Compilation of an Arabic Children's Corpus

Latifa Al-Sulaiti, Noorhan Abbas, Claire Brierley, Eric Atwell, Ayman Alghamdi
University of Leeds
School of Computing, University of Leeds, LS2 9JT, UK
E-mail: lsulaiti9@hotmail.com, noorhanabbas@yahoo.co.uk, C.Brierley@leeds.ac.uk, E.S.Atwell@leeds.ac.uk ,
scaaa@leeds.ac.uk

Abstract

Inspired by the *Oxford Children's Corpus*, we have developed a prototype corpus of Arabic texts written and/or selected for children. Our *Arabic Children's Corpus* of 2950 documents and nearly 2 million words has been collected manually from the web during a 3-month project. It is of high quality, and contains a range of different children's genres based on sources located, including classic tales from *The Arabian Nights*, and popular fictional characters such as *Goha*. We anticipate that the current and subsequent versions of our corpus will lead to interesting studies in text classification, language use, and ideology in children's texts.

Keywords: Arabic, Children's Corpus, Genre Classification

1. Introduction

When the *Oxford Children's Corpus* (OCC) was published in 2012, it was heralded as the first of its kind (Wild, Kilgarriff and Tugwell 2012). The preferred definition of children's literature adopted for that corpus is material written especially for children, and/or selected for children by parents, teachers, and publishers. Version 1.0 of the OCC contains some 30 million tokens of English text for 5-14 year olds, organised into four basic genres (*fiction*; *non-fiction*; *children's own writing*; and *unclassified*), and with an emphasis on 21st-century material (*ibid*).

Children's corpora as defined above appear to be few and far between. A small corpus of children's texts for English (less than 1 million words) was extracted from the *British National Corpus* (BNC) by Thompson and Sealey (2007) for corpus-based insight into the distinguishing features of children's versus adult fiction. There is also a children's literature section in the *Corpus of Translated Finnish* (Mauranen 2000), where the main source languages are English and Russian; this has been used in studies of translationese in Finnish children's texts (Puurtinen 2003; 1998). Finally, children's fiction is included in the parallel *Bulgarian-Polish Corpus* (Dimitrova and Koseska-Toszewa 2009).

Rather like the OCC for English, the *Arabic Children's Corpus* (ACC) introduced in this paper is a first for the Arabic language. Surveys of Arabic children's literature do exist. Perhaps the most comprehensive is Al-Hajji's *Bibliographical Guide* (1990), which covers an estimated 12,000 books published for children across the Arab region between 1950-1999 (Mdallel 2003). Al-Hazza (2006) offers an update on new and selected publications which genuinely reflect contemporary Arab culture; and Peterson (2005) offers an intriguing survey of Arabic children's magazines, most notably: *Majid*; *Alaa Eldin*; *Al-Arabi Alsaghir*; and *Bolbol*.

Compilation (and later annotation) of the ACC is an ongoing project. However, version 1.0 of the corpus, with some 1,877,615 word tokens, and drawn exclusively from the Internet, already attempts to classify a wider range of children's genres than any of the above datasets.

2. Why Collect a Corpus of Children's Literature?

The OCC was primarily compiled for the purposes of lexicography, to inform children's dictionaries at Oxford University Press (OUP). A large corpus was needed to help refine headword lists, and to identify collocates, senses, and naturally-occurring examples of target terms used in context (Wild, Kilgarriff and Tugwell 2012). Corpus comparison of children's versus adult literature (where a representative sample of the latter was drawn from the BNC) in SketchEngine (Kilgarriff *et al* 2014) established that children's dictionaries should not just be simplified versions of adult dictionaries, and should be based on children's corpora since these properly reflect the language and content that children are exposed to (Wild, Kilgarriff and Tugwell 2012).

Critical Linguistics or Critical Discourse Analysis (Wodak and Meyer 2001) is a major field of enquiry with regard to children's literature. Researchers are interested in the interrelationship of syntax, readability and ideology (Puurtinen 1998; Mdallel 2003). Strategies such as nominalisation and passivisation in translations of children's texts are subtle effects which tend to obscure agency, and hence responsibility (Puurtinen 1998). At the same time, they introduce linguistic complexity, and may detract from readability, defined as the ease and naturalness with which such texts may be read aloud by children themselves (*ibid*).

