ARL-TN-0735 • Feb 2016

US Army Research Laboratory

ARL Arabic Dependency Treebank

by Stephen C Tratz

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TN-0735 ● FEB 2016

US Army Research Laboratory

ARL Arabic Dependency Treebank

by Stephen C Tratz
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* February 2016	**2. REPORT TYPE** Final	**3. DATES COVERED (From - To)**

4. TITLE AND SUBTITLE ARL Arabic Dependency Treebank	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Stephen C Tratz	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory ATTN: RDRL-CII-T 2800 Powder Mill Road Adelphi, MD 20783-1138	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TN-0735
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This technical note describes the US Army Research Laboratory (ARL) Arabic Dependency Treebank (AADT) for the purpose of documenting its release. The AADT was derived from existing Arabic treebanks distributed by the Linguistic Data Consortium using constituent-to-dependency conversion software written at ARL. Earlier versions of the AADT, as well as parsers trained from it, have been used in several published ARL research efforts, and, by releasing the data, we hope to facilitate additional Arabic language processing research by the greater community.

15. SUBJECT TERMS

Arabic, treebank, parsing

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Stephen C Tratz
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	20	**19b. TELEPHONE NUMBER (Include area code)** 301-394-2305

Contents

Approved for public release; distribution is unlimited.

INTENTIONALLY LEFT BLANK.

1. Overview

This technical note provides a brief description of the US Army Research Laboratory (ARL) Arabic Dependency Treebank (AADT) version 1.0 for the purpose of documenting its release. AADT was automatically derived from 4 existing Linguistic Data Consortium (LDC) resources—the latest versions of the Arabic Treebank (ATB) parts 1, 2, and 3, as well as the Arabic Treebank Broadcast News dataset (Maamouri, et al.; LDC2010T13, LDC2011T09, LDC2010T08, LDC2012T07). The original ATB contains over 2,000 news stories produced by a handful of Arabic news services. Each story was annotated by the LDC, with every token receiving its appropriate part-of-speech tag and morphological segmentation, and every sentence being annotated with its constituent parse. For AADT, we created dependency parses for latest version of the conversion process briefly described in Section 4.5 of the paper "A Cross-Task Flexible Transition Model for Arabic Tokenization, Affix Detection, Affix Labeling, POS Tagging, and Dependency Parsing" (Tratz, 2013). An earlier version of this dependency treebank was also used in the paper "Resumptive Pronoun Detection for Modern Standard Arabic to English MT" (Tratz, 2014). The LDC is one of the foremost sources of annotated data used in computational linguistics, and, by releasing this dependency treebank back to them for redistribution, we hope to facilitate Arabic natural language processing research by the greater community.

The remainder of this technical note defines the dependency tree file format (Section 2), and presents the part-of-speech tag (Section 3) and dependency label (Section 4) schemes used throughout the AADT.

2. File Format

The files are in an 11-column tab-separated format with one or more blank lines between sentences. All files are UTF-8 encoded. An example is presented below.

1	2	3	4	5	6	7	8	9	10	11
0/0	core	-1/0	ROOT	mDy	>amoDaY_1	امضى	AmDY	VB_PV	spend/accomplish/...	-
1/0	core	0/0	subj	rkb	rAkib_1	ركاب	rkAb	NOUN	riders/passengers	-
2/0	pref	-	-	-	-	ال	Al	DET	the	-
2/1	core	1/0	idafa	Tyr	TA}irap_1	طائر	TA}r	NOUN	aircraft/airplane	-
2/2	suff	-	-	-	-	ة	p	NSUFF_FEM_SG	[fem.sg.]	-
3/0	pref	-	-	-	-	ال	Al	DET	the	-
3/1	core	2/1	amod	sEd	saEuwdiy~_1	سعودي	sEwdy	ADJ	Saudi	-
4/0	core	0/0	obj	lyl	layolap_1	ليل	lyl	NOUN	night/evening/soire	-
4/1	suff	-	-	-	-	ت	t	NSUFF_FEM_SG	[fem.sg.]	-
4/2	core	4/0	idafa	-	-	هم	hm	PRON_AG_3_MASC_PL	their	-
5/0	pref	-	-	-	-	ال	Al	DET	the	-
5/1	core	4/0	amod	mDy	mADiy_1	ماضي	mADy	ADJ	past/bygone	-
5/2	suff	-	-	-	-	ة	p	NSUFF_FEM_SG	[fem.sg.]	-
6/0	core	0/0	prep	fy	fiy_1	في	fy	PREP	in	-
7/0	core	6/0	pcomp	fndq	funoduq_1	فندق	fndq	NOUN	hotel	-
…	…	…	…	…	…	…	…	…	…	…

