arTenTen: a new, vast corpus for Arabic

Yonatan Belinkov
MIT, USA
belinkov@mit.edu

Nizar Habash
Columbia University, USA
nh2142@columbia.edu

Adam Kilgarriff
Lexical Computing Ltd, UK
adam@lexmasterclass.com

Noam Ordan
University of Saarland, Germany
noam.ordan@gmail.com

Ryan Roth
Columbia University, USA
ryanr@ccls.columbia.edu

Vit Suchomel
Masaryk University, Cz., and Lexical Computing Ltd, UK
xsuchom2@fi.muni.cz

Without data, nothing. Corpora are the critical data resources for computational linguistics, particularly in current times where data-driven methods have moved centre-stage. Since 2003, the key resource for Arabic has been Arabic Gigaword, created and distributed by the Linguistic Data Consortium (Graff 2003). It has regularly been updated and is now in its fifth edition. It contains exclusively newswire text.

In this paper, we present arTenTen, a web-crawled corpus of Arabic, gathered in 2012, and a member of the TenTen Corpus Family (Jakubíček et al 2013). arTenTen comprises 5.8 billion words. It was crawled using Spiderling (Suchomel and Pomikalek 2012). It has been further carefully cleaned, including duplicate removal, using the JusText and Onion tools (Pomikalek 2011). We are currently (May 2013) in the process of tokenising, lemmatising and part-of-speech tagging arTenTen with the leading MADA tool version 3.2 (Habash and Rambow 2005; Habash et al. 2009). We will load arTenTen and its annotations into the Sketch Engine -- http://www.sketchengine.co.uk -- a leading corpus query tool, where it will be available for all to investigate.

MADA's in-context annotations are automatically converted into 30 features that are searchable in Sketch Engine. The features include the word in Arabic script and Buckwalter's transliteration (Habash et al., 2007), its full diacritization, lemma (diacritized and undiacritized in Arabic script and in transliteration), stem, part-of-speech (POS), the full Buckwalter Analyzer tag (Buckwalter 2004), values and POS tags for four possible proclitic slots, the values of eight inflection features -- person, aspect, voice, mood, gender, number, state and case, enclitic value and POS tag, English gloss and whether the word had a spelling variation. The following is an example word وبفكرة `and with an idea' together with it 30 features:

```
وبفكرة wbfkrp wabifikorapK fikorap_1 فِكْرَة fkrp فكرة fikor noun wa/CONJ+bi/PREP+fikor/NOUN +ap/NSUFF_FEM_SG+K/CASE_INDEF_GEN null null wa conj bi prep null null null null null null f s i g NULL NULL idea;notion;concept lex
```

Once arTenTen is fully encoded, we will compare it with Arabic Gigaword and an earlier web-crawled corpus available from Leeds University -- http://corpus.leeds.ac.uk/internet.html -- (Sharoff 2006). We also plan to explore arTenTen's composition in relation to Modern Standard Arabic and the dialects, using, amongst other things, Buckwalter and Parkinson's Frequency Dictionary (2011) and the keywords method presented in (Kilgarriff 2012).

References

Buckwalter, T. Buckwalter Arabic Morphological Analyzer v2.0. LDC catalog number LDC2004L02.

Buckwalter, T. and Parkinson, D. 2011. A Frequency Dictionary of Arabic. Rouledge Frequency Dictionary Series.

Graff, D. 2003. Arabic Gigaword. LDC Catalog No.: LDC2003T12. Linguistic Data Consortium.

Habash, N. and Rambow, O. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In Proc. of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan.

Habash, N., Rambow, O. and Roth, R. 2009. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proc. of the International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Habash, N., Soudi, A. and Buckwalter, T. 2007. "On Arabic Transliteration." In A. van den Bosch and A. Soudi, eds., *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. 2013. The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster.

Kilgarriff, A. 2012. Getting to Know your Corpus. Proc. Text, Speech, Dialogue Conference. Brno, Czech Republic: Springer.

Pomikalek, J. 2011. Removing Boilerplate and Duplicate Content from Web Corpora. PhD thesis, Masaryk University, Brno.

Pomikalek, J. and Suchomel, V. 2012. Efficient web crawling for large text corpora. Proc. of the 7th Web as Corpus Workshop (WAC7), Lyon, France.

Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.), WaCky! Working papers on the Web as Corpus. Bologna, Italy.