Technical report

Statistics and inter-annotator agreement calculations of the Palestinian dialect corpus –Curras

Mustafa Jarrar Faeq Alrimawi

{mjarrar, falrimawi}@birzeit.edu

Birzeit University

Palestine

This report presents various calculations and statistics about the Palestinian dialect (Curras) that we presented in this article (http://www.jarrar.info/publications/JHRAZ17.pdf). First, we present the calculations conducted to evaluate the inter-annotator agreement. Second, we present stop words and the most frequent words in the corpus.

Cite: Mustafa Jarrar, Faeq Alrimawi: Statistics and inter-annotator agreement calculations of the Palestinian dialect corpus -Curras. Technical Report. Birzeit University, Palestine, August, 2015.

Inter-annotator agreement (kappa)

Prefix

The table below shows the prefix tags that annotators used. The table also shows information about agreement and disagreement between the annotators.

Tag	A1	A2	Agreement	Disagreement	A1×A2
NULL*	1,007	1,013	996	28	1,020,091
DET	257	261	257	4	67,077
PROG_PART	84	83	82	3	6,972
IV3MS	75	75	72	6	5,625
CONJ	67	58	56	13	3,886
IV1S	42	35	33	11	1,470
IV3FS	35	34	32	5	1,190
PREP	35	41	32	12	1,435
IV1P	17	14	14	3	238
IV2MS	10	15	7	11	150
IV3P	6	2	2	4	12
DEM_PRON	3	3	3	0	9
IV2FS	3	3	3	0	9
IV3MP	3	7	3	4	21
JUS_PART	3	0	0	3	0
FUT_PART	2	1	1	1	2
EMPHATIC_PART	1	1	0	2	1
IV2MP	1	1	1	0	1
IV3FP	1	1	1	0	1
PART	1	4	0	5	4
Sum	**1,653****	**1,653**	**1,595**	**116**	**1,108,194**

*NULL: word has no prefix

**This number is higher than the number of words annotated (1,529) since some words might have more than one prefix

From this table we can calculate the kappa value for the prefix as follows:

Observed agreement =

$$\frac{\sum \text{agreements}}{\sum \text{items}} = \frac{1{,}595}{1{,}653} = 0.9649123$$

Expected agreement =

$$\frac{\sum A1 \times A2}{(\sum items)^2} = \frac{1{,}108{,}194}{1{,}653^2} = 0.4055740$$

Kappa =

$$\frac{observed\ agreement - expected\ agreement}{1 - expected\ agreement} = \frac{0.9649123 - 0.4055740}{1 - 0.4055740} = 0.9409721$$

Complex prefix

Tag	A1	A2	Agreement	Disagreement	A1×A2
NULL	1,006	1,010	996	24	1,016,060
DET	237	241	237	4	57,117
IV3MS	47	47	47	0	2,209
CONJ	45	37	35	12	1,665
PREP	27	33	24	12	891
PROG_PART+IV1S+	22	18	16	8	374
PROG_PART+IV3MS+	20	21	18	5	420
PROG_PART+IV3FS+	20	17	17	3	340
IV1S	16	14	13	4	224
IV1P	12	11	11	1	132
CONJ+DET+	10	9	9	1	90
IV3FS	9	11	9	2	99
PREP+DET+	7	7	7	0	49
IV2MS	7	4	4	3	28
PROG_PART+IV1P+	4	2	2	2	8
PROG_PART+IV3P+	4	2	2	2	8
PROG_PART+IV2MS+	3	11	3	8	33

CONJ+IV3MS+	3	3	3	0	9
DEM_PRON+DET+	3	0	0	3	0
PROG_PART+IV3MP+	2	4	2	2	8
IV2FS	2	2	2	0	4
CONJ+IV3FS+	2	2	2	0	4
CONJ+PROG_PART+IV1S	2	2	2	0	4
CONJ+PROG_PART+IV3FS	2	2	2	0	4
FUT_PART+IV3MS+	2	1	1	1	2
JUS_PART+IV3FS+	2	0	0	2	0
IV3MP	1	2	1	1	2
IV2MP	1	1	1	0	1
CONJ+IV1P+	1	1	1	0	1
CONJ+PROG_PART+IV3MS	1	1	1	0	1
PREP+IV3MS+	1	1	1	0	1
PROG_PART+IV2FS+	1	1	1	0	1
PROG_PART+IV3FP+	1	1	1	0	1
IV3P	1	0	0	1	0
PART	1	0	0	1	0
CONJ+PROG_PART+IV3P	1	0	0	1	0
EMPHATIC_PART+IV3MS+	1	0	0	1	0
JUS_PART+IV1S+	1	0	0	1	0
PROG_PART+IV1MS+	1	0	0	1	0
DEMO+DET+	0	3	0	3	0
PART+IV3FS+	0	2	0	2	0
SUB_CONJ	0	1	0	1	0
EMPHATIC_PART+IV1S+	0	1	0	1	0
PART+DET+	0	1	0	1	0
PART+IV3MS+	0	1	0	1	0
CONJ+PROG_PART+IV3MP	0	1	0	1	0
Sum	**1,529**	**1,529**	**1,471**	**116**	**1,078,432**