The values of the 11 columns are as follows:

1) Unique identifier for a particular word/affix. The first number indicates the whitespace/punctuation-separated token it belongs to; the second number indicates the morpheme within the token.

2) One of 3 values (core—the “core” part of a word; “pref”—prefix; “suff”—suffix). The term core was chosen to avoid linguistically loaded terms such as *stem* or *root*. It is worth noting that clitics are split off from the remainder of the word and are marked as cores to indicate their word-level status within the conversion. Since clitics appear frequently in Arabic, it is not unusual for a single token to have multiple “core” lines associated with it. Although the definite determiner *Al* is considered a clitic, it is labeled as a “pref” for the sake of convenience—there is never any question as to where it attaches in a dependency tree.

3) Identifier of the governing word.

4) Label of the dependency on the edge.

5) The root of the token. Although the ATB’s integrated format specifies the SAMA (Maamouri et al., LDC2010L01) lemma identifiers for the words, it does not provide the root that the lemma is derived from (Most roots in Arabic are sequences of 3 or 4 characters). Therefore, the values in this field were populated automatically using a utility program that accesses the SAMA database.

6) The lemma identifier in the SAMA database for the given word.

7) Original text.

8) Transliterated text (transliterated using the popular Buckwalter transliteration scheme).

9) Part-of-speech label.

10) Gloss (definition).

11) Sparsely populated field used to indicate co-indexing for resumptive pronouns/affixes.

Thus, taken together, fields 1, 2, and 3 define the labeled dependency edges between all the “core” elements of the sentence. The “pref” and “suff” morphemes are implicitly linked with their adjacent cores, and many of their fields are left empty (indicated by a hyphen).

3. Part-of-Speech (POS) Tag Scheme

The POS tag scheme is similar to the scheme used by the ATB but has a variety of modifications. One important note is that any portion of the original POS label that corresponds to an unwritten portion of a word is simply dropped. For example, if the original label for the token was DET+NOUN+CASE_DEF_GEN but the final *kaSra* short vowel diacritic—the typical indicator of genitive case—was not written, the DET label would appear on one line with the *Al* definite determiner, the NOUN label would appear on a line with the core noun text, and the CASE_DEF_GEN portion would simply be dropped because it does not correspond to a written morpheme. Most of the mappings are 1-to-1 and should be fairly clear to anyone who is already familiar with the ATB POS tag scheme; for example, PVSUFF_SUBJ:2FS is rewritten as PS_2_FEM_SG. A list of all the POS tags is provided in Appendix A.

4. Dependency Label Scheme

For ease of understanding, many of the dependency labels have names that are similar or identical to the most similar dependency labels in the popular Stanford English dependency label scheme (de Marneffe & Manning, 2008). However, this is to not to say that they may always be interpreted identically. Also, some labels are specific to Arabic, including *idafa*, *fidafa*, and *kccmp*. A complete listing of the AATB dependency labels is given in Appendix B.

5. References

Badawi El-Said, Carter Michael G, Gully Adrian. Modern written arabic: A comprehensive grammar. Routledge, 2004.

de Marneffe, Marie-Catherine, Manning Christopher D. The stanford typed dependencies representation. COLING 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation. Association for Computational Linguistics, 2008.

Maamouri Mohamed, Bies Ann, Kulick Seth, Gaddeche Fatma, Mekki Wigdan, Krouna Sondos, Bouziri Basma, Zaghouani Wajdi. Arabic treebank: Part 1 v 4.1 LDC2010T13. Web Download. Philadelphia: Linguistic Data Consortium, 2010.

Maamouri Mohamed, Bies Ann, Kulick Seth, Gaddeche Fatma, Mekki Wigdan, Krouna Sondos, Bouziri Basma, Zaghouani Wajdi. Arabic treebank: Part 2 v 3.1 LDC2011T09. Web Download. Philadelphia: Linguistic Data Consortium, 2011.

