

Beyond Gender, Number, and Rationality: Modeling Extended Arabic Morpho-Syntactic Agreement with Aggregation

Bailasan Zaina^{1,2}

Salam Khalifa^{1,3}

Nizar Habash¹

¹Computational Approaches to Modeling Language Lab, New York University Abu Dhabi

²Georgetown University

³Stony Brook University

btz3@georgetown.edu, {salam.khalifa, nizar.habash}@nyu.edu

Abstract

Arabic gender–number morpho-syntactic agreement exhibits systematic variation beyond functional gender, number, and rationality (GNR) (Alkuhlani and Habash, 2011). We identify and classify agreement patterns involving plural-to-unit and unit-to-members shifts, and introduce *aggregation* (AGG), a lexeme-level feature and corpus-derived lexical-semantic resource that captures these extended agreement affordances. We extend the GNR agreement framework with aggregation and evaluate it on over one million agreement pairs from ATB1/2/3 and BAREC. AGG improves agreement prediction over a strong baseline by 1.4% absolute on average (38% error reduction). Remaining errors largely reflect analyzer, parser, and treebank limitations. Interpretable agreement assessment of this kind can support the development of pedagogical applications and benchmarking the grammaticality of generative model outputs.¹

1 Introduction

Arabic is a morphologically rich language whose words enter into agreement and assignment relations over several features, including gender, number, case, and definiteness (Habash, 2010). This paper focuses on gender and number agreement.

Basic Arabic agreement largely follows feature *matching*: the target (e.g. an adjective or verb) matches the noun that controls it in gender and number, with the well-known exception that irrational (non-human) plurals default to feminine-singular agreement. As Alkuhlani and Habash (2011) showed, what matters is the controller’s *functional* (grammatical) gender and number rather than its surface form, since broken plurals and similar forms diverge from the features they control. Together, functional gender, number, and rationality

C	D	A	A	G	Aggregation
I	I	I	R	R	Rationality
S	S	P	P	P	Number
M	M	M	M	M	Gender
الشجر	الفرق	الكتب	الوزراء	العرب	← controller
trees	team	books	ministers	Arabs	↓ target
✓	✓				MS قال
		✓		+	FS قالت
	+		✓	✓	MP قالوا
✓	✓				MS مهم
		✓		+	FS مهمة
	+		✓	✓	MP مهمون

Table 1: Agreement options for five controllers and two targets. A single ✓ marks default gender-number-rationality (GNR) agreement; + marks an extended agreement option not predicted by default GNR. Abbreviations: M masculine, F feminine, S singular, P plural, R rational, I irrational; A atomic, C collective, G groupable, D divisible.

(GNR) predict the default agreement target in most cases. Some controllers, however, license agreement options that fall outside this default. Table 1 previews the problem with five definite, masculine controllers: most of their agreement options follow from GNR (✓), but a few are extended options (+) that GNR cannot predict, such as a plural construed as a single unit or a unit accessed through its members. These extended options are the focus of this paper, and we model them with a new lexeme-level feature, *aggregation* (AGG).

Modeling interpretable agreement is useful for pedagogical applications (Saddiki et al., 2018) and for benchmarking the grammatical output of generative models (Alrajhi et al., 2022; Zbib et al., 2026).

Our contributions are threefold. **First**, we introduce aggregation (AGG) as a new lexical agreement feature for Arabic computational linguistics, together with a corpus-derived AGG lexi-

¹Data and Code are publicly available: https://github.com/CAMEL-Lab/Arabic_GNRA_Lexicon

con for noun entries; and we implement an extended agreement rule system that captures both default and extended agreement options. **Second**, we broaden the empirical scope of Arabic agreement validation by expanding the construction inventory beyond the three phenomena targeted in prior work (verb-subject agreement, noun-adjective agreement, and numeral-noun agreement) to include subject-verb agreement, demonstratives, relative pronouns, subject-predicate constructions, *ḥāl* (circumstantials), ordinals, and fine-grained numeral-noun subtypes. We also refine the treatment of numeral-noun agreement by distinguishing fine-grained numeral subtypes. **Finally**, we conduct a large-scale evaluation of the above mentioned constructions over more than one million agreement pairs across numerous genres.

2 Linguistic Background

2.1 Agreement, Controllers, and Targets

Arabic distinguishes feature *assignment* (e.g. a verb assigning case to its subject) from feature *agreement* (a target reflecting a controller’s features); we model agreement only. The basic relevant controller features are *gender* (masculine/feminine), *number* (singular/dual/plural), and *rationality* (rational/human vs. irrational/nonhuman). Each noun carries *form* features, which describe the surface morphology of the observed wordform, and *functional* features, which represent the agreement-relevant gender and number values used by agreement targets. The two diverge for broken plurals and similar forms, so agreement is best graded against functional features.

2.2 Simple Match and GNR Agreement

We organize agreement prediction into three families, realized as five configurations (§5). **Simple match** copies the controller’s gender and number to the target, using either form or functional features. **GNR** agreement applies construction-specific rules over gender, number, and rationality, again over form or functional features; rationality is what derives the deflected feminine-singular agreement of irrational plurals (e.g. كلاب *kilāb* ‘dogs’ controls feminine singular), and functional-feature GNR is the strongest configuration in prior work. The third family, which we call **GNRA**, extends GNR with aggregation; we introduce it in §4 after motivating the gap it fills.

