TECHNICAL REPORT

No. CCLS-12-03

Title: A Modern Standard Arabic Closed-Class Word List

Authors: Wael Salloum and Nizar Habash

A Modern Standard Arabic Closed-Class Word List

Wael Salloum and Nizar Habash
Center for Computational Learning Systems
Columbia University
{wael,habash}@ccls.columbia.edu

This document describes a list of Modern Standard Arabic closed-class words, which can be used as a stop list for a variety of natural language processing applications. The list contains 740 inflected words and clitics in the Arabic Treebank (ATB) tokenization scheme (Maamouri et al., 2004; Habash, 2010). The inflected words are based on 309 lemmas from the Standard Arabic Morphological Analyzer, SAMA (Graff et al., 2009). To get a copy of the full list, please contact the authors.

Approach

We created this closed-class word list by first identifying the closed-class lemmas in SAMA and then generating the inflected forms in the ATB tokenization from these lemmas. We used the Arabic analysis and generation engine ALMOR (Habash, 2007) to execute the morphological generation. This is the same engine used in the Arabic disambiguation and tokenization tool MADA+TOKAN (Habash et al., 2009). ALMOR uses the SAMA databases.
To determine the list of lemmas, we used the following set of resources:

1. ALMOR Part-of-Speech (POS) tag. We included lemmas that have closed-class POS tags, e.g., prep, adv, noun_quant, conj, interj, pron_dem.
2. CATiB list. We used a list of closed-class words that was created as part of the Columbia Arabic Treebank (CATiB) project (Habash and Roth, 2009; Habash et a., 2009). We mapped these words to their lemmas by analyzing them with SAMA and then manually selecting the appropriate lemma for each word. For example, SAMA produces for the word *$bh* three distinct lemmas: *$abah_1* ` resemblance', *$ab~ah_1* `compare, liken', and *$iboh_1* `like, almost'; we manually selected *$iboh_1*. Other words from the CATiB list include: *>badA, Hasob, Ha*ow, HawAlay, mEA, qubAlap, Albat~ap*.
3. Words ending with mA. Examples include *<in~amA, >an~amA, Eam~A, EindamA, HasobamA, HiynamA, baEodamA*.
4. Special verbs such as *kAn* and its sisters are included.

In addition to the lemmas, we include the set of 18 proclitics and enclitics that are tokenized in the ATB tokenization scheme, e.g., *w+* `and', *+hA* 'her'.

The List

The entries in the list consist of five columns:

1. ATB Alif/Yah normalized surface form in UTF-8
2. ATB Alif/Yah normalized surface form in Buckwalter transliteration (Buckwalter, 2004)
3. Unnormalized lemma form in Buckwalter transliteration (Buckwalter, 2004)
4. POS tag
5. English gloss

Examples of different categories are listed below.

Interjection Examples

اه	Ah	>ah_1	interj	ah!;ouch!
اجل	Ajl	>ajal_1	interj	yes;indeed;certainly

Verb Examples

كان	kAn	kAn-u_1	verb	be;was;were
كانت	kAnt	kAn-u_1	verb	be;was;were
كانا	kAnA	kAn-u_1	verb	be;was;were
كانوا	kAnwA	kAn-u_1	verb	be;was;were
اكون	Akwn	kAn-u_1	verb	be;was;were
اكن	Akn	kAn-u_1	verb	be;was;were
نكون	nkwn	kAn-u_1	verb	be;was;were
تكون	tkwn	kAn-u_1	verb	be;was;were
تكن	tkn	kAn-u_1	verb	be;was;were
تكونوا	tkwnA	kAn-u_1	verb	be;was;were
تكونون	tkwnwn	kAn-u_1	verb	be;was;were

Pronominal Examples

انا	AnA	>anA_1	pron	I
نحن	nHn	naHonu_1	pron	we
الذي	Al*y	Al~a*iy_1	pron_rel	which;who;whom_[masc.sg.]
الذين	Al*yn	Al~a*iy_1	pron_rel	which;who;whom_[du.]
الذين	Al*yn	Al~a*iy_1	pron_rel	who;whom_[pl.]
التي	Alty	Al~a*iy_1	pron_rel	which;who;whom_[fem.sg.]
اللواتي	AllwAty	Al~a*iy_1	pron_rel	who;whom_[fem.pl.]

Preposition Examples

الا	AlA	<il~A_2	prep	except
الي	Aly	<ilaY_1	prep	to;towards

عدا	EdA	EadA_1	prep	except_for
علي	Ely	EalaY_1	prep	on;above
عن	En	Ean_1	prep	(away)_from,_off,_about,_on,_over
عما	EmA	Ean_1	prep	about_+_that
حاشا	HA$A	HA$A_1	prep	except
حتي	Hty	Hat~aY_1	prep	until;up_to
في	fy	fiy_1	prep	in
فيم	fym	fiy_1	prep	in_+_what

Nominal Examples

امام	AmAm	>amAm_1	noun	front;forward
عبر	Ebr	Eabor_1	noun	across;over;via;by_means_of;crossing
عند	End	Einod_1	noun	with/at
عندما	EndmA	Einod_1	noun	the_time_when_+_[def.acc.]_+_that
بين	byn	bayona_1	noun	between/among
شبه	$bh	$iboh_1	noun	like;almost;semi-
شطر	$Tr	$aTor_2	noun_quant	Part,_portion,_division,_section
اكثر	Akvr	>akovar_1	noun_quant	more,_most,_majority
اي	Ay	>ay~_1	noun_quant	any
ايها	AyhA	>ay~ap_1	noun_quant	any
ضعف	DEf	DiEof_1	noun_quant	double;multiple
عشر	E$r	Eu$ur_1	noun_quant	(one)_tenth

References

T. Buckwalter. Buckwalter arabic morphological analyzer version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0, 2004.

D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter. Standard Arabic Morphological Analyzer (SAMA) Version 3.1, 2009. Linguistic Data Consortium LDC2009E73.

N. Habash. Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers, 2010.

N. Habash, O. Rambow and R. Roth. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, 2009.

N. Habash. Arabic Morphological Representations for Machine Translation. In A. van den Bosch and A. Soudi, editors, Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer, 2007.

N. Habash, R. Faraj, and R. Roth. Syntactic Annotation in the Columbia Arabic Treebank. In Proceedings of MEDAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2009.

N. Habash and R. Roth. CATiB: The Columbia Arabic Treebank. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 221–224, Suntec, Singapore, 2009.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In NEMLAR Conference on Arabic Language Resources and Tools, pages 102–109, Cairo, Egypt, 2004.