This is the final authors' version of: Hardie, A and Ibrahim, W (in press) **Exploring and classifying the Arabic copula and auxiliary *kāna* via enhanced part-of-speech tagging**. *Corpora*. In case of any difference between this version and the journal's typeset version, the latter is to be considered definitive.

Exploring and categorising the Arabic copula and auxiliary *kāna* via enhanced part-of-speech tagging

Andrew Hardie[*] and Wesam Ibrahim[†]

[*] Linguistics and English Language, Lancaster University, UK

[†]Department of Basic Sciences, Community College, Princess Nourah bint Abdulrahman University, Saudi Arabia;
and Department of Foreign Languages, Faculty of Education, Tanta University, Egypt

a.hardie@lancaster.ac.uk ; wmibrahim@pnu.edu.sa

Abstract

Arabic syntax has yet to be studied in detail from a corpus-based perspective. The Arabic copula *kāna*, 'be', functions additionally as an auxiliary, creating *periphrastic tense-aspect constructions*; but the literature on these functions is far from exhaustive. To analyse *kāna* within the million-word Leeds *Corpus of Contemporary Arabic*, part-of-speech tagging (using novel, targeted enhancements to a previously described program which improves the accessibility for linguistic analysis of the output of Habash et al.'s 2012 MADA disambiguator for the Buckwalter Arabic morphological analyser) is applied to disambiguate copula and auxiliary at a high rate of accuracy. Concordances of both are extracted, and 10% samples (499 instances of copula *kāna*, 387 of auxiliary *kāna*) are manually analysed to identify surface-level grammatical patterns and meanings. This raw analysis is then systematised according to the more general patterns' main parameters of variation; special descriptions are developed for specific, apparently fixed-form expressions (including two phraseologies which afford expression of verbal and adjectival modality). Overall, substantial new detail, not mentioned in existing grammars, is discovered (e.g. the quantitative predominance of the past imperfect construction over other uses of auxiliary *kāna*); there exists notable potential for these corpus-based findings to inform and enhance not only grammatical descriptions, but also pedagogy of Arabic as a first or second/foreign language.

1. Introduction[1]

The Arabic grammatical tradition is long-established and sophisticated (Owens, 1990, 1997). Yet in comparison to contemporary linguistic approaches to description of grammar, this tradition offers less attention to matters of *syntax* as opposed to *morphology*. Given the complexity of derivation and inflection in Arabic, this is no surprise; a similar focus on morphology over syntax is observable in other classical grammatical traditions, such as the Sanskrit (e.g. the *Aṣṭādhyāyī* of Pāṇini: Cardona, 1976) or the Greek (e.g. the *Tekhnē Grammatikē* of Dionysius Thrax: Forbes, 1933:112).

An example is the tense-aspect-mood system. Famously, in Classical and Modern Standard Arabic, verbs exhibit two main finite forms, described as perfect/imperfect aspect or

[1] Hardie's work on this paper was supported by the ESRC *Centre for Corpus Approaches to Social Science* (CASS) (grant reference ES/R008906/1). Ibrahim's work on this paper was supported by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University, through the *Fast-track Research Funding Program*.

as past/present tense; the present tense secondarily exhibits mood inflection, and likewise future tense is formed from the present by prefixation. Much research in Arabic grammar has focused on the question of whether the distinction between two main forms is one of tense or of aspect (see *inter alia* Ammann, 2002; Ouali, 2018).[2] As well as these inflections, Arabic possesses syntactically-marked tense-aspect constructions – that is, periphrastic constructions combining a main verb with some auxiliary element(s) to express some tense-aspect which is not morphologically marked on any single word. Periphrastic constructions are central to, and much-studied within, the tense-aspect systems of languages such as English and Chinese. English's periphrastic perfect and progressive aspects, and passive voice, are well-understood. But equivalent periphrastic constructions in Arabic have attracted rather less attention from most quarters. Typical pedagogical and reference grammars note their existence but say almost nothing about them. Formal and functional approaches to Arabic syntax usually, albeit not always, treat them as peripheral to questions concerning the inflected forms (see section 2).

In the age of corpus-based methodologies, description of such periphrastic constructions can feasibly be informed by a mass of natural language data. Due to the high frequency of grammatical, as opposed to lexical, units, even a relatively small corpus yields sufficient examples for empirical description of different functions and estimation of their relative frequencies. Yet such corpus-based analysis is another area in which little work has been done to date on Arabic syntax. Some studies have utilised corpus data to address closely-defined topics in Arabic grammar, notably Sartori's (2019) study of conditional sentences and the sequences of tense they exhibit. In more general work, Ryding (2005: xviii-xix,9) reports basing her reference grammar on a "database" of contemporary Arabic prose texts, which would appear to constitute a small corpus despite Ryding not labelling it as such. Her approach to this data is to treat it essentially as a repository of examples, in the sense discussed by McEnery and Hardie (2012:173). Consequently, Ryding's work includes none of the distinctive types of findings (especially frequencies) which corpus methods afford. Similarly, while Bahloul (2008:1,3) reports using a "corpus", the dataset referred to (23 short texts) is too small for meaningful quantitative analysis. Thus, corpus methods have yet to be applied to *general* topics in Arabic syntax in any significant way.