Children's texts in general are noted for their pedagogic and didactic overtones (*ibid*), and this pertains to Arabic children's literature as well (Mdallel 2003). Based on Al-Hajji's bibliography (1999), one of the most popular genres besides overtly religious texts (*e.g.* stories about the Prophet's life and the lives of other prophets such as Moses and Jesus) is historical fiction featuring Islamic heroes (Mdallel 2003). This is corroborated in a study of Arabic children's magazines (Peterson 2005), where comic characters are said to model the 'hybrid identity' of young Arabs as both Muslim and modern, via juxtaposition of Islamic and consumer values.

Building on Puurtinen's work cited above, Mdallel

(2003) offers further evidence of translation as a vehicle for enculturation. She notes that while Russian and Chinese publishers specialise in translating their children's classics into Arabic, there are far fewer Arabic translations of modern Western literature for children. The influx of translated books is also said to be detrimental to the spread of local literature in the region (*ibid*). However, on a positive note, Al-Hazza and Lucking (2012) welcome English translations of high-quality, contemporary Arab children's fiction in American classrooms to encourage cultural pluralism. Titles cited include: *The Day of Ahmed's Secret* plus *Sami and the Time of the Troubles* (Heide and Gilliland 1990; 1992); and also historical fiction such as: *A Peddler's Dream* (Shefelman 1992) and *Saladdin: Noble Prince of Islam* (Stanley 2002).

3. Sourcing the Corpus

The first stage of designing a corpus is to decide on the sources of texts. Since the Internet is rich in Arabic texts and easily accessible, it was the main source for corpus collection. Two native Arabic speakers on the team identified around 70 websites from which to collect texts. Websites that contained mainly audio and video files were excluded. Scanned PDF files were also excluded since there is no sufficiently accurate Arabic Optical Character Recognition (OCR) system for the low resolution files we came across; although most of the OCRs available online claim high accuracy, tests reveal that their output needs editing (Abbas, Atwell and Al-Sulaiti 2016).

The list of websites was filtered using specified context words such as:

English	Arabic
Arabic children’s stories	قصص اطفال عربيه
Arabic children’s magazines	مجلات اطفال عربيه
Arabic children’s plays	مسرحيات اطفال عربيه
Arabic children’s songs	اناشيد اطفال عربيه
Arabic children’s blog	مدونة أطفال عربية
Arabic children’s forums	منتديات اطفال عربية

Figure 1: Search terms used for locating sources

Most of the websites found were online children’s magazines, children’s websites, forums which include small sections for children’s stories, and blogs. These websites are created by individuals who either include their own writing or collect stories written by others. Usually these websites contain a good selection of stories written by well-known authors. The texts identified cover a variety of categories such as folk tales, religious stories, fiction, biographies, plays, poems, instructions, scientific, medical, general knowledge and warfare texts *et cetera*. Most of the texts found were suitable for older children, aged seven and above. However, texts for beginners are very few and some are scanned PDF files or audio files. To make the corpus more balanced, it was decided to manually type these texts for younger children.

Once suitable websites were determined, each researcher embarked on text collection, making sure to keep a record of the title and source of each text in an Excel file. This ensured that the list of texts collected by the two researchers was not the same because some websites copy stories from other sources.

4. Collecting the Corpus

Our initial plan for collecting the corpus was to use bespoke web-as-corpus software, namely: WebBootCat (Baroni *et al* 2006). This offers two alternatives for corpus collection: (i) via seed terms used as queries in Google, where subsequent hits then constitute a first-pass specialist corpus; (ii) and by uploading URLs. In practice, however, we found that sometimes, when the corpus was downloaded from the website, the text was not complete; when compared to the original document either the beginning or the end of the text was missing. This is related to the fact that when the tool tries to remove unwanted data such as navigation menus, links, ads, headers and footers – all referred to as boilerplate – a slice of the text is also removed; the tool cannot separate out redundant from pertinent content, hence some of the latter is lost as well. We also found that, on occasion, the tool failed to process potentially relevant websites which instead came up as errors.

In terms of the second option (*i.e.* uploading URLs), extracting from a website means that all the data on the website will be extracted. This worked well for websites that were solely/mainly for children, and for longer stories, where the URLs only pointed to children’s material; but some websites are forums and contain all sorts of data. Therefore, using URLs was not appropriate for our purposes since everything would be extracted.

Based on the tests conducted in WebBootCat, and the problems outlined above, we found both tool options for automatic corpus collection unsuitable for our purposes. However, the ‘Create Corpus’ tool in SketchEngine (Kilgarriff *et al* 2014) enables researchers to gather texts in a range of different formats and then upload them into SketchEngine format. If it is a .doc(x) file, output text appears without paragraph tags, but if it is from a website, Create Corpus copies paragraph tags from the html.