2.3 Agreement Constructions

Alkuhlani and Habash (2011) modeled three constructions (noun–adjective, verb–subject, and numeral–noun). We replicate their constructions and extend with six additional constructions that span adjectival, verbal, pronominal, and numeral targets and differ in which features they copy. All nine constructions are presented in Table 2

2.4 The problem of Extended Agreement

The agreement families above describe both default agreement and the additional (+) options shown in Table 1. What remains missing is a way to characterize which lexemes license these additional options and what semantic construals they make available. In such cases, a rational plural may be construed as one non-individuated unit, while a unit-like group noun may be accessed through its individual members. Traditional Arabic grammar already recognizes related noun classes, especially اسم الجمع *ism al-jam* ‘collective/group noun’ and اسم الجنس الجمعي *ism al-jins al-jam* ‘collective generic noun’; the latter is often discussed together with اسم الوحدة *ism al-waḥda* ‘singulative’, which names a single unit/member of the class denoted by a collective generic noun. These traditional categories show that Arabic grammar distinguishes group, collective, and individual-unit readings, but they do not by themselves predict the agreement alternations targeted here. The additional options are not predictable from gender, number, and rationality alone; they depend on the lexical-semantic properties of the noun, which we formalize as *aggregation* in §4.

3 Related Work

3.1 Arabic Agreement Modeling

Arabic’s templatic and concatenative morphology motivates a distinction between form-based and functional morphology (Habash, 2010; Alkuhlani and Habash, 2011). Alkuhlani and Habash (2011) enriched the Penn Arabic Treebank (Part 3) with functional gender, number, and rationality and modeled three constructions, showing that agreement cannot be reduced to surface feature copying; Mar-ton et al. (2011) use such functional features to improve dependency parsing. Alkuhlani and Habash (2011)’s line of work is the closest to our work: we replicate its form/function and GNR rule sets, but extend it along three axes (more constructions,

Construction	Example	Controller	Target	Agreement Type
Noun-Adj	^B الجميلة ^A البنت <i>al-bint^A al-jamlā^B</i> 'the pretty ^B girl ^A ,	A: Noun	B: Adj (Mod)	gender, number
Verb-Subj	^B الأولاد ^A قرأ <i>qara'a^A al-awlād^B</i> 'the boys ^B read ^A ,	B: Noun (Subj)	A: Verb	gender; number=sg
Subject-Verb	^B قرأوا ^A الأولاد <i>al-awlād^A qara'ū^B</i> 'the boys ^A read ^B ,	A: Noun (Subj)	B: Verb	gender, number
Subject-Pred	^B كبير ^A البيت <i>al-bayt^A kabīr^B</i> 'the house ^A is big ^B ,	A: Noun (Subj)	B: Pred	gender, number
Numeral-Noun	^B أعضاء ^A خمسة <i>khamṣat^A a'dā^B</i> 'five ^A members ^B ,	B: Noun (counted)	A: Numeral	gender polarity
Demonstrative	^B الكتاب ^A هذا <i>hādhā^A al-kitāb^B</i> 'this ^A book ^B ,	B: Noun	A: Dem	gender, number
Relative	^B الذي ^A الرجل <i>al-rajul^A alladhī^B</i> 'the man ^A who ^B ,	A: Noun	B: Rel. pron	gender, number
Ḥāl	^C ضاحكا ^B الرجل ^A جاء <i>jā'a^A al-rajul^B ḍāḥikan^C</i> 'the man ^B came ^A laughing ^C ,	B: Noun (Subj)	C: Ḥāl	gender, number
Ordinal	^B الثاني ^A الفصل <i>al-faṣl^A al-thānī^B</i> 'the second ^B chapter ^A ,	A: Noun	B: Ordinal	gender, number

Table 2: The nine agreement constructions evaluated. Each example tags its words with superscripts (^A, ^B, ^C) that link to the Controller and Target columns. The subject-predicate family also covers كان *kāna*- and إِنَّ *'inna*-type predications.

	KH'2011	This paper
Inflectional Features	Form Gen Num Func Gen Num	From Gen Num Func Gen Num
Lexical Features	Rationality -	Rationality Aggregation
Constructions	Three	Nine
Datasets	PATB Part 3 -	PATB Parts 1/2/3 BAREC
Agreement Pairs	39,561	1,021,686

Table 3: Comparing our work with Alkuhlani and Habash (2011) (KH'2011) across different dimensions.

larger and cross-genre corpora, and aggregation) and address the extended-agreement cases (the + cells of Table 1) that it does not model. See Table 3 for full comparison.

3.2 Semantic Grounding of Aggregation

AGG is not a new semantic theory but a computationally usable operationalization of distinctions long discussed for nominal denotation: atomicity, collectivity, individuation, and group- versus member-level access (Chierchia, 1998; Pearson,

2011; D'Anna, 2020). Pearson (2011) shows that group nouns display both unit-level and member-level behavior and that "committee" nouns differ from "collection" nouns, so collectivity alone does not license member access. For Arabic, D'Anna (2020) treats agreement with collectives as reflecting holistic, distributive, or non-individuated construals of the controller. We are inspired by this literature in defining aggregation (§4). However, our contribution is a computational, lexeme-level agreement feature rather than a semantic theory.