In this paper, we attempt a first corpus-based empirical description of periphrastic tense-aspect constructions in Arabic. We investigate the verb *kāna*, 'be'. Like verbs meaning 'be' in many other languages, *kāna* functions as both copula and auxiliary. We analyse these functions in the Leeds *Corpus of Contemporary Arabic* (CCA: Al-Sulaiti and Atwell, 2006), one million words of written Modern Standard Arabic (MSA). We devise and evaluate enhancements to an approach to part-of-speech tagging introduced in earlier work (Ibrahim and Hardie, 2019) to distinguish copula and auxiliary examples, and thence apply manual concordance analysis to categorise the observed patterns of use. First, we review relevant background information in the prior literature (section 2). We then explain how we enhanced the existing tagging system to achieve automatic disambiguation of copula and auxiliary functions (section 3). Section 4 presents our exploration of the data, and outlines the usage patterns that we observed. Our general finding is that descriptions of *kāna*'s behaviour in the literature are accurate, but inadequate: the corpus data allows a more complete and nuanced description than hitherto possible.

[2] We take no position on this question, but use *past*/*present tense* for descriptive convenience.

2. Copula and auxiliary verbs in Arabic

The Arabic verb *kāna*, 'be',[3] is one of a group of verbs traditionally labelled '*kāna* and her sisters' or the 'sisters of *kāna*' (literally translating *kāna wa-ʾaḫawātuhā / aḫawātu kāna*; see Ryding, 2005:634) for their shared behaviour as copula verbs. As well as *kāna*, the group includes *ʾaṣbaḥa*, 'become (in the morning)', *ʾaḍḥā*, 'become (before noon)', *ẓalla*, 'continue, remain', *bāta*, 'become', *ʾamsā*, 'become (in the evening)', *ṣāra*, 'become', *laysa*, 'not be', *mā zāla*, 'continue', *mā bariḥa*, 'continue', and *mā dāma*, 'continue'. *Kāna* is the prototypical and most common member of this group;[4] all but a handful are fairly rare, as the frequency list of any lemmatised Arabic corpus immediately shows. To situate our exploration of *kāna* relative to the literature, we survey what is said about it (and particularly auxiliary uses) in English-language reference and pedagogical grammars of MSA, before considering literature in theoretical syntax.

Arabic copulas are largely unremarkable cross-linguistically: they link a subject to a subject complement, conveying such meanings as 'become', 'remain', and 'stay' (Ryding, 2005:176-177, 634-636; Abu-Chacra, 2007:195-196). *Kāna*, unlike other copulas, is not used in the present tense, where the *nominal clause* construction (subject plus subject complement with no verb) occurs instead. Only when tense other than present needs to be conveyed is *kāna* used to express copula 'be' (Ryding, 2005:59, 63).

The use of *kāna* and 'sisters' as copulas is prominently and consistently explained in grammars of Arabic published in English. However, as in many languages, the common and general copula *kāna* also serves as an auxiliary verb, forming *periphrastic tense-aspect constructions* – also called *verbal complexes* (Bahloul, 2008; Cuvalay-Haak, 1997) or kāna-*compounds* (Marmorstein, 2016:123-131) – to express some particular combination of tense and aspect which the language does not mark inflectionally.

Reference grammars often say relatively little about auxiliary *kāna*. The most detailed account is Ryding's (2005:446-449). Ryding describes a number of tense-aspect constructions involving *kāna*, under the rubric of "compound verbs". She characterises past-tense *kāna* followed by a present-tense main verb as the "past progressive", functioning "[t]o convey the idea of continued or habitual action in the past"; she further notes that "experiential verbs" that indicate "knowing, feeling or understanding" often appear in "the past continuous [*sic*] tense rather than the simple past in Arabic", unlike in English. Next, Ryding discusses the "pluperfect or past perfect", i.e. past-tense *kāna* plus past-tense main verb, which expresses "an anterior action, i.e., an action in the past that is over with and which serves as a background action for the present". She adds that "[t]he particle *qad* may be optionally inserted just before the main verb", although "[r]arely is *qad* used when the verb is negative"; *qad* is a multifunctional particle, interpreted as emphasising aspectual or modal meaning (Ryding, 2005:450). Present- or future-tense *kāna* plus past-tense main verb is described by Ryding (2005:449) as "future perfect", conveying "a state or action expected to be completed in the future". Finally, Ryding notes that past-tense *kāna* plus future-tense main verb expresses "unreal condition or a contrary-to-fact condition", that is, "an action that would or could have taken place, but actually did not".[5]

It is worth underlining the brevity of Ryding's accounts of these constructions. The "future perfect" and "unreal condition" are covered in two and three lines respectively, the

[3] Arabic verb lemmata are conventionally labelled with the third-person singular masculine past-tense form. Thus *kāna* is both the 'name' of this lemma, and the form meaning '(he) was'.