We therefore decided that the best method of collecting a children’s corpus was to do this manually. In this way, we ensured that texts would be complete and no unwanted material included. Our approach was to examine each site for suitable texts, and to copy and paste material into a file, making note of provenance data such as URL, author, and title of text. Texts were then reformatted and saved as Word files (see Section 6).

5. Classification of Genres

Our corpus presents a new and diverse snapshot of Arabic Children’s Literature in the 21st century. The genre categorisation scheme emerged from corpus collection, and reflects the variety of children’s texts available on the web. Version 1.0 of our corpus classifies texts via two overarching categories: *Fiction* or *Non-Fiction*, and further specifies genres according to the following types (Figure 2).

Fiction	**Non-Fiction**
Adventure Stories Animal Stories Comic Books Contemporary Fiction Detective Stories Educational Fairy Tales Fantasy Fiction Folk Tales Ghost Stories Historical Fiction Humour Moral Tales Nature Stories Plays Poetry & Nursery Rhymes Riddles Science-Fiction	Biography History Informative Texts Religion Science War Other

Figure 2: Genre classification scheme used in the *ACC*

However, classifying texts in terms of their primary genre does not preclude further sub-categorisation. How, for instance, would one categorise the *Harry Potter* series? They are certainly contemporary; but as well as depicting ordinary life as we know it (school, teachers, holidays, home, parents, siblings *etc*), they also conjure a fantasy world of magical powers and beneficent/malicious creatures; and they are not even just children's books, since mums and dads (and people on the train) read them too! Therefore, our genre categorisation scheme is likely to be refined in subsequent versions of the corpus, since the ACC is an ongoing project. One idea is to introduce a more differentiated hierarchy of genres and metadata.

< doc

url = "http://vb.3dlat.net/showthread.php?t=183153"

Title = "قصة الارنب مشمش"

Author = "Unknown"

Genre = "Fiction,Fiction::Moral,Fiction::Animal"

Dialect = "Egyptian" >

Figure 3: Example of story header for one corpus file

6. Presentation of the Corpus: Formatting and Storage

Version 1.0 of our Arabic Children's Corpus has been uploaded into the SketchEngine corpus query tool in preparation for further investigation and analysis. This involved several steps as follows. Stories were collected from different websites and all text formatting, including bold, italics, highlights, *et cetera*, was removed (see Section 4). Then, for each story, a header was added containing the following fields: URL of the story; story title; author; genre; and dialect (see Figure 3).

Sketch Engine creates a configuration file for each uploaded corpus. This configuration file is located in the registry directory with a filename which is the corpus identifier/name on the system. It contains basic information about the corpus such as language and encoding, and also definition of the attributes that correspond to the metatags added to the story files (*cf.* Figure 3). The structure and attribute names in the actual data have to correspond to the corpus configuration file. These attribute definitions are constructed by the corpus owners to enable SketchEngine to handle the metadata properly.

The beginning and the end of each story text was also marked with the paragraph html tags <p> and </p>. Marking the text in this way facilitated using the Onion tool (provided by SketchEngine) to remove duplicate texts when compiling the corpus. All the corpus files were also checked manually to remove any duplicate texts. Figure 4 shows paragraph mark-up for a single document mapped to the header information given in Figure 3.

<p>

كان فيه مرة ارنوب جميل اسمه مشمش كان بيحب ياكل الجزر

وبيحب يلعب في الحديقة

وفي يوم من الايام مامته قالتله يا مشمش يا حبيبي

قالها نعم ياماما

قالتله خد جنيه وروح هات خس وجزرمن السوق عشان اعمل الغداء

قالها حاضر ياماما

وخرج الارنوب مشمش راح السوق

وهو في طريقه للسوق قابل ارنوب صاحبه اسمه سمسم

فسلم عليه

سمسم قال لارنوب انت رايح فين يا مشمش

مشمش قاله انا رايح السوق اجيب خس وجزر لماما عشان تعمل

الغداء

سمسم قاله انا رايح اشتري شيكولاته ايه رايك ما تيجي تشتري

شيكولاته معايا

قاله لا انا مش معايا غير فلوس الخس والجزر

</p>

</doc>

Figure 4: Paragraph mark-up for the same document defined in Figure 3

A simple hierarchy for the genre field has been introduced for this pilot version of the corpus. Multiple values in the genre field can be structured into a hierarchy, so headers in the story files were therefore modified to create this tree-like structure (*cf.* Figure 5).