Observed agreement =

$$\frac{\sum \text{agreements}}{\sum \text{items}} = \frac{1{,}471}{1{,}529} = 0.9620667$$

Expected agreement =

$$\frac{\sum A1 \times A2}{(\sum items)^2} = \frac{1{,}078{,}432}{1{,}529^2} = 0.4612940$$

Kappa =

$$\frac{observed\ agreement - expected\ agreement}{1 - expected\ agreement} = \frac{0.9620667 - 0.4612940}{1 - 0.4612940} = 0.9295844$$

Stem

For the stem part, the tags that were used by the annotators are shown in the table below.

Tag	A1	A2	Agreement	Disagreement	A1×A2
NOUN	637	631	604	60	401,947
IV	191	187	184	10	35,717
ADJ	106	113	93	33	11,978
PREP	97	106	91	21	10,282
NOUN_PROP	97	76	75	23	7,372
PV	59	61	55	10	3,599
CV	53	55	50	8	2,915
INTERJ	30	35	26	13	1,050
VOC_PART	21	26	21	5	546
SUB_CONJ	22	22	21	2	484
NOUN_QUANT	24	20	20	4	480
PRON_1S	21	21	21	0	441
ADV	20	16	12	12	320
INTERROG_ADV	12	23	12	11	276
FOREIGN	13	21	13	8	273
REL_PRON	16	17	14	5	272
NEG_PART	20	13	13	7	260
CONJ	11	12	4	15	132
PRON_3MS	8	6	6	2	48
INTERROG_PRON	6	8	5	4	48
DEM_PRON_F	5	5	5	0	25
PRON_1P	5	5	5	0	25
PRON_2MS	4	5	4	1	20

DEM_PRON_MS	4	5	3	3	20
NOUN_NUM	6	3	3	3	18
DEM_PRON	4	4	4	0	16
PSEUDO_VERB	2	7	0	9	14
ADJ_COMP	4	3	3	1	12
PRON_2P	3	3	3	0	9
ADJ_NUM	2	2	2	0	4
FUT_PART	2	2	2	0	4
PRON_3FS	2	2	2	0	4
RESTRIC_PART	2	2	2	0	4
IV_PASS	1	1	1	0	1
PRON_3P	1	1	1	0	1
NULL	0	7	0	7	0
DEM_PRON_MP	0	2	0	2	0
PART	0	1	0	1	0
REL_ADV	11	0	0	11	0
VERB	3	0	0	3	0
DEM_PRON_P	2	0	0	2	0
PRON_2FS	1	0	0	1	0
PV_PASS	1	0	0	1	0
Sum	**1,529**	**1,529**	**1,380**	**298**	**478,617**

Suffix

Tag	A1	A2	Agreement	Disagreement	A1×A2
NULL	1060	1029	1022	45	1,090,740
NSUFF_FEM_SG	155	168	151	21	26,040
POSS_PRON_1S	34	37	33	5	1,258
CVSUFF_SUBJ:2MS	34	32	32	2	1,088
POSS_PRON_3MS	31	19	19	12	589
PRON_3MS	5	19	5	14	95
NSUFF_MASC_PL	17	17	17	0	289
NEG_PART	16	16	13	6	256
NSUFF_FEM_PL	16	16	16	0	256
PRON_1S	15	16	15	1	240
POSS_PRON_2MS	14	15	14	1	210
PVSUFF_SUBJ:1S	15	11	10	6	165
PVSUFF_SUBJ:3FS	10	11	9	3	110
POSS_PRON_3FS	11	10	10	1	110
PREP	9	9	8	2	81