[4] In the CCA, *kāna* is the tenth most frequent lemma overall (frequency 8,854) and most common verb lemma. The next most frequent copula, *ʾaṣbaḥa*, 'become', has frequency 808, just one-tenth that of *kāna*.

[5] Sources which encompass Colloquial Arabics report more than four *kāna*-based constructions (e.g. Ouali, 2018 identifies nine). However, Colloquial Arabic is outside our scope here.

pluperfect in a single albeit longer paragraph, with only "past progressive" receiving slightly more solid treatment over two pages. While Ryding's account is valuable beyond her definitions for the multiple examples she provides, we must suspect that this compressed presentation does not convey the whole picture.

Other MSA grammars present the copula and auxiliary functions of *kāna* in ways parallel to, but briefer and less satisfactory than, Ryding. For instance, Abu-Chacra (2007:240-241) lists four "tenses" with auxiliary *kāna* (the same four constructions that Ryding presents) and gives one or two examples of each, but not any further explanation. His list is: "Past perfect (pluperfect)"; "Past progressive or habitual"; "Future in the past (future of perfect)"; "Past in the future (perfect of future)". Abu-Chacra notes the optional *qad* in the pluperfect as being "inserted to emphasize the finality of the action or for reasons of style".

Alhawary's (2011:84-86) account is longer but less complete. He lists the construction of past-tense *kāna* plus present-tense main verb *twice*, under "Past Continuous Tense" and "Past Habitual Tense", corresponding to two English translations, *was doing* versus *used to do*. He also mentions "past perfect tense", consisting of past-tense *kāna* and past-tense main verb, whilst subsequently discussing sentences with a sequence of tenses across clauses; he defines the function of such a sequence as "expressing two events/verbs in the past, one having happened before the other".

Wickens (1980:73) introduces the "pluperfect" (past-tense *kāna* plus past-tense main verb), and adds two points on its use: that the clause subject often appears between the two verbs, and that the particle *qad* "maybe added to this construction (as well as substituting for it): it may precede the first verb or the second". Then he introduces the sequence of present-tense *kāna* and past-tense main verb as equivalent to "the English so-called future perfect, 'he will have written'", and notes that the same points apply to this as to the pluperfect. Wickens does not mention the structures that Ryding labels "past progressive" and "unreal condition".

We summarise the functions of auxiliary *kāna* reported by these sources as follows:

- **Past progressive/continuous/habitual**: past-tense *kāna* with present-tense main verb; translated as *was VERBing/used to VERB.*
- **Past perfect/pluperfect**: past-tense *kāna* with past-tense main verb, with particle *qad* between the verbs typically described as optional; translated as *had VERBed*.
- **Future perfect/past-in-the-future**: present- or future-tense *kāna* with past-tense main verb, again with optional *qad*; translated as *will have VERBed*.
- **Future-in-the-past/Unreal condition/contrary-to-fact**: past-tense *kāna* plus future-tense main verb; translated as *would have VERBed*.

In theoretical linguistics, *kāna* is usually treated rather differently. In formalist/generativist studies (notably Aoun et al. 2010 but see also Bahloul, 2008; Benmamoun and Choueiri, 2013; Ouali, 2018), it tends only to attract attention within debate on the tense-versus-aspect interpretation of the finite verb forms and consequences thereof for generativist theoretical constructs such as the Inflection Phrase, Tense Phrase and Aspect Phrase. As Benmamoun and Choueiri's (2013) overview of generative research into Arabic syntax demonstrates, issues such as negation and the nature of the Arabic 'subject' are of more direct concern. When auxiliary *kāna* does enter the discussion, however, it is explained in terms not of forming periphrastic constructions, but rather as having the consistent function of adding an indication of past time to a main verb that lacks time marking and inherently encodes only aspect. For instance, Bahloul (2008:136) argues that in Arabic

> the complex temporal relations, present, past, or future, do not affect and are not affected by the basic verbal form. Instead, these relations are entirely governed by auxiliaries and modals.

> This amounts to saying that, while auxiliaries and modals control the temporal and modal features of the verbal complex, the verbal form denotes basic invariant features."

That is, the presence or absence of *kāna* adds time reference to a finite verb which expresses only aspect (a view which implies the primacy of aspect in distinguishing the two finite forms).