Maamouri Mohamed, Bies Ann, Kulick Seth, Krouna Sondos, Gaddeche Fatma, Zaghouani Wajdi. Arabic treebank: Part 3 v 3.2 LDC2010T08. Web Download. Philadelphia: Linguistic Data Consortium, 2010.

Maamouri Mohamed, Bies Ann, Kulick Seth, Krouna Sondos, Tabassi Dalila, Ciul Michael. Arabic treebank - Broadcast News v1.0 LDC2012T07. Web Download. Philadelphia: Linguistic Data Consortium, 2012.

Maamouri Mohamed, Graff David, Bouziri Basma, Krouna Sondos, Bies Ann, Kulick Seth. LDC Standard Arabic Morphological Analyzer (SAMA) version 3.1 LDC2010L01. Web Download. Philadelphia: Linguistic Data Consortium, 2010.

Ryding Karin C. A reference grammar of modern standard arabic. Cambridge University Press, 2005.

Tratz Stephen. A cross-task flexible transition model for arabic tokenization, affix detection, affix labeling, POS tagging, and dependency parsing. Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages. 2013.

Tratz Stephen, Voss Clare, Laoudi Jamal. Resumptive pronoun detection for modern standard arabic to english MT. Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL. 2014.

Appendix A. Part-of-Speech Tag List

The following is a complete list of the part-of-speech tags that appear within the AADT. For convenience, the list is split into 3 sublists based upon whether the tag is used with elements labeled as “core,” “pref,” or “suff,” respectively.

A-1 Part-of-Speech Labels for ‘core’

ABBREV	Abbreviation
ADJ	Adjective
ADJ_COMP	Comparative or superlative adjective
ADJ_NUM	Ordinal number
ADV	Adverb
CONJ	Coordinating conjunction
CONNEC_PART	Connective particle
#DEM_PRON – Demonstrative pronoun	
DEM_PRON	(Non-specific gender and number)
DEM_PRON_FEM	Feminine
DEM_PRON_FEM_DU	Feminine dual
DEM_PRON_FEM_SG	Feminine singular
DEM_PRON_MASC_DU	Masculine dual
DEM_PRON_MASC_SG	Masculine singular
DEM_PRON_PL	Plural
DIALECT	Dialect
EMPHATIC_PART	Emphatic particle
EXCLAM_PRON	Exclamatory pronoun
FOCUS_PART	Focus particle
FOREIGN	Foreign transliterated word
FUT_PART	Future particle
INNA	Inna or one of her sisters
INTERJ	Interjection
INTERROG_PART	Interrogative particle
JUS_PART	Jussive particle
LATIN	Latin script token
NEG_PART	Negative particle
NOUN	Noun
NOUN_NUM	Cardinal number
NOUN_PROP	Proper noun
NOUN_QUANT	Quantifier noun
NUMERIC_COMMA	Numeric comma (letter *reh* used as comma)
PART	Particle
PARTIAL	Partial
PREP	Preposition (true prepositions only)
#PRON – Pronoun	
PRON_AG_1_PL	Accusative/genitive 1st person plural
PRON_AG_1_SG	Accusative/genitive 1st person singular
PRON_AG_2_DU	Accusative/genitive 2nd person dual
PRON_AG_2_FEM_PL	Accusative/genitive 2nd person feminine plural