4 Aggregation

GNR specifies the *default* agreement pattern of a lexeme, but, as §2.4 showed, some nouns license additional options. We capture this with **Aggregation (AGG)**, a lexeme-level feature on nouns that is independent of gender, number, and rationality and characterizes how a lexeme's denotation is internally organized. GNR gives the default agreement; AGG specifies the affordances for *extended* agreement. AGG has four values for nouns, schematized in Figure 1. The figure separates nouns that remain

AGG	Default Construal	Extra Construal	Extra
A	individuated entity	–	–
C	non-individuated aggregate	–	–
G	rational plural individuals	collective / non-individuated unit	FS
D	unit-like group	rational/human members	MP

Table 4: Directionality of the two reactive AGG values. Groupable construes a rational plural as one unit (extra FS); Divisible accesses a unit’s members (extra MP). Atomic and Collective license no extra agreement.

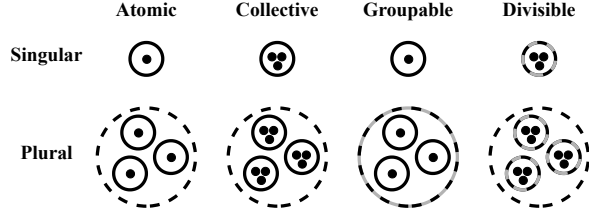


Figure 1: The four AGG values as denotational schemata, crossed with singular and plural. A grey ring is a *coercible* boundary licensing extended agreement: Atomic and Collective have none; Groupable adds a grey envelope over a rational plural ($\rightarrow$ FS); Divisible opens a unit to its members ($\rightarrow$ MP).

within default GNR agreement from nouns that license an additional directional construal. **Atomic** (A) and **Collective** (C) nouns license no agreement beyond GNR: an Atomic noun is an ordinary individuated entity, while a Collective noun is a non-individuated aggregate whose members are not independently accessible for agreement; member reference instead uses a singulative. The two *extended-agreement* values carry a directionality (Table 4). **Groupable** (G) lexemes have a rational-plural default but allow that plural to be construed as a single non-individuated unit, additionally licensing feminine-singular agreement (MP $\rightarrow$ FS). **Divisible** (D) lexemes have a unit-like default but allow access to their constituent members, additionally licensing masculine plural agreement when those members are rational/human (MS/FS $\rightarrow$ MP). Because AGG is lexeme-level, a value characterizes the whole lexeme even when the affordance surfaces only on particular number forms. Table 5 summarizes the full AGG inventory, including *Not Applicable* (N) for non-nouns. The agreement-reactive nouns are exactly the Groupable and Divisible cases.

The example in Figure 2 shows a Divisible controller in context: the noun المجموعة *al-majmū’a* ‘the group’ takes default feminine-singular agreement on the adjective الأخيرة *al-’axīra* ‘the last’ (from GNR), but member-access masculine-plural agreement on the verb فازوا *fāzū* ‘won’, because the group is accessed through its individual members.

Controller			Target Agreement			
Lemma	Gloss	Base	G N R A	GNR	GNRA	
موظف	employee	موظفون	M P R A	MP	–	
وزير	minister	وزراء	M P R A	MP	–	
كلب	dog	كلاب	M P I A	FS	–	
شجر	trees	أشجار	M P I C	FS	–	
رجل ¹	man	رجال	M P R G	MP	FS	
رجل ²	leading figure	رجال	M P R G	MP	FS	
عربي	Arab	عرب	M P R G	MP	FS	
ناس	people	ناس	F S I D	FS	MP	
قوم	nation	قوم	M S I D	MS	MP	
فريق	team	فريق	M S I D	MS	MP	

Table 5: Aggregation values for example controllers, grouped by aggregation value (A, C, G, D). A and C are *non-AGG-active* (default-only); G and D are *AGG-active*. “GNR” is the default agreement for the cited baseword, and “GNRA” is the extra agreement that Groupable and Divisible additionally license (“–” = none).

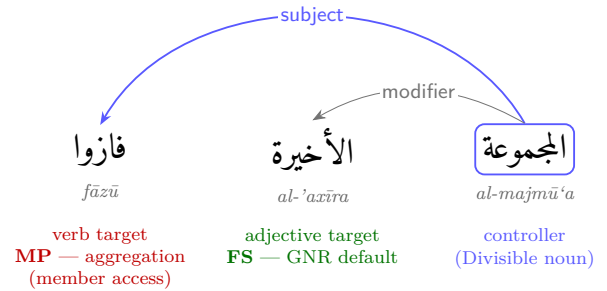


Figure 2: Example with mixed agreement patterns.

5 Methodology

5.1 Rule Configurations

We evaluate five configurations, the first four replicating Alkuhlani and Habash (2011). **Simple match** comprises (1) *Form Match* and (2) *Func Match*, where the target copies the controller’s form or functional gender and number. **GNR** comprises (3) *Form Rules* and (4) *Func Rules*, construction-specific agreement rules over gender, number, and rationality applied to form or functional features;

Func Rules is the strongest replicated baseline. The fifth, **GNRA** or (5) *Extended*, adds aggregation. All five are evaluated on ATB1/2/3 and BAREC; the main comparison is Func Rules vs. Extended. The full extended rule inventory is in Table 6.