Comparable views are found outside of purely formalist syntactic theory. Marmorstein (2016:68) judges that "[t]he auxiliary verb *kāna* operates as a temporal or a modal adapter: it adjusts the predicate to the deictic point of reference […] so that the predicate is left to indicate aspectual distinctions". Ammann (2002:328) asserts that "[i]n the complex constructions of copula + lexical verb, what the inflection of the copula marks in Arabic is clearly absolute tense" – as opposed to the main verb, which marks *relative* tense after a copula or absolute tense when alone (thus, unlike Marmorstein, favouring the primacy of tense over aspect). Both these scholars do in addition catalogue the periphrastic constructions with *kāna* in ways more-or-less compatible with the reference grammars' presentation. Holes (2004) likewise considers the function of auxiliary *kāna* to be adding time marking to a main verb that inherently encodes only aspect, noting that past-tense *kāna* "has an anteriorizing effect on any verb to which it is preposed, whatever its aspectual value" (Holes, 2004:233). When the auxiliary is present rather than past (*yakūnu* being the present-tense form of *kāna*), the function is to indicate "unrealized, or nonfactual action" (Holes, 2004:234). This view can even account for past-tense copula *kāna* as an example of this anteriorisation (implicitly interpreting it as the addition of auxiliary *kāna* to a nominal clause, rather than a past-tense copula: Holes, 2004:232). Holes additionally observes that when preceded by *qad*, the *yakūnu*-plus-past-tense combination indicates possibility: "may have done X" (at some future point). Despite favouring the primacy of aspect over tense for finite verbs, Holes (2004:217) notes that the historic aspect distinction is presently evolving to one of tense in contemporary Arabic, and observes:

> Reading contemporary written Arabic, one has the impression that the use of auxiliaries to form "compound tenses" is much more widespread than was true in the writing of the early nineteenth century, and certainly, going back further, when compared to medieval prose.
> (Holes, 2004:234)

A similar stance is taken by the theorist who has dealt most thoroughly with *kāna*-based constructions, Cuvalay-Haak (1997; summarised in Cuvalay, 1994), working within the framework of Dik's Functional Grammar. Cuvalay-Haak's view of *kāna* that it is a "supportive" verb, added when a tense-mood-aspect (TMA) value needs to be expressed, but there is no verb to mark it on – that is, in cases where the main verb already expresses one TMA value, or where there is no verb at all (i.e. in nominal clauses, resulting in copula *kāna*). Moreover, "the distinctions that are closest to the stem in underlying clause structure have priority for being expressed on the lexical verb, thus forcing 'outer' operator value to be expressed on the auxiliary" (Cuvalay-Haak, 1997:201), so that normally the main verb expresses aspect and the auxiliary tense. Thus, despite her use of a more complex theoretical apparatus, Cuvalay-Haak's view concurs with that of Marmorstein, Ammann, and Holes.

The model of auxiliary *kāna* as pure marker of (absolute) tense is appealing in that it simplifies the explanation required, as four (or more) constructions no longer need separate accounts. However, in corpus-based research, verbal constructions are often observed to exhibit distinctive behaviour not reducible to the sum of their components' tense-aspect-modality features. Therefore, we will here treat auxiliary *kāna* as creating multiple distinct tense-aspect constructions, each of which can be separately analysed. Given the evident variation of

terminology, we adopt the following labels: *past imperfect* for "past progressive/continuous/habitual"; *pluperfect* and *future perfect* in preference to "past perfect" and "past-in-the-future"; and *past counterfactual* rather than "unreal condition" or "future in the past".

3. Enhancing part-of-speech tagging for *kāna*

In previous work (Ibrahim and Hardie, 2019) we outline a system for accessible part-of-speech (POS) annotation in Arabic. Our approach utilises the MADA software (Habash and Rambow, 2005; Habash et al., 2009, 2012), itself built upon the Buckwalter morphological analyser (Buckwalter, 2004), but re-codes the output to make it easily usable within software such as CQPweb (Hardie, 2012), facilitating research by linguists without programming background. One distinction that we *added* to the Buckwalter/MADA annotation was between main verbs (VV…) and auxiliary verbs (VX…). This distinguishes copula and auxiliary uses of *kāna* and other *sisters of kāna*.. Cuvalay (1994:272-281), who refers to the sisters as "defective verbs", notes that they do indeed have uses within multi-verb constructions, so that the VV/VX distinction will likely be valuable for them.

Although our initial version of this system introduced the VV/VX distinction, we did not undertake detailed evaluation of its accuracy. Before researching *kāna*, therefore, we revisited the system and considered (a) how accurately it disambiguates copula and auxiliary functions and (b) whether that accuracy rate could be improved. We looked only at *kāna*, excluding other *sisters of kāna*, in keeping with this study's focus.

We anticipated many errors in copula/auxiliary disambiguation, based on the experience of earlier work tagging other languages. English POS tagsets often do not distinguish auxiliary/non-auxiliary *be*, *have*, and *do* simply because automated taggers struggle to make this distinction. For instance, CLAWS (Garside et al., 1987) applies tags beginning with VB, VH and VD to *be/have/do*,[6] regardless of function – except when applying the detailed C8 tagset,[7] which typically necessitates manual intervention. We were surprised to find that VV/VX accuracy for *kāna* was quite high (78.9%), as table 1 shows. This assessment is based on a small sample of 161 instances of verbs with lemma *kāna*; table 1 breaks down the accuracy rates according to a broad characterisation of grammatical context (our actual analysis, below, applies much more detailed contextual classification).

