers. In *Proceedings of the 27th ACM International Conference on Design of Communication*, SIGDOC '09, page 29–36, New York, NY, USA. Association for Computing Machinery.

Kristian Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):927–932.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Taha Yasseri, András Kornai, and János Kertész. 2012. A practical approach to language complexity: A wikipedia case study. *PLOS ONE*, 7(11):1–8.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

8. Language Resource References

Sandra Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles, California. Association for Computational Linguistics.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Stefan Bott, Horacio Saggion, and Simon Mille. 2012. Text simplification tools for Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1665–1671, Istanbul, Turkey. European Language Resources Association (ELRA).

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA. Association for Computational Linguistics.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Helena de Medeiros Caseli, Tiago de Freitas Pereira, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Caroline Gasperin, and Sandra Maria Aluísio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79, Kiyv, Ukraine. Association for Computational Linguistics.

L. Familiar and T. Assaf. 2017. *Saud al-Sanousi's Saaq al-bambuu: the authorized abridged edition for students of Arabic*. G - Reference,Information and Interdisciplinary Subjects Series. Georgetown University Press.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France. European Language Resources Association.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? assessing the readability of Basque texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. 2022. JADES: New text simplification dataset in Japanese targeted at non-native speakers. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 179–187, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria. Association for Computational Linguistics.

Sigrid Klerke and Anders Søgaard. 2012. DSim, a Danish parallel corpus for text simplification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey. European Language Resources Association (ELRA).

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of* *the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Namoos Hayat Qasmi, Haris Bin Zia, Awais Athar, and Agha Ali Raza. 2020. SimplifyUR: Unsupervised lexical text simplification for Urdu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3484–3489, Marseille, France. European Language Resources Association.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.*, 6(4).

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimplesenteval-2021 shared task: Evaluating sentence simplification for russian. In *Proceedings of the Computational Linguistics and Intellectual Technologies Conference*, pages 607–617.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.