5.2 The Extended System

The extended rule set preserves the same mapping architecture as the baselines: rules take controller-side features as input and output the agreement features licensed on the target. The baseline rules use the controller’s form features $F(N)$ or functional GNR features $G(N)$ to derive the default; the extended system keeps the functional GNR mappings as the default layer and adds further controller-side inputs, including, but not limited to, aggregation, to license additional target mappings (Table 6). Thus AGG is used as an input condition, parallel to gender, number, and rationality, not as an after-the-fact interpretation of errors. AGG is the central new feature of the layer but not its entirety. A controller–target pair is scored as correct when the gender–number features observed on the target are among the agreement options licensed by the corresponding rule configuration for that controller; otherwise it is scored as incorrect. A configuration may license more than one target option, and the observed target need only belong to the licensed set.

Separately, a small class of 24 lemmas (e.g. سوق *sūq* ‘market’, طريق *ṭarīq* ‘road’) admits two valid genders. We treat this as a lexical-resource matter: such nouns receive parallel masculine and feminine entries, so either agreement pattern counts as correct. This is not an aggregation feature, but an independent lexical alternation that feeds the GNRA agreement rules.

5.3 The AGG Lexicon

We developed a targeted, corpus-derived lexical-semantic AGG lexicon. Candidate nouns were extracted automatically from the evaluated ATB and BAREC agreement pairs and flagged when their features or observed agreement behavior suggested an alternation not derivable from gender, number, and rationality alone. We manually reviewed the candidates and distinguished AGG-reactive nouns, namely Groupable and Divisible lexemes, from default (non-reactive) nouns, including Atomic and Collective cases as well as other inanimate or abstract nouns whose agreement remains derivable

from GNR. Concretely, a lemma is Groupable if its rational plural can control feminine-singular unit agreement, and Divisible if its singular unit form can control masculine-plural member agreement. Annotation is lexeme-level, and cases with unstable judgments were flagged and kept out of the central evidence.

The primary annotation was carried out by one author, a native Arabic-speaking PhD researcher in Arabic linguistics with relevant training and prior annotation experience. Following a review of the literature, we piloted the annotation on approximately 3,000 automatically extracted noun candidates in their agreement contexts; the pilot led us to refine the initial form-based labels into the final lemma-sensitive Atomic/Collective/Groupable/Divisible representation. Candidates were reviewed in context and classified using corpus evidence where available and semantic-class criteria otherwise. Difficult cases and the final category definitions were reviewed collaboratively by all three authors. The process took approximately 60–70 hours over several weeks.

The AGG-reactive portion of the resource contains 864 lemma entries. Because lemma types recur across many surface forms and agreement pairs in the evaluated data, the AGG-reactive subset is small in type coverage but high in token impact, accounting for 20.3% of noun tokens in the agreement pairs. The annotated lemmas fall into recognizable semantic clusters: Groupable lemmas are dominated by human plural classes (ethnonyms such as العرب *al-‘arab* ‘the Arabs’; religious and social groups such as اليهود *al-yahūd* ‘the Jews’; professional and status groups such as العلماء *al-‘ulamā* ‘scholars’, القضاة *al-quḍāt* ‘judges’; and kinship plurals such as الإخوة *al-‘ixwa* ‘brothers’), while Divisible lemmas are dominated by unit-denoting collectives accessed through human members (teams and groups such as فريق *farīq* ‘team’; committees and bodies such as اللجنة *lajna* ‘committee’, المجلس *majlis* ‘council’; and organizations such as حزب *ḥizb* ‘party’, وفد *wafd* ‘delegation’).

5.4 Corpus Synchronization and Data

We use two corpora for evaluation. **ATB** (ATB1/2/3) is the replication corpus, with 58,501 evaluated controller–target agreement pairs; [Alkuhlani and Habash \(2011\)](#) used only PATB Part 3. The

Controller Condition (Input)							Target Result	
Form Gen $F(N)$	Form Num $F(N)$	Form Rat $F(N)$	Func Gen $G(N)$	Func Num $G(N)$	Func Rat $G(N)$	AGG value	Adj. Agr	Member Access
*	*	*	M	S	*	*	MS	—
*	*	*	M	D	*	*	MD	—
*	*	*	M	P	R	*	MP	—
*	*	*	M	P	I	*	FS	—
*	*	*	F	S	*	*	FS	—
*	*	*	F	D	*	*	FD	—
*	*	*	F	P	R	*	FP	—
*	*	*	F	P	*	*	FS	—
*	S	*	*	P	R	G	FS	BLOCKED
*	*	*	*	*	I	D	MP	LICENSED

Table 6: **Noun–adjective extended rule inventory.** The left block gives the rule input conditions: form features ($F(N)$), functional features ($G(N)$), and AGG value; the result block gives the licensed adjective agreement. The final column is a semantic readout, not an input condition: feminine singular agreement with a Groupable (G) noun corresponds to blocked member access, and masculine plural agreement with a Divisible (D) noun to licensed member access. “*” marks an unconstrained feature value; “—” marks cases with no AGG-induced reading. Target abbreviations combine M/F (masculine/feminine) with S/D/P (singular/dual/plural); e.g., MD = masculine dual.