Context	Correct tag	N tags correct	N examples	% correct
Copula + NP/adjective	VV	30	48	62.5%
Copula + PP/adverbial	VV	25	39	64.1%
Copula + Clause	VV	7	8	87.5%
Copula +Modal + Clause	VV	7	7	100.0%
Auxiliary + Verb	VX	58	59	98.3%
Total		*127*	*161*	*78.9%*

Table 1. Initial accuracy of VV/VX disambiguation

The system works by assigning VX to all possible auxiliaries, and then examining surrounding context for evidence of copula status. Originally, VX was changed to VV *if* there was no second verb between the VX and the next clear end-of-clause (as indicated by a conjunction or punctuation tag). Our evaluation shows this rule to be too weak. Frequently, a second verb *does* appear before the next conjunction/punctuation but the two verbs do not actually form an

[6] http://ucrel.lancs.ac.uk/claws7tags.html
[7] http://ucrel.lancs.ac.uk/claws8tags.pdf

auxiliary-main pair. In such cases, VX was not changed to VV when it should have been. We adjusted the system to look at only the next five tokens (or less if a conjunction/punctuation tag is seen). Looking directly after the prospective auxiliary would *not* be sufficient, because of the many instances where a subject NP occurs between auxiliary and main verbs. Looking at five tokens balances the need to look beyond what immediately follows and the need to avoid looking *so* far ahead that unconnected verbs are encountered.

We tested the modified system on a new sample of selected sentences: some exemplifying copula *kāna* plus NP or adjective, some exemplifying copula *kāna* plus PP or adverbial. These are the contexts where accuracy as assessed above was unacceptably low. Table 2 gives VV/VX disambiguation success rates *for these difficult contexts*, comparing the original system's performance on this new sample to that of the revised system.

Context	System	N tags correct	N examples	% correct	+%
Copula + NP/adjective	Original	38	60	63.3	25.0
	Revised	53	60	88.3	
Copula + PP/adverbial	Original	35	56	62.5	19.6
	Revised	46	56	82.1	

Table 2. Revised accuracy of VV/VX disambiguation in problematic cases; +% = percentage point improvement

The modified disambiguation rule improved VV/VX disambiguation in these problematic contexts to levels comparable to the other three contexts in the initial evaluation. While this is gratifying, it does not preclude the possibility that additional improvements could eliminate yet further errors. Some such potential improvements became evident in the course of our analysis utilising the now-current state of the system (for instance, treating relativiser *mā* as a conjunction-like stop point).

We were content to test the original system's accuracy, and devise improvements, through evaluation of a small sample. But to report the system's new *status quo*, a more robust evaluation was necessary. Thus, we assessed a random sample of 1,000 instances of *kāna* (see table 3 for results; this larger sample is broken down by correct tag rather than context). POS tagger evaluations typically report accuracy in excess of 95%. However, such reported rates are for *all words* and *all tags*. This includes many 'easy wins'; the most frequent words in written English, *the–of–and*, each have only one possible analysis (article/preposition/conjunction) and make up 10 or 11% of all tokens. By contrast we assess one task only: functional disambiguation of a single highly-frequent verb. For this task, the overall accuracy rate of 90.5% is highly satisfactory. We have not attempted to assess tagger accuracy over all tokens; since our system re-codes and slightly extends the MADA output, performance on everything *except* VV versus VX should be roughly that reported for MADA.

Correct tag	N in sample	N tagged correctly	% correct
VV	585	505	86.3%
VX	415	400	96.3%
Entire sample	1000	905	90.5%

Table 3. Evaluation of revised VV/VX disambiguation for *kāna*

4. Investigating *kāna*

4.1. Data and method

The corpus on which we base our analyses is the *Corpus of Contemporary Arabic* (CCA), created at the University of Leeds by Latifa Al-Sulaiti (Al-Sulaiti and Atwell, 2006). While other Arabic corpora are available, the CCA is ideal for our research. It incorporates a range of genres despite its small size, and thus represents general written MSA more adequately than larger, but more narrowly sampled, datasets. Critically, it is available for full-text download,[8] rather than only for online search. This allows us to run it through MADA and our own re-coding program described above; it also means that our research is replicable, in that this corpus can be used by other researchers to check, refute, or build on our work.

The CCA does have certain drawbacks. While it covers many genres, it does not do so evenly, as Al-Sulaiti and Atwell report (2006:161). Moreover, although texts were sourced from multiple countries, the preponderance of data originates from the Gulf States (Qatar and the UAE primarily, Saudi Arabia and Kuwait secondarily). This is suboptimal, as MSA *does* vary regionally, albeit far less than Colloquial Arabic. The second author of the present study is a native speaker of Egyptian Arabic; thus, when a pattern was found that seemed unusual or odd, it was not easy to be completely certain whether it was really an abnormal construction, or simply one present in the MSA of the Gulf but not the MSA of Egypt.