< doc

url = "http://al-hakawati.net/arabic/stories_Tales/story23.asp"

Title = "الياقوتة العجيبة"

Author = "Unknown"

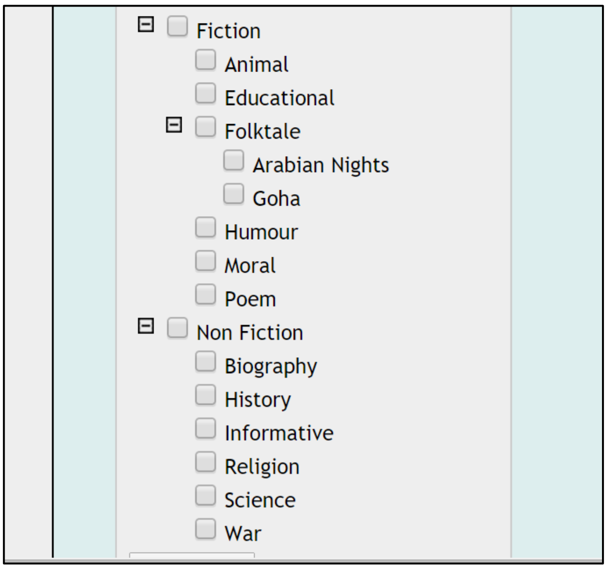
Genre = "Fiction,Fiction::Folktale,Fiction::Folktale::Arabian Nights"

Dialect = "MSA" >

Figure 5: Story header for one corpus file showing 3-level hierarchy for the genre: *Folktales*

Most genres can be defined via 2-levels, but the example in

Figure 5 shows an *Arabian Nights* story classified via 3 levels for the *Folktales* genre: Fiction, Folktale, and Arabian Nights. Figure 6 displays the hierarchy of genres in this current version of the corpus.



- ⊟ □ Fiction
  - □ Animal
  - □ Educational
  - ⊟ □ Folktale
    - □ Arabian Nights
    - □ Goha
  - □ Humour
  - □ Moral
  - □ Poem
- ⊟ □ Non Fiction
  - □ Biography
  - □ History
  - □ Informative
  - □ Religion
  - □ Science
  - □ War

Figure 6: Screenshot from SketchEngine showing simple tree structure for text genres

As previously mentioned, our Arabic Children's Corpus is an ongoing project; but current statistics for version 1.0 are as follows (Figure 7).

Text Type	Number of Tokens
Corpus	1, 877,615
Fiction	711,615
Non Fiction	310, 000
Arabian Nights	806, 000
Goha Stories	50,000
No of stories/documents	2950

Figure 7: Statistics for the *Arabic Children's Corpus* (version 1.0)

7. Further Issues in Text Collection and Classification

We have encountered a number of issues in text collection and classification during this 3-month project. One issue is provenance data: some texts have no authorship attribution; some sites copy from other sites so the actual source of the text cannot be identified; and it is also difficult to ascertain whether some texts are original or translated. There are also issues to do with written quality of texts: some texts contain spelling mistakes (*e.g.* two words are connected, or a word is divided into two, or an incorrect letter is used). We have also found missing prepositions, and redundant words added in a sentence. Errors such as these necessitate further proofreading as another aspect of manual corpus collection.

It is unusual for collectors of corpora to proofread corpus texts, as normally the corpus is used as evidence of real language usage; but corpora aimed at language teaching and learning may be an exception, as we need correct exemplars of the language to be taught (Alfaifi *et al* 2013; Atwell 1987). Another issue is vowelisation: some texts are fully vowelled (*i.e.* contain short vowels and other diacritics), while others are not; this may create problems for analysis later (*e.g.* use of a concordancer). Finally, though we consider genre to be an important feature for inclusion in the corpus, texts are very difficult to categorise. For example, a biography can contain anecdotal stories, or a story with animals as characters may preach a moral lesson.

8. Conclusions and Further Work

We have compiled the first corpus of Arabic children's texts, defined as texts written or selected especially for children. The data, being collected manually, is of a high quality and covers a wide range of genres.

Collecting the corpus presented a major manual challenge. There is no straightforward way of characterising children's texts, and automated methods in WebBootCat (*e.g.* using seed terms to identify and upload URLs) were not always successful in filtering out unsuitable and/or unwanted material. The platform chosen for corpus formatting and storage is SketchEngine, a market leader enabling export of corpus data in a choice of different formats.

We plan to develop and refine this corpus further in the context of a major, applications-based project involving automatic corpus annotation and analysis. Some of our immediate ideas on corpus composition include:

- addition of classic texts by well-known Arabic children's writers;
- revision and further differentiation of genre categories and hierarchy;
- approaching copyright holders if we decide to make this corpus freely available at some point (though it may be impractical to get copyright for each site);
- collecting a subcorpus of children's language, following the design of the *Oxford Children's Corpus*, to enable psycholinguistic and pedagogic research into children's language use.

