Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 192–198

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

Simplification of Arabic Masterpieces for Extensive Reading: A Project Overview

Muhamed Al Khalil[a,*], Nizar Habash[a], Hind Saddiki[a,b]

[a]*New York University Abu Dhabi, Abu Dhabi, UAE*
[b]*Mohammadia School of Engineering, Mohammed V University in Rabat, Rabat, Morocco*

Abstract

We describe the motivation and outline of a project for the simplification of Arabic masterpieces for extensive reading, a collaboration between researchers in Arabic literature, pedagogy and natural language processing, with the purpose of formulating simplification guidelines for Arabic fiction targeting school-aged learners; then using them to guide human simplification efforts with support from state-of-the-art computational natural language processing technology.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: fiction; reading; simplification; education; K-12; Modern Standard Arabic;

1. Introduction

Despite its large speaker base and geopolitical importance, resource availability for the Arabic language remains lacking compared to English or other major world languages. This directly impacts its usage and learnability, especially considering the dearth of independent reading materials in Arabic for young learners. With an average illiteracy rate of 27% in Arab countries vs. 16% in the rest of the world[1], we cannot overstate the need for more independent reading materials suitable for different age groups to reinforce success in school [1], especially with the

* Corresponding author. Tel.: +971-2-628-4112
E-mail address: muhamed.alkhalil@nyu.edu
[1] ALESCO statement on Arab Literacy Day (January 8 2017): http://www.alecso.org/site/

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.109

ongoing civil conflicts and their social disruptions in the Middle East leaving more youth with no access to formal education.

In the context of natural language processing (NLP), we see a similar gap for Arabic: research efforts are mostly focused on tools and resources for Modern Standard Arabic (MSA) news or (to a limited extent) dialectal Arabic in social media ([2-4] among others). While there are large publicly available, machine-readable collections for newswire like Gigaword [5], or for dialectal Arabic (DA) such as the Gumar Corpus [6], we would be hard-pressed to find similarly substantial resources or computational tools geared towards other registers of Standard Arabic, namely in fiction or education.

In this paper, we present an overview of the SAMER project (Simplification of Arabic Masterpieces for Extensive Reading), which brings together researchers in Arabic pedagogy, literature, and NLP in order to build a corpus of curricular readings in Arabic; formulate data-driven guidelines and models for simplification, and use them to guide the simplification of a collection of novels. By the end of the project, we expect to have a sizeable corpus in Arabic based on K12 curricula, a fiction corpus of original and simplified novels, in addition to a graded reader scale and a suite of computational tools (simplification framework and reading-level identification system). Finally, methodology and lessons learned from the simplification process itself form a new approach to developing graded readers by leveraging advanced NLP tools, one that could facilitate the creation of more extensive reading materials among publishers of Arabic.

2. Related work

2.1. Arabic language: reading pedagogy

Researchers on language acquisition agree on the importance of reading for attaining academic success [7-9]. Benefits specific to extensive reading outside school are well-documented [1, 10, 11], namely:

- Promoting vocabulary growth and verbal fluency,
- Improving test scores in all subjects across the curriculum,
- Enhancing learning ability and grammatical knowledge,
- Enriching general content knowledge.

Yet, the catalogues of prominent publishers indicate an immense disparity between extensive reading books available for Arabic and for other languages. For instance, Collins Big Cat offer nearly 700 works for English and less than 100 for Arabic [12, 13], while Oxford University Press presents an even smaller selection of under 50 books for Arabic [14]. Other than a few independent Arabic publishers with modest catalogues, the only notable effort operating in the Middle East is a reader leveling system developed by Taha[2] in collaboration with the Arab Thought Foundation, which is being adopted by some publishers for levelling children's books in Arabic. However, while this is a commendable step towards a standardized levelling framework, in practice this classification is often done intuitively regardless of text readability as it lacks enforceable guidelines and tools. We expect resources such as our curricular corpus to compensate for that shortcoming (fill/bridge gap/void) by supporting practitioners' intuition and expertise with empirical and quantifiable data.

Graded (or levelled) readers are books for extensive reading designed to be suitable for a target age group. They are created either from original material, a slow and laborious creative effort, or by simplifying popular fiction to fit the target level. Publishers usually provide rules and guidelines for simplifying fiction into graded readers while still relying on the simplifier's intuition and expertise: the simplifier is given directives to favor vocabulary from an approved list, respect a specific headword or lemma count for a target level, or avoid certain complex grammatical constructs [15]. With the recent advances in NLP, we can offer even greater support for simplifying graded readers through computational methods for discovering target vocabulary or measuring the difficulty of a text or excerpt.

[2] Hanada Taha Leveling System: http://hanadataha.com/leveling-system-3/