POSS_PRON_1P	8	9	8	1	72
NSUFF_MASC_DU	8	8	8	0	64
IVSUFF_SUBJ:3MP	7	7	7	0	49
CVSUFF_SUBJ:2FS	5	7	5	2	35
PVSUFF_SUBJ:3MS	3	7	2	6	21
PVSUFF_SUBJ:2MS	2	7	1	7	14
PVSUFF_SUBJ:3MP	7	6	6	1	42
CASE_INDEF_ACC	6	6	6	0	36
IVSUFF_DO:2MS	6	6	6	0	36
IVSUFF_DO:3MS	6	6	5	2	36
CVSUFF_SUBJ:2MP	5	6	5	1	30
PRON_2MS	8	5	5	3	40
POSS_PRON_2FS	5	4	4	1	20
IVSUFF_DO:3FS	4	4	4	0	16
PRON_1P	4	4	3	2	16
PRON_3FS	4	4	4	0	16
POSS_PRON_2MP	3	4	2	3	12
PRON_2MP	3	4	3	1	12
PRON_2FS	2	4	2	2	8
IVSUFF_SUBJ:3FS	0	4	0	4	0
CVSUFF_DO:1S	5	3	3	2	15
IVSUFF_DO:2MP	3	3	3	0	9
IVSUFF_DO:3MP	3	3	3	0	9
IVSUFF_SUBJ:2FS	3	3	3	0	9
PVSUFF_SUBJ:1P	3	3	3	0	9
IVSUFF_DO:1S	2	3	1	3	6
CVSUFF_DO:3FS	2	2	2	0	4
CVSUFF_DO:3MS	2	2	2	0	4
IVSUFF_SUBJ:P	2	2	2	0	4
NSUFF_FEM_DU	2	2	2	0	4
POSS_PRON_3MP	2	2	2	0	4
PVSUFF_SUBJ:2FS	2	2	2	0	4
PVSUFF_SUBJ:2MP	2	2	2	0	4
PVSUFF_DO:2MS	1	2	0	3	2
PVSUFF_DO:3FS	1	2	1	1	2
CVSUFF_DO:1P	1	1	1	0	1
IVSUFF_DO:2FS	1	1	1	0	1
IVSUFF_SUBJ:2MP	1	1	1	0	1
IVSUFF_SUBJ:3FP	1	1	1	0	1
PVSUFF_DO:2MP	1	1	1	0	1
CASE_DEF_ACC	0	1	0	1	0
PVSUFF_DO:1S	0	1	0	1	0
PVSUFF_DO:3MS	0	1	0	1	0
POSS_PRON_2P	2	0	0	2	0
CASE_DEF_GEN	1	0	0	1	0
Sum	**1,581**	**1,581**	**1,496**	**170**	**1,122,196**

Complex suffix

Tag	A1	A2	Agreement	Disagreement	A1×A2
NULL	1,045	1,028	1,002	69	1,074,260
NSUFF_FEM_SG	147	157	141	22	23,079
POSS_PRON_1S	30	32	29	4	960
POSS_PRON_3MS	27	15	15	12	405
CVSUFF_SUBJ:2MS	25	19	14	16	475
NSUFF_MASC_PL	16	17	16	1	272
NSUFF_FEM_PL	15	15	15	0	225
POSS_PRON_2MS	13	13	13	0	169
PVSUFF_SUBJ:1S	13	9	8	6	117
POSS_PRON_3FS	11	10	10	1	110
PRON_1S	10	11	10	1	110
PVSUFF_SUBJ:3FS	9	10	8	3	90
PVSUFF_SUBJ:3MS	9	10	2	15	90
NSUFF_MASC_DU	8	8	8	0	64
POSS_PRON_1P	8	8	8	0	64
NEG_PART	8	5	5	3	40
IVSUFF_SUBJ:P	8	2	2	6	16
PVSUFF_SUBJ:3MP	7	6	6	1	42
CASE_INDEF_ACC	6	6	6	0	36
IVSUFF_DO:3MS	6	5	5	1	30
PRON_2MS	6	5	5	1	30
PRON_3MS	5	19	5	14	95
IVSUFF_DO:2MS	5	5	5	0	25
POSS_PRON_2FS	5	4	4	1	20
PRON_1P	3	4	3	1	12
IVSUFF_DO:3FS	3	3	3	0	9
PRON_3FS	3	3	3	0	9
PREP+PRON_1S	3	3	3	0	9
NSUFF_FEM_SG+POSS_PRON_3MS	3	3	3	0	9
CASE_DEF_NOM	3	0	0	3	0
IVSUFF_DO:2P	3	0	0	3	0
PVSUFF_SUBJ:2MS	2	6	1	6	12
CVSUFF_SUBJ:2FS	2	4	2	2	8
IVSUFF_DO:3MP	2	3	2	1	6
PRON_2FS	2	3	2	1	6
NSUFF_FEM_SG+POSS_PRON_1S	2	3	2	1	6
IVSUFF_SUBJ:2FS	2	2	2	0	4
NSUFF_FEM_DU	2	2	2	0	4
POSS_PRON_2MP	2	2	1	2	4