ATB data derives from the Penn Arabic Treebank (PATB; Maamouri et al., 2004). Its dependency view follows CATiB (Habash and Roth, 2009), converted from gold treebank annotation, and its functional morphology comes from CALIMASr and the CAMEL resources (Taji et al., 2018; Obeid et al., 2020; Khairallah et al., 2024).

BAREC is the extension corpus, based on the BAREC-10M resource (Elmadani et al., 2026). From this corpus, our extraction pipeline yields 963,185 evaluated controller–target agreement pairs. Unlike ATB, whose syntactic layer is converted from gold treebank annotation, BAREC is automatically parsed with CAMEL Parser 2.0 (Elshabrawy et al., 2023).

For each corpus, extraction aligns a syntactic layer with a morphological layer: the syntax determines which controller–target pairs exist, and the morphology supplies the form and functional features. For ATB, we synchronize the CATiB view with CALIMA analyses. For BAREC, we apply the same logic to the automatically parsed CATiB view and CAMEL Morph analyses, with an added document-level alignment stage. Synchronizing BAREC’s two views aligns 514,677 of 514,879 sentences in 20,535 matched documents (99.96%), and a further 37,745 tokens are dropped for surface-form disagreement. We do not model definiteness, which is out of scope. To assess extraction quality, we manually validated a stratified random sample of 500 retained BAREC controller–target pairs spanning all nine constructions; 4 pairs were incorrectly extracted, an estimated extraction error rate

of 0.8%. This estimate concerns extraction among retained pairs and is separate from downstream analyzer and parser noise (see §7).

6 Results

6.1 Metrics

We report accuracy for each rule configuration and error reduction from the strongest functional baseline (Func Rules) to the Extended system. Since the evaluated constructions differ greatly in size, we use a macro-average across construction–corpus rows as the headline summary, giving each row equal weight. The error reduction in the average row is computed from the macro-average accuracies, rather than by averaging the row-level error-reduction values.

6.2 Overall Results

Table 7 reports accuracy for the five configurations per construction and corpus. The contrast of Alkuhlani and Habash (2011) reproduces: simple match is weakest, GNR rules capture most cases, and the Extended system is the maximum in every row. On average, the Extended system improves over the strongest functional baseline by 1.4 points absolute ($96.2 \rightarrow 97.7$), corresponding to a 37.6% average reduction in remaining errors; on individual constructions the gain reaches 4.4 accuracy points (BAREC noun–adjective, $92.6 \rightarrow 97.0$). The gains capture a well-defined class of agreement alternations, namely collective-unit and member-access readings, that GNR alone cannot express.

Construction	Corpus	N	Simple		GNR		GNRA	Error Redux (%)
			Form Match	Func Match	Form Rules	Func Rules	Extended	
Noun–Adjective	ATB	35,153	69.1	70.9	82.3	98.1	99.1	51.8
	BAREC	542,561	66.8	66.5	77.1	92.6	97.0	59.8
Verb–Subject	ATB	11,350	76.8	75.5	86.9	95.2	95.5	4.8
	BAREC	212,667	77.8	75.1	85.2	94.3	96.4	36.4
Subject–Verb	ATB	1,520	77.2	82.6	84.5	95.8	97.0	27.8
	BAREC	27,948	78.0	82.8	83.4	98.5	99.8	88.4
Subject–Predicate	ATB	386	71.8	74.6	84.7	97.2	97.9	27.4
	BAREC	6,763	77.2	80.2	81.6	94.1	95.4	22.5
Numeral–Noun	ATB	767	16.0	4.0	84.6	96.6	96.7	3.8
	BAREC	9,513	24.2	9.3	77.5	92.6	92.9	4.3
Demonstrative	ATB	2,830	74.5	75.4	85.8	98.5	99.8	86.3
	BAREC	61,729	74.5	74.0	83.0	96.2	98.3	54.6
Relative	ATB	4,413	63.5	67.4	77.8	95.8	96.7	23.1
	BAREC	67,141	61.1	60.3	75.2	96.2	97.0	20.6
Hāl	ATB	686	92.4	94.3	92.9	96.4	99.1	75.8
	BAREC	8,146	75.8	83.8	76.4	99.6	99.7	12.8
Ordinal	ATB	1,396	95.6	97.2	95.6	97.2	99.9	94.7
	BAREC	26,717	97.1	97.7	97.1	97.7	99.7	88.0
Average			70.5	70.6	84.0	96.2	97.7	37.6

Table 7: Per-construction accuracy (%) for the five configurations on ATB and BAREC. **Form/Func Match** = target copies the controller’s form/functional gender-number; **Form/GNR Rules** = construction-specific agreement rules over form features, or over functional gender, number, and rationality; **Extended** = our system, conditioning on form, functional/GNR, and aggregation features.

BAREC Text Category	N	GNR	GNRA	Red.
Educational Materials	273,377	14,301	5,705	60.1
Literature	272,625	14,638	5,722	60.9
Media & Culture	235,528	11,271	4,722	58.1
Academic	160,586	15,702	11,060	29.6
Encyclopedic	18,406	1,107	474	57.2
Religion & Philosophy	2,098	130	56	56.9
Art & Music	565	81	54	33.3
Total	963,185	57,230	27,793	51.4

Table 8: BAREC Final-stage errors by Text Category, aggregated across all constructions. **GNR** and **GNRA** are the error counts of the GNR Func Rules baseline and the Extended system; **Red.** (%) is error reduction.