PRON_AG_2_FEM_SG	Accusative/genitive 2nd person feminine singular
PRON_AG_2_MASC_PL	Accusative/genitive 2nd person masculine plural
PRON_AG_2_MASC_SG	Accusative/genitive 2nd person masculine sing.
PRON_AG_3_DU	Accusative/genitive 3rd person dual
PRON_AG_3_FEM_PL	Accusative/genitive 3rd person feminine plural
PRON_AG_3_FEM_SG	Accusative/genitive 3rd person feminine singular
PRON_AG_3_MASC_PL	Accusative/genitive 3rd person masculine plural
PRON_AG_3_MASC_SG	Accusative/genitive 3rd person masculine singular
PRON_NOM_1_PL	Nominative 1st person plural
PRON_NOM_1_SG	Nominative 1st person singular
PRON_NOM_2_FEM_SG	Nominative 2nd person feminine singular
PRON_NOM_2_MASC_PL	Nominative 2nd person masculine plural
PRON_NOM_2_MASC_SG	Nominative 2nd person masculine singular
PRON_NOM_3_DU	Nominative 3rd person dual
PRON_NOM_3_FEM_PL	Nominative 3rd person feminine dual
PRON_NOM_3_FEM_SG	Nominative 3rd person feminine singular
PRON_NOM_3_MASC_PL	Nominative 3rd person masculine plural
PRON_NOM_3_MASC_SG	Nominative 3rd person masculine singular
PUNC	Punctuation
RC_PART	Result clause particle (introduces apodosis)
RESTRIC_PART	Restrictive particle
RI_ADV	Relative or interrogative adverbial
#RI_PRON – Relative or interrogative pronoun	
RI_PRON_DEF	Definite
RI_PRON_FEM_DU_AG_DEF	Feminine dual accusative/genitive definite
RI_PRON_FEM_DU_NOM_DEF	Feminine dual nominative definite
RI_PRON_FEM_PL_DEF	Feminine plural definite
RI_PRON_FEM_SG_DEF	Feminine singular definite
RI_PRON_INDEF	Indefinite
RI_PRON_MASC_DU_AG_DEF	Masculine dual accusative/genitive definite
RI_PRON_MASC_DU_NOM_DEF	Masculine dual nominative definite
RI_PRON_MASC_PL_DEF	Masculine plural definite
RI_PRON_MASC_SG_DEF	Masculine singular definite
SUB_CONJ	Subordinating conjunction
TRANSERR	Transcription error
TYPO	Typo
VB_CV	Command verb
VB_IV	Imperfect verb

VB_IV_PASS	Imperfect verb passive voice
VB_PV	Perfect verb
VB_PV_PASS	Perfect verb passive voice
VERB_PART	Verb particle (*laqado* and *qado*)
VOC_PART	Vocative particle

A-2 Part-of-Speech Labels for ‘pref’

DET	Determiner (NOTE: On occasion, the Al determiner will appear unattached; in this relatively rare situation, the determiner is considered a ‘core’.)
#IP – Imperfect verb prefix	
IP_1_PL	1st person plural
IP_1_SG	1st person singular
IP_2_DU	2nd person dual
IP_2_FEM_PL	2nd person feminine plural
IP_2_FEM_SG	2nd person feminine singular
IP_2_MASC_PL	2nd person masculine plural
IP_2_MASC_SG	2nd person masculine singular
IP_3_FEM_DU	3rd person feminine dual
IP_3_FEM_PL	3rd person feminine plural
IP_3_FEM_SG	3rd person feminine singular
IP_3_MASC_DU	3rd person masculine dual
IP_3_MASC_PL	3rd person masculine plural
IP_3_MASC_SG	3rd person masculine singular

A-3 Part-of-Speech Labels for ‘suff’

#CASE – Case marker	
CASE_DEF_ACC	Definite accusative
CASE_DEF_GEN	Definite genitive
CASE_DEF_NOM	Definite nominative
CASE_INDEF_ACC	Indefinite accusative
CASE_INDEF_GEN	Indefinite genitive
CASE_INDEF_NOM	Indefinite nominative
#CS – Command verb suffix	
CS_2_DU	2nd person dual
CS_2_FEM_SG	2nd person feminine singular
CS_2_MASC_PL	2nd person masculine plural
CS_2_MASC_SG	2nd person masculine singular
#IS – Imperfect verb suffix	
IS_2_FEM_SG_i	2nd person feminine singular indicative
IS_2_FEM_SG_s_j	2nd person feminine singular subjunctive/jussive
IS_DU_i	Dual indicative
IS_DU_s_j	Dual subjunctive/jussive
IS_FEM_PL	Feminine plural
IS_MASC_PL_i	Masculine plural indicative
IS_MASC_PL_s_j	Masculine plural subjunctive/jussive
IS_i	Indicative
IS_j	Jussive
IS_s	Subjunctive