PVSUFF_SUBJ:1P+NEG_PART	2	2	2	0	4
CVSUFF_SUBJ:2FS+CVSUFF_DO:1S	2	2	2	0	4
POSS_PRON_1S+NEG_PART	2	2	2	0	4
CVSUFF_SUBJ:2MS+CVSUFF_DO:3MS	2	1	1	1	2
CVSUFF_SUBJ:2P	2	0	0	2	0
POSS_PRON_3P	2	0	0	2	0
CVSUFF_SUBJ:2P+CVSUFF_DO:3FS	2	0	0	2	0
CVSUFF_SUBJ:2MS+CVSUFF_DO:1S	2	0	0	2	0
PRON_2MP	1	3	1	2	3
PVSUFF_SUBJ:2MP	1	2	1	1	2
NSUFF_FEM_SG+POSS_PRON_2MP	1	2	1	1	2
CVSUFF_DO:1P	1	1	1	0	1
IVSUFF_DO:2FS	1	1	1	0	1
PVSUFF_DO:2MS	1	1	0	2	1
PVSUFF_SUBJ:1P	1	1	1	0	1
PVSUFF_SUBJ:2FS	1	1	1	0	1
PREP+PRON_3FS	1	1	1	0	1
PVSUFF_SUBJ:3FS+NEG_PART	1	1	1	0	1
PVSUFF_SUBJ:1S+PVSUFF_DO:3FS	1	1	1	0	1
IVSUFF_DO:3FS+NEG_PART	1	1	1	0	1
IVSUFF_SUBJ:2FS+NEG_PART	1	1	1	0	1
PVSUFF_SUBJ:2FS+PREP+PRON_1S	1	1	1	0	1
CVSUFF_SUBJ:2FS+PREP+PRON_1S	1	1	1	0	1
NSUFF_FEM_PL+POSS_PRON_3MS	1	1	1	0	1
NSUFF_FEM_SG+POSS_PRON_2MS	1	1	1	0	1
CASE_DEF_GEN	1	0	0	1	0
IVSUFF_DO:1S	1	0	0	1	0
IVSUFF_DO:3P	1	0	0	1	0
IVSUFF_SUBJ:FP	1	0	0	1	0
POSS_PRON_2P	1	0	0	1	0
PRON_2P	1	0	0	1	0
PVSUFF_SUBJ:2P	1	0	0	1	0
NSUFF_FEM_SG+POSS_PRON_2P	1	0	0	1	0
IVSUFF_SUBJ:P+IVSUFF_DO:1S	1	0	0	1	0
PVSUFF_SUBJ:3MS+PREP+PRON_2MS	1	0	0	1	0
PRON_1P+NEG_PART	1	0	0	1	0
PREP+PRON_2MS	1	0	0	1	0
NSUFF_MASC_PL_GEN	1	0	0	1	0
PREP+PRON_2P	1	0	0	1	0
PVSUFF_SUBJ:1S+PVSUFF_DO:2P	1	0	0	1	0
CVSUFF_SUBJ:2P+CVSUFF_DO:1S	1	0	0	1	0