6.3 Where the Extension Helps

The Extended system improves most where a few systematic, aggregation-driven triggers dominate, as in noun–adjective, subject–verb, ordinal, and demonstrative; it improves least where the residual errors are *heterogeneous*, spread across many unrelated, lexeme- and analysis-specific causes rather than a few systematic triggers a single rule can capture (§7). Table 8 stratifies BAREC by Text Category; residual errors concentrate in *Academic* text, which is dominated by specialized vocabulary and heavier automatic-analysis noise rather than genuine agreement failures.

6.4 Specificity of the Extended System

A natural concern is whether the Extended system gains accuracy merely by licensing more alternatives. To address this, we compare, for each controller word and feature configuration, the number of distinct target gender–number options the system licenses (*permissiveness*) with the number of distinct options actually attested in the corpus for that configuration (*diversity*). A configuration is *More Specific* when the system licenses fewer options than are attested, *Exact* when the two counts coincide, and *Less Specific* when the system licenses more options than are attested. This measure compares the sizes of the licensed and attested option sets, rather than whether the individual licensed options match the attested ones; prediction accuracy is evaluated separately: it asks how tightly the licensed options track the attested alternatives, weighted by pair occurrences.

Table 9 reports occurrence-weighted results for all nine constructions on both corpora. Two patterns emerge. First, relative to the GNR baseline, the Extended system shifts part of the *More Specific* mass to *Exact*: on BAREC noun–adjective, *More Specific* decreases from 59.7% to 56.0% of pair occurrences while *Exact* rises from 40.3% to 42.4%, and the system becomes *Less Specific*

Construction	Corpus	GNR			Extended		
		More	Ex.	Less	More	Ex.	Less
Noun–Adj.	ATB	7.9	92.1	0.0	6.3	67.7	25.9
	BAREC	59.7	40.3	0.0	56.0	42.4	1.6
Verb–Subj.	ATB	9.2	90.8	0.0	6.4	78.7	15.0
	BAREC	46.9	53.1	0.0	40.6	55.8	3.6
Subj.–Verb	ATB	2.7	97.3	0.0	0.9	78.4	20.7
	BAREC	2.6	97.4	0.0	1.8	88.8	9.3
Dem.	ATB	0.8	99.2	0.0	0.8	84.4	14.8
	BAREC	20.2	79.8	0.0	18.2	75.8	6.0
Rel.	ATB	2.4	97.6	0.0	2.3	97.6	0.1
	BAREC	18.6	81.4	0.0	18.4	81.2	0.4
Subj.–Pred.	ATB	0.0	100	0.0	0.0	84.6	15.4
	BAREC	9.8	90.2	0.0	8.1	84.0	7.9
Num.–Noun	ATB	11.9	88.1	0.0	11.9	88.1	0.0
	BAREC	57.4	42.6	0.0	56.7	43.1	0.2
Ord.	ATB	1.4	98.6	0.0	1.4	95.1	3.4
	BAREC	4.1	95.9	0.0	3.5	93.1	3.4
Ḥāl	ATB	0.1	99.9	0.0	0.1	91.3	8.6
	BAREC	51.4	48.6	0.0	51.4	45.5	3.1

Table 9: Specificity of the GNR baseline and the Extended system, as percentages of controller–target pair occurrences (More/Exact/Less Specific – See §6.4).

in only 1.6% of pair occurrences. The BAREC verb–subject and numeral constructions show the same profile, whereas in constructions with little attested variation (preverbal subjects, demonstratives, predicates) the additional licensed options largely go unattested, so Exact decreases in favor of small Less Specific shares ($\leq 9.3\%$ on BAREC). Second, the two corpora distort specificity in opposite directions. ATB is roughly fifteen times smaller than BAREC, so most controllers occur with only one attested option even where the alternation is grammatical; this both suppresses ATB’s More Specific shares and counts every licensed-but-unattested option against it, yielding the systematically larger Less Specific shares (up to 25.9% for noun–adjective). Conversely, BAREC’s scale attests genuine alternation but also accumulates rare analyzer errors, and a single stray target form on a frequent controller counts its entire occurrence mass as More Specific: the ḥāl share, for instance, stems almost entirely from one high-frequency configuration attested with two stray plural tokens, whereas the noun–adjective, numeral, and verb–subject shares are dominated by controllers attested with multiple options at substantial rates.

Lexicon Source	Test Corpus	Accuracy
ATB	ATB	98.0
BAREC	ATB	98.0
ATB+BAREC	ATB	98.0
Func Rules	ATB	96.8
BAREC	BAREC	97.4
ATB	BAREC	97.3
ATB+BAREC	BAREC	97.4
Func Rules	BAREC	95.8

Table 10: Cross-corpus lexicon evaluation: macro-average accuracy (%) across the nine constructions when the AGG lexicon is restricted to vocabulary from one corpus and evaluated on each test corpus. Func Rules is the lexicon-free functional baseline.

6.5 Numeral Constructions

The numeral–noun rows behave differently from the others: Arabic numerals 3–10 take *gender polarity*, the gender opposite to the counted noun’s singular, so the numeral’s surface gender is the reverse of what plain matching expects. This is why the Simple Match columns are unusually low for numeral–noun (e.g. Func Match 4.0/9.3): copying the controller’s gender is systematically wrong under polarity, whereas the GNR and AGG configurations encode it.