#NS – Nominal suffix	
NS_FEM_DU_AG	Feminine dual accusative/genitive
NS_FEM_DU_AG_POSS	Feminine dual accusative/genitive possessed
NS_FEM_DU_NOM	Feminine dual nominative
NS_FEM_DU_NOM_POSS	Feminine dual nominative possessed
NS_FEM_PL	Feminine plural
NS_FEM_SG	Feminine singular
NS_MASC_DU_AG	Masculine dual accusative/genitive
NS_MASC_DU_AG_POSS	Masculine dual accusative/genitive possessed
NS_MASC_DU_NOM	Masculine dual nominative
NS_MASC_DU_NOM_POSS	Masculine dual nominative possessed
NS_MASC_PL_AG	Masculine plural accusative/genitive
NS_MASC_PL_AG_POSS	Masculine plural accusative/genitive possessed
NS_MASC_PL_NOM	Masculine plural nominative
NS_MASC_PL_NOM_POSS	Masculine plural nominative possessed
#PS – Perfect verb suffix	
PS_1_PL	1st person plural
PS_1_SG	1st person singular
PS_2_FEM_SG	2nd person singular
PS_2_MASC_PL	2nd person masculine plural
PS_2_MASC_SG	2nd person masculine singular
PS_3_FEM_DU	3rd person feminine dual
PS_3_FEM_PL	3rd person feminine plural
PS_3_FEM_SG	3rd person singular
PS_3_MASC_DU	3rd person masculine dual
PS_3_MASC_PL	3rd person masculine plural
PS_3_MASC_SG	3rd person masculine singular

INTENTIONALLY LEFT BLANK.

Appendix B. Dependency Labels List

The following is a complete list of the dependency labels that appear within the AADT.

adv Adverbial modifier; includes adverbs as well as other words used adverbially including various preposition-like nouns and other words.

advcl Adverbial clause or adverbial clause-like structure.

amod Adjectival modifier.

appos Apposition. (Ryding, 2005; pp. 224–227)

cc Connects a coordinating conjunction with a preceding conjunct. Note that lakinna, a sister of inna that translates as “however” or “but,” may occur with this dependency despite being labeled as an INNA and not a CONJ.

ccomp Clausal complement.

combo Combination. This is currently only used with a handful of multi-word coordinating conjunction expressions.

conj Connects a conjunct with a preceding coordinating conjunction or conjunct.

cop Complement of copula.

dep Other or unknown dependency.

det Determiner.

fidafa “False” idafa. Unlike typical idafa constructions, which are headed by nouns, these are headed by adjectives. (Ryding, 2005; pp. 221–223)

flat Flat. This is used for names and similar phenomena that lack syntactic structure, or at least any syntactic annotation.

icc Initial coordinating conjunction. Arabic sentences frequently begin with a coordinating conjunction. For this and similar situations, *icc* is used instead of connecting the head of the sentence to the coordinating conjunction via a *conj* dependency.

idafa Idafa construction. Note that in the AADT, the "Quotation or title relationship"'(cf. Ryding, 2005; p.210) is treated as apposition rather than idafa.

intj Interjection.

iobj Indirect object.

kccomp Clausal complement of kAna. This is separated from the more general *ccomp* dependency because the verb kAna, following another verb, can be used to express continued or habitual action in the past (Ryding, 2005; pp. 446–447).

neg Negation.

obj Direct object.

ocomp Object complement.

parataxis Parataxis. Used to connect to sentences together that are written next to each other but that are not connected by an explicit coordinating conjunction.

part Particle modifier. Used with a variety of different particles, including the future particle. Note that NEG_PART will typically appear with a *neg* dependency and FOCUS_PART is treated as if it were a preposition.

pcomp Object/complement of a true preposition.

prep Preposition modifier. Links a true preposition to its governor.

punct Punctuation.

reladv Relative adverbial modifier.

ricomp Complement of a relative or interrogative pronoun/adverb.

sc Subordinating conjunction. Used with subordinating conjunctions other than “inna and her sisters.”

subj Subject. This may occur without a verb, as with equational sentences.

tmz Tamyiiz. (Ryding, 2005)

tpc Topicalized element (not including topicalized subjects).

voc Vocative.

xrrcl Relative clause with an explicit relativizer.

zrrcl Relative clause without an explicit relativizer (zero relativizer).

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CIO LL
IMAL HRA MAIL & RECORDS
MGMT

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

1 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CII T
STEPHEN TRATZ