IVSUFF_SUBJ:P+IVSUFF_DO:2MS	1	0	0	1	0
IVSUFF_SUBJ:3MP	0	5	0	5	0
IVSUFF_DO:2MP	0	3	0	3	0
IVSUFF_SUBJ:3FS	0	3	0	3	0
PVSUFF_SUBJ:3MS+NEG_PART	0	3	0	3	0
CVSUFF_SUBJ:2MP	0	2	0	2	0
POSS_PRON_3MP	0	2	0	2	0
CVSUFF_SUBJ:2MP+CVSUFF_DO:3FS	0	2	0	2	0
CVSUFF_SUBJ:2MS+IVSUFF_DO:1S	0	2	0	2	0
NSUFF_FEM_SG+CASE_DEF_NOM	0	2	0	2	0
CASE_DEF_ACC	0	1	0	1	0
IVSUFF_SUBJ:2MP	0	1	0	1	0
IVSUFF_SUBJ:3FP	0	1	0	1	0
PVSUFF_DO:3FS	0	1	0	1	0
PVSUFF_SUBJ:3MS+PVSUFF_DO:3MS	0	1	0	1	0
PVSUFF_SUBJ:2MS+PVSUFF_DO:1S	0	1	0	1	0
IVSUFF_SUBJ:3MP+IVSUFF_DO:1S	0	1	0	1	0
PREP+POSS_PRON_2MS	0	1	0	1	0
PVSUFF_SUBJ:3MS+PVSUFF_DO:2MS	0	1	0	1	0
POSS_PRON_1P+NEG_PART	0	1	0	1	0
PREP+PRON_2FS	0	1	0	1	0
PREP+PRON_2MP	0	1	0	1	0
PVSUFF_SUBJ:1S+PVSUFF_DO:2MP	0	1	0	1	0
CVSUFF_SUBJ:2MP+CVSUFF_DO:1S	0	1	0	1	0
IVSUFF_SUBJ:3FS+IVSUFF_DO:3MS	0	1	0	1	0
IVSUFF_SUBJ:3MP+IVSUFF_DO:2MS	0	1	0	1	0
CVSUFF_SUBJ:2MP+CVSUFF_DO:3MS	0	1	0	1	0
Sum	**1529**	**1529**	**1394**	**270**	**1100908**

Observed agreement =

$$\frac{\sum \text{agreements}}{\sum \text{items}} = \frac{1{,}394}{1{,}529} = 0.911706998$$

Expected agreement =

$$\frac{\sum A1 \times A2}{(\sum items)^2} = \frac{1100908}{1{,}529^2} = 0.470907987$$

Kappa =

$$\frac{observed\ agreement - expected\ agreement}{1 - expected\ agreement} = \frac{0.911706998 - 0.470907987}{1 - 0.470907987}$$
$$= 0.83312354$$

POS

Tag	A1	A2	Agreement	Disagreement	A1×A2
noun	650	653	623	57	424,450
verb	316	310	302	22	97,960
adj	98	113	85	41	11,074
prep	98	109	94	19	10,682
noun_prop	96	79	75	25	7,584
pron	48	43	43	5	2,064
interj	29	30	23	13	870
part_voc	21	26	21	5	546
conj_sub	22	22	21	2	484
noun_quant	24	20	20	4	480
adv	20	16	12	12	320
adv_interrog	12	23	12	11	276
pron_rel	16	17	15	3	272
part_neg	20	13	13	7	260
pron_dem	12	14	9	8	168
conj	11	12	4	15	132
pron_interrog	6	9	6	3	54
noun_num	6	3	3	3	18
verb_pseudo	2	7	0	9	14
adj_comp	4	3	3	1	12
adj_num	3	2	2	1	6
part_restrict	2	2	2	0	4
part_fut	2	1	1	1	2
abbrev	0	0	0	0	0
adv_rel	11	0	0	11	0
part	0	2	0	2	0
Sum	**1,529**	**1,529**	**1,389**	**280**	**557,732**

Person

Tag	A1	A2	Agreement	Disagreement	A1×A2
na	1,174	1,182	1,172	12	1,387,668
3	165	169	154	26	27,885
1	102	91	85	23	9,282
2	88	87	73	29	7,656
Sum	**1,529**	**1,529**	**1,484**	**90**	**1,432,491**

Aspect

Tag	A1	A2	Agreement	Disagreement	A1×A2
na	1,221	1,225	1,219	8	1,495,725
i	191	190	184	13	36,290
p	62	63	56	13	3,906
c	55	51	47	12	2,805
Sum	**1,529**	**1,529**	**1,506**	**46**	**1,538,726**

Gender

Tag	A1	A2	Agreement	Disagreement	A1×A2
m	1,050	886	862	212	930,300
na	228	356	212	160	81,168
f	251	287	237	64	72,037
Sum	**1,529**	**1,529**	**1,311**	**436**	**1,083,505**

Number

Tag	A1	A2	Agreement	Disagreement	A1×A2
s	1,161	1,135	1,095	106	1,317,735
na	225	268	209	75	60,300
p	132	119	108	35	15,708
d	11	7	7	4	77
Sum	**1,529**	**1,529**	**1,419**	**220**	**1,393,820**