6.6 Cross-Corpus Lexicon Evaluation

To assess whether the gains from AGG depend on constructing the lexicon from the same corpus used for evaluation, we restrict the AGG lexicon to vocabulary extracted from one corpus and evaluate on the other. Table 10 reports the macro-average accuracy across the nine constructions for each lexicon-source and test-corpus combination.

On BAREC, the ATB-only lexicon achieves 97.3% accuracy, compared with 97.4% for both the BAREC-derived and combined ATB+BAREC lexicons, while the Func Rules baseline reaches 95.8%. Thus, 1.5 of the 1.6 percentage-point improvement over the functional baseline is retained when BAREC does not contribute its vocabulary to lexicon construction. On ATB, all three AGG lexicon conditions yield approximately 98.0%, compared with 96.8% for Func Rules. These results indicate that the gains are not primarily dependent on deriving the AGG vocabulary from the same corpus on which it is evaluated, substantially reducing the circularity concern while not eliminating all possible corpus-derived bias.

7 Discussion and Residual Error Analysis

The Extended system improves accuracy across constructions, and the remaining errors are heterogeneous. They fall into three broad types: a small set of model-relevant hard cases, construction-specific issues, and upstream resource/analyzer errors. The last dominate: most residual disagreements are not agreement-rule failures but artifacts of the input resources, echoing Alkuhlani and Habash (2011), whose own residual analysis attributed roughly half of remaining disagreements to annotation errors and a further fifth to treebank errors.

7.1 Construction-specific Residuals

A transparent residual case is numeral polarity (above). The relevant rule is categorical: numerals 3–10 take the gender opposite to the singular counted noun. Thus, with feminine سنة *sana* ‘year’ the expected form is ست سنوات *sitt sanawāt* ‘six years’, not ست سنوات *sittat sanawāt*. The latter nonetheless occurs in the data and is counted as an error, although the deviation originates with the source text rather than with the model (a gold/source error).

7.2 Upstream Resource and Analyzer Errors

A large share of the residual is upstream noise. The largest source is missing functional gender on named entities and place names, which the analyzer defaults to masculine: the toponym إسرائيل *isrā’īl* ‘Israel’ (likewise واشنطن *wāshintun* ‘Washington’, إيران *īrān* ‘Iran’) is tagged masculine singular yet controls feminine-singular targets (أظهرت إسرائيل *‘aẓharat isrā’īl* ‘Israel showed’), so the system records a mismatch. A second family is feminine forms tagged masculine, including *ta-marbuta* nouns such as جامعة *jāmi‘a* ‘university’, حكومة *ḥukūma* ‘government’, and منطقة *manṭiqa* ‘region’. A third is broken plurals assigned their lemma’s singular features, e.g. الأحياء *al-‘ahyā* ‘the neighborhoods’, tagged masculine singular like its lemma حي *ḥayy* ‘neighborhood’ instead of masculine plural. These call for resource-level fixes, and the rule system doubles as a diagnostic that surfaces them; for named entities we report the gaps rather than patching them with ad-hoc genders.

7.3 Out-of-scope Cases

Some constructions are excluded by reproducible filters because they require phrase- or subtree-level computation rather than word-level features. One case is quantificational NPs, where agreement depends on the quantifier and the quantified noun jointly. For example, كل طالب حضر *kull ṭalib ḥaḍara* ‘every student attended’ involves singular agreement with the quantified noun, whereas كل الطلاب حضروا *kull al-ṭullāb ḥaḍarū* ‘all the students attended’ involves a plural quantified noun and plural agreement. Another case is conjunction agreement, where the target may agree with one conjunct or with the whole coordinated phrase: in جاء محمد وفاطمة *ja’a Muḥammad wa-Faṭima* ‘Muhammad and Fatima came’, the verb may show closest-conjunct agreement with محمد *Muḥammad*, or, in SV order, agree with the coordinated phrase as a dual.

8 Conclusion and Future Work

We recast Arabic agreement so that GNR gives the default and aggregation, a lexeme-level noun feature, licenses extended agreement, with Groupable and Divisible modeling the opposite directions of plural-to-unit and unit-to-members. We contribute the AGG feature, an aggregation-aware rule system, a corpus-derived AGG lexicon, and a large-scale evaluation over more than one million agreement pairs from ATB1/2/3 and BAREC, on which AGG improves agreement prediction by up to 4.4 accuracy points over GNR features.

Future work includes phrase-level phenomena such as coordination and quantification, integrating AGG into Arabic NLP resources such as CAMEL Tools, expanding lexical coverage, and applying the framework to dialectal and learner Arabic, grammatical error correction and explanation (Elhady et al., 2026), and the evaluation of MT- and LLM-generated Arabic. Future work includes phrase-level phenomena such as coordination and quantification, integrating AGG into Arabic NLP resources such as CAMEL Tools, expanding lexical coverage, applying the framework to dialectal and learner Arabic, grammatical error correction and explanation (Elhady et al., 2026), the evaluation of MT- and LLM-generated Arabic, and automatic induction of AGG values from distributional or contextual representations, reducing the need for manual lexicon expansion.

Limitations

Our evaluation depends on upstream resources, including the morphological analyzer, treebank annotations, and automatically derived structure for BAREC. As a result, some apparent agreement errors reflect resource or extraction artifacts rather than failures of the agreement model itself. Direct comparison with Alkuhlani and Habash (2011) is limited by differences in corpora and construction coverage. Finally, the AGG lexicon is targeted and corpus-derived, leaving unattested nouns unannotated and some classifications judgment-sensitive; some agreement phenomena, such as coordination and quantification, require phrase-level computation beyond the current word-level scope.

Ethics Statement

The primary annotation of the AGG lexicon was conducted by one author; difficult cases, agreement rules, and the final category definitions were reviewed collaboratively by all authors to promote consistency and reduce individual annotator bias. The resulting lexical annotations, rule definitions, and evaluation scripts will be made publicly available to support transparency, reproducibility, and further research in Arabic NLP, subject to the licensing restrictions of the underlying corpora. We used AI writing assistance only for language-level editing of the paper, as permitted by the ACL Policy on Publication Ethics; all analyses, annotations, experiments, and results were produced and verified by the authors.

Acknowledgments

We thank the anonymous reviewers for their careful reading and constructive feedback, which helped improve the clarity and presentation of this work. We also thank Khalid N. Elmadani for helpful discussions and support related to the BAREC resource. We are grateful to the members of CAMEL Lab for their feedback and support throughout this project.

References

Sarah Alkuhlani and Nizar Habash. 2011. A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 357–362, Portland, Oregon, USA. Association for Computational Linguistics.

- Wafa Abdullah Alrajhi, Hend Al-Khalifa, and Abdulmalik AlSalman. 2022. Assessing the linguistic knowledge in Arabic pre-trained language models using minimal pairs. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 185–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Gennaro Chierchia. 1998. Plurality of mass nouns and the notion of “semantic parameter”. In Susan Rothstein, editor, *Events and Grammar*, pages 53–103. Kluwer Academic Publishers, Dordrecht.
- Luca D’Anna. 2020. Collectives in the Qur’ān revisited: Another possibility of semantic agreement. *Journal of Semitic Studies*, 65(1):147–169.
- Khaled Elhady, Omar Kallas, Nizar Habash, and Bashar Alhafni. 2026. ArabiGEE: A hierarchical taxonomy for Arabic grammatical error explanation. *Preprint*, arXiv:2606.10765.
- Khalid N. Elmadani, Adel Mahmoud Wizani, Hanada Taha Thomure, and Nizar Habash. 2026. A large and balanced multi-domain arabic corpus annotated for morphology, syntax, and readability. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11761–11775, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Ahmed Elshabrawy, Muhammed AbuOdeh, Go Inoue, and Nizar Habash. 2023. CamelParser2.0: A state-of-the-art dependency parser for Arabic. In *Proceedings of ArabicNLP 2023*, pages 170–180, Singapore (Hybrid). Association for Computational Linguistics.
- Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic Treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224, Suntec, Singapore. Association for Computational Linguistics.
- Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*, volume 10 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.
- Christian Khairallah, Salam Khalifa, Reham Marzouk, Mayar Nassar, and Nizar Habash. 2024. Camel morph MSA: A large-scale open-source morphological analyzer for Modern Standard Arabic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2683–2691, Torino, Italia. ELRA and ICCL.
- Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.
- Yuval Marton, Nizar Habash, and Owen Rambow. 2011. Improving Arabic dependency parsing with form-based and functional morphological features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1586–1596, Portland, Ore-

gon, USA. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMEL tools: An open source Python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Hazel Pearson. 2011. A new semantics for group nouns. In *Proceedings of the 28th West Coast Conference on Formal Linguistics (WCCFL 28)*, pages 160–168, Somerville, MA. Cascadilla Proceedings Project.

Hind Saddiki, Nizar Habash, Violetta Cavalli-Sforza, and Muhamed Al Khalil. 2018. Feature optimization for predicting readability of Arabic L1 and L2. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–29, Melbourne, Australia. Association for Computational Linguistics.

Dima Taji, Salam Khalifa, Ossama Obeid, Fadhl Eryani, and Nizar Habash. 2018. An Arabic morphological analyzer and generator with copious features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 140–150, Brussels, Belgium. Association for Computational Linguistics.

Mohamad Bilal Zbib, Hasan Abed Al Kader Hammoud, Ammar Mohanna, Nadine Rizk, Fatima Karnib, Sina Moukaled, and Bernard Ghanem. 2026. [AraLing-Bench: A human-annotated benchmark for evaluating Arabic linguistic capabilities of large language models](#). In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 385–393, Rabat, Morocco. Association for Computational Linguistics.

A License

We list below the licenses of the data and tools used in this work, all of which are employed in accordance with their intended use.

- PATB / ATB ([Maamouri et al., 2004](#)): Linguistic Data Consortium (LDC) License
- BAREC-10M Corpus ([Elmadani et al., 2026](#)): Creative Commons Attribution Share Alike 4.0
- CAMEL Tools ([Obeid et al., 2020](#)): MIT License
- CAMEL Morph ([Khairallah et al., 2024](#)): MIT License
- CamelParser2.0 ([Elshabrawy et al., 2023](#)): MIT License