First, we created two complete concordances: (a) for any word with lemma *kāna* and a tag beginning in VX; (b) for any word with lemma *kāna* and a tag beginning in VV. These represent the auxiliary and copula data respectively. Some examples had to be reclassified during analysis, due to residual POS tagging errors (since, of course, the enhancements discussed above did not achieve 100% perfect accuracy). The virtue of searching by lemma, rather than word-form, is that all possible inflections of *kāna* were captured: both tenses, all subject-agreement inflections, and all possible combinations with proclitics and enclitics. The CCA contains 8,854 tokens of lemma *kāna* (5,246 as VV and 3,608 as VX), spread across 89 different word-forms (see Appendix). Then, we reduced the concordances to a size amenable to manual analysis, by taking a 10% random sample of each. Every retained concordance line was examined for patterns evident around the instance of *kāna*, looking particularly at the meaning expressed in context and the form and function of elements in the remainder of the clause (*kāna* being almost always clause-initial). For both copula and auxiliary, the full set of examples was classified according to these patterns, which were tabulated and counted.[9] This was initially merely a superficial classification based on grouping similar examples; at a subsequent stage, we considered the patterns identified in terms of their similarities to one another, allowing us to schematise their interrelationships. The following section presents the patterns identified, with examples and frequencies. We give corpus frequencies as absolute values; as the CCA is a one-million word corpus,[10] these roughly equal per-million relative frequencies. Counts of concordance lines are given as absolute numbers and as percentages of the total examples under consideration.

[8] Data was downloaded from http://www.comp.leeds.ac.uk/eric/latifa/research.htm (offline at time of writing due to reorganisation of the University of Leeds's website).

[9] The initial reading of, and low-level categorisation of patterns within, the concordance lines was undertaken by the second author. Subsequently, the refinement and correction of this initial analysis was undertaken by both authors collaboratively. Cases where the appropriate grammatical analysis was not immediately clear were handled via detailed discussion during which multiple reference works on Arabic grammar were consulted as necessary (all cited under *References*); and additional context for the concordance examples was scrutinised to resolve ambiguities of interpretation.

[10] To be precise (to 3 significant figures) there are 937,000 tokens in the corpus.

The term *pattern* in corpus-based analysis of (lexico-)grammar may refer either generically to anything, concrete or abstract, repeatedly observed in a concordance, or to more strictly defined and theoretically coherent concepts, as for instance in Pattern Grammar (Hunston and Francis, 1999). We did not bind our methods fully to the framework of Pattern Grammar. Although our syntactic analysis, like Hunston and Francis', is of surface-level units, we did not reject traditional clause-function categories such as object and subject/object complement; we would argue that our analysis in fact demonstrates the value of these categories for Arabic. However, where drawing functional distinctions would have taken us too far from our central analytic goals (for instance, distinguishing argument and adjunct PPs), we were content to label entities within patterns purely formally. Conversely, our use of NP and PP (noun phrase and preposition phrase) to label nominals should not be taken as implying adherence to formal theories of phrase structure.

4.2. Results

4.2.1. Copula *kāna*

The data for *kāna* as copula (VV) consists of 499 examples (a number of mis-tagged auxiliaries having been moved to the VX data). 493 can be treated in terms of twelve distinct single-clause patterns – all expressing the copula function, and thus consisting (minimally) of some form of *kāna* together with some subject complement. These are, by and large, classifiable according to two formal criteria. The first is the form of the subject complement: adjectival, nominal (i.e. NP), oblique-nominal (i.e. PP), or clausal; all our examples of clausal complements were headed by complementiser *ʾan*, suggesting that if other types of subordinate clause are used as complements after *kāna*, they are rarer than *ʾan*-clauses.

The second criterion is the form and order of the other elements. In all our examples, *kāna* is at the start of the clause, before subject or complement. But the remainder of the clause varies in three ways. First, the subject may be explicit (an NP) or implicit/zero (inferable from context and the subject-agreement inflection of *kāna*). Second, an explicit subject may either precede the subject complement (typical order) or follow it (inverted order). Third, an adverbial (adverb or PP) may be present or not present. We observe only four different value-combinations for these parameters. Other value-combinations which we strongly suspect to be possible (e.g. explicit subject with adverbial PP) can be presumed, from their absence in this data, to be rare in written MSA. Table 4 illustrates the actually-observed value-combinations.

Value-combination #	Explicit subject	Inverted order	Adverbial
1	+	–	–
2	+	+	–
3	–	–	–
4	–	–	+

Table 4. Value-combinations on parameters of formal variation alongside copula *kāna*

#The value-combinations 1 to 4 thus formed the basis, but not the entirety, of the notation that we developed to describe the twelve patterns we observed. In table 5, each pattern's structure is presented alongside a label in this notation. As well as a number 1 to 4, the pattern labels add an abbreviation for the form of the subject complement (A: Adjectival, N: nominal, O: oblique-nominal,[11] Cl: clausal). The complete list of actually-observed patterns is therefore : A1, A2, A3, A4; N1, N3, N4; O1, O2, O3; and Cl1. No instances of the theoretically possible

[11] By *oblique*, we mean a nominal marked with non-core case, which for Arabic means prepositional marking.