We plan to use our Arabic Children's Corpus for research on language suitable for use in books for child readers ; for example, we can compare it with other Arabic corpora representing adult language, by extracting vocabulary and formulaic sequences (Alghamdi *et al* 2016) in texts written for children, and comparing these with vocabulary and formulaic sequences in Arabic texts written for adults.

In conclusion, we believe that with further development, our Arabic Children's Corpus will constitute an excellent source of data for a range of Arabic language teaching and linguistics research, and development of reading books for Arab children.

9. References

Abbas, N., Atwell, E. and Al-Sulaiti, L. 2016. 'An Empirical Evaluation of 10 Arabic Optical Character Recognition Tools'. *International Journal of Computational Linguistics (IJCL)*, 7. 1-14.

Alfaifi, A., Atwell E, Abuhakema G. 2013. 'Error Annotation of the Arabic Learner Corpus: A New Error Tagset' in: Language Processing and Knowledge in the Web, LNCS Lecture Notes in Computer Science vol. 8105, pp.14-22. Springer.

Alghamdi, A., Atwell, E. and Brierley, C. 2016. 'An empirical study of Arabic formulaic sequence extraction methods' in *Proceedings of LREC Language Resources and Evaluation Conference.*

Al-Hajji, F.A. 1990. 'al-Dalil al-Biblioghrafi Likitab at-Tifl al-'Arabi. Bibliographical Guide to Arab Children's Books'. Sharjah. *Dairatu al-Thakafa wal-'alam*.

Al-Hazza, T.C. 2006. 'Arab Children's Literature: An Update'. In *Book Links*. 15.3. 11-12.

Al-Hazza, T.C. and Lucking, B. 2012. 'Celebrating Diversity through Explorations of Arab Children's Literature'. In *Childhood Education*. 83.3. 132-135.

Atwell, E. 1987. 'How to detect grammatical errors in a text without parsing it' in: Maegaard, B., (editor) *Proceedings of EACL '87 Third conference on European chapter of the Association for Computational Linguistics*, 38-45.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. 2006. 'WebBootCaT: instant domain-specific corpora to support human translators'. In *Proceedings of EAMT European Association for Machine Translation*. 247-252.

Dimitrova, L. and Koseska-Toszewa, V. 2009. 'Bulgarian-Polish Corpus'. In *Cognitive Studies*. 9. 133-141.

Heide, F.P. and Gilliland, J. 1990. *The day of Ahmed's secret*. New York. Lothrop, Lee and Shepard Books.

Heide, F.P. and Gilliland, J. 1992. *Sami and the time of the troubles.* New York. Clairon Books.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Vojtěch, K., Michelfeit, J., Rychlý, P. and Suchomel, V. 2014. 'The Sketch Engine: ten years on'. In *Lexicography*. 1.1. 7-36.

Mauranen, A. 2000. 'Strange strings in translated language. A study on corpora'. In Olohan, M. (ed.), *Intercultural Faultlines: Research Models in Translation Studies. I. Textual and Cognitive Aspects*. Manchester: St. Jerome. 119-141.

Mdallel, S. 2003. 'Translating Children's Literature in the Arab World'. In *Meta: Translators' Journal*. 48.1-2. 298-306.

Peterson, M.A. 2005. 'The *Jinn* and the Computer: Consumption and identity in Arabic children's magazines'. In *Childhood*. 12.2. 177-200.

Puurtinen, T. 2003. 'Genre-specific Features of Translationese? Linguistic Differences between Translated and Non-translated Finnish Children's Literature'. In *Literary and Linguistic Computing*. 18.4. 389-406.

Puurtinen, T. 1998. 'Syntax, Readability and Ideology in Children's Literature'. In *Meta: Translators' Journal*. 43.4. 524-533.

Thompson, P. and Sealey, A. 2007. 'Through children's eyes?' In *International Journal of Corpus Linguistics*. 12.1. 1-23.

Shefelman, J. 1992. *A peddler's dream*. Austin, Texas. Houghton Mifflin.

Stanley, D. 2002. *Saladin: Noble prince of Islam*. New York. HarperCollins.

Wild, K., Kilgarriff, A. and Tugwell, D. 2012. 'The Oxford Children's Corpus: Using a Children's Corpus in Lexicography'. In *International Journal of Lexicography*. doi: 10.1093/ijl/ecs017. OUP. 1-29.

Wodak, R. and Meyer, M. (eds). 2001. *Methods of Critical Discourse Analysis*. London. SAGE.