N2 (nominal complement + explicit subject + inverted order) or O4 (oblique-nominal complement + implicit/zero subject + adverbial) were observed, perhaps because these might result in ambiguity. In N1, only order distinguishes subject and complement NPs; it is unsurprising that we found no examples with inverted order (the missing N2). Meanwhile, O4 would put two PPs in a row (the subject complement and an adverbial) which might easily be ambiguous. There are no instances of Cl2, Cl3 or Cl4; but given the rarity of clausal subject complements overall (with just three examples of Cl1) these gaps cannot be considered important. Finally, beyond these eleven, we identify a functionally-specific sub-type of O3 (ModO3), explained below. Following the tabulation of these patterns, their structures and frequencies (see table 5) we move on to consider various points of interest emerging from these analyses.

Label	Structure and function	Freq.	%
A1	*/kāna/* — NP_{subj} — Adj *Adjectival subject complement*	73	14.8
A2	*/kāna/* — Adj — NP_{subj} *Adjectival subject complement*	1	0.2
A3	*/kāna/* — $Zero_{subj}$ — Adj *Adjectival subject complement*	111	22.5
A4	*/kāna/* — $Zero_{subj}$ — PP — Adj *Adjectival subject complement with adverbial PP*	2	0.4
N1	*/kāna/* — NP_{subj} — NP *Nominal subject complement*	82	16.6
N3	*/kāna/* — $Zero_{subj}$ — NP *Nominal subject complement*	75	15.2
N4	*/kāna/* — $Zero_{subj}$ — PP — NP *Nominal subject complement with adverbial PP*	6	1.2
O1	*/kāna/* — NP_{subj} — PP *Oblique-nominal subject complement*	32	6.5
O2	*/kāna/* — PP — NP_{subj} *Oblique-nominal subject complement*	37	7.5
O3	*/kāna/* — $Zero_{subj}$ — PP *Oblique-nominal subject complement*	60	12.2
ModO3	*/kāna/* — $Zero_{subj}$ — PP-modal — *ʾan* — Clause *Modal variant of O3*	11	2.2
Cl1	*/kāna/* — NP_{subj} — *ʾan* — $Clause_{comp}$ *Clausal subject complement*	3	0.6
	Total	493	100

Table 5. Patterns observed for copula *kāna*

The patterns in table 5 (and throughout) can be read as follows.

- Each pattern represents a linear sequence of 'slots' joined with wide dashes.
- Arabic words in italics are concrete components.
- /Slash brackets/ represent lemmata. Thus */kāna/* represents any of the corpus's 89 inflectional forms of *kāna*, e.g. *wakānat*, *sayakūnu*, *kunnā*, …
- A main verb is represented by V with tense in subscript, e.g. V_{past}.

- NPs headed by nouns or non-clitic pronouns are represented as NP; if an NP is the subject/object, subscript 'subj/obj' is added.
- Encliticised pronouns are represented as $Clitic_{obj}$; not all clitic pronouns are objects but those in our patterns all are.
- An implicit/zero subject pronoun is represented as $Zero_{subj}$; unlike other elements, $Zero_{subj}$ has no actual position, so for convenience it is inserted after the verb.
- PP, Adj, and Adv represent PPs, adjectives, and adverbs respectively.
- Subordinate clauses are represented by Clause, with subscript 'subj/obj/comp/adv' if the clause has subject/object/subject complement/adverbial role respectively. Where a specific conjunction introduces a Clause, it is given in the pattern as a concrete element; otherwise, the Clause is implied to include a slot for a varying *incipit*.
- Subject/object complement slots of non-specific formal type are represented as $Comp_{subj}$ and $Comp_{obj}$.
- Finally, semantic restrictions on a slot are indicated with single words appended to elements with hyphens: e.g. PP-modal to indicate that the PP in question always has modal meaning in this pattern.

Figure 1 charts the distribution of instances across patterns. It shows that the preferred values for subject behaviour are the first and third: NP_{subj} followed by complement, and $Zero_{subj}$ followed by complement (so the complement actually follows *kāna*). The former is the most used with nominal and clause complements, the latter with adjective and oblique-nominal complements. Patterns involving subject/complement inversion, or adverbial PPs, are rare. Looking deeper, whereas the patterns for adjectives and nouns are similar, things are different for the Oblique-nominals. When such a complement occurs with an NP_{subj}, subject/complement inversion is more common than not (albeit the standard-error bars overlap almost entirely), even though the pattern with $Zero_{subj}$ is most frequent of all. A log-likelihood test shows that the interaction between the two variables of complement type and subject-behaviour as shown in figure 1 is highly significant (log-likelihood=119.4, d.f.=9, $p=1.178 \times 10^{-21}$).

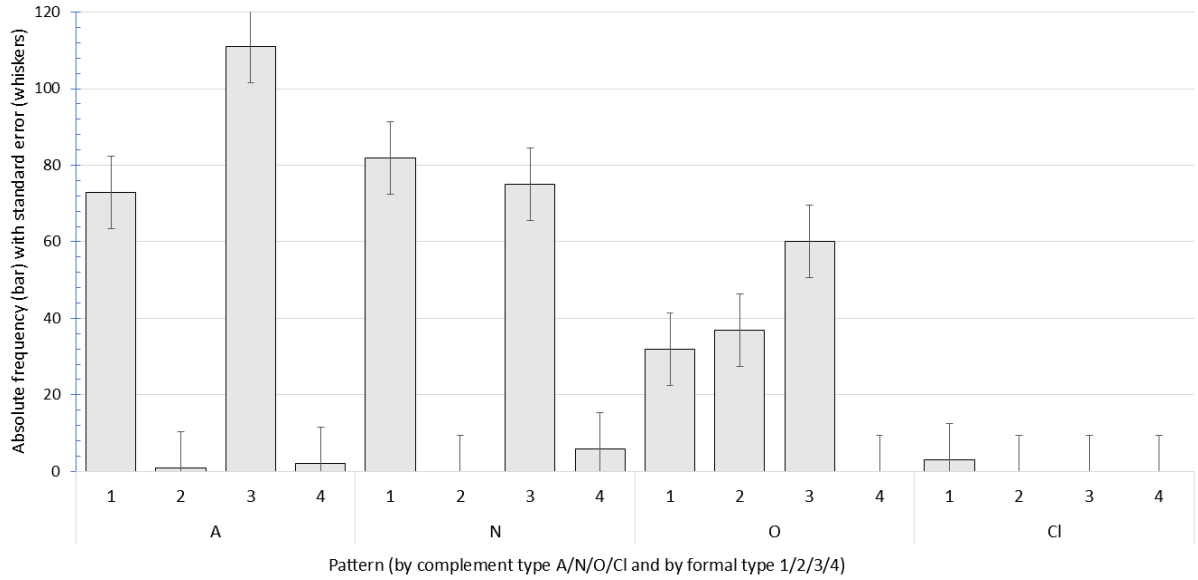

Figure 1. Frequency magnitudes of sixteen copula patterns (including 5 unobserved)

The most common patterns, A3 and N1, are exemplified in (1) and (2) respectively.[12]

(1) له مناسبا كان لكنه

```
lakina=hu kāna         munāsibān      la=hu
but=3SGM  be.PAST.3SGM suitable.INDEF for=3SGM
```

"But it was suitable for him."

(2) التسلية من نوعا التقليد يكون وقد

```
wa=qad  yakūnu       attaqlīd      nawʾān     mina attasliyah
and=QAD be.PRES.3SGM imitation.DEF type.INDEF ABL  entertainment.DEF
```

"And maybe imitation is a kind of entertainment."

Example (2) includes *yakūnu*, the present of *kāna*, normally not used as a copula; its use here is motivated by the construction consisting of *waqad* plus present-tense verb, which expresses epistemic modality (possibility).

The ModO3 pattern deserves specific discussion. Structurally, it is O3: *kāna* has copula function and the PP is a subject complement. The clause with complementiser *ʾan* can be regarded as either the actual subject, or in our preferred view an example of subject extraposition, *kāna*'s implicit subject being co-referential with the clause, and equivalent to English dummy *it*. However, functionally and phraseologically, this pattern can be defined more narrowly due to the specific nature of the PP and the overall meaning conveyed.

Two kinds of PP complement are observed in ModO3. The first combines ablative preposition *mina*, 'from, of', with a definite-marked adjective as nominal head (i.e. it modifies no noun). Critically, this adjective always has modal meaning. The following adjectives occur in our data: *ṣaʿb*, 'difficult'; *mumkin*, 'possible'; *ṭabīʿī*, 'natural, expected', *mutaʾakad*, 'certain'; and *muftaraḍ*, 'supposed, believed'. The function of this adjectival PP is equivalent to a bare modal adjective in English; the structure can be literally translated as 'it was (difficult/possible/certain/…) that (finite clause)', although often an idiomatic English translation differs. Example (3) illustrates *ṭabīʿī*, expressing epistemic modality: high confidence regarding the proposition of the extraposed subordinate clause.

(3) وكان من الطبيعي أن تنتشر في الأشهر الأخيرة صورة " جول "

```
wa=kāna          mina aṭṭabīʿī     ʾan  tantašira              fī
and=be.PAST.3SGM ABL  natural.DEF  COMP spread.PRES.SUBJ.3SGF in
alʾašhur     alʾaḫīrah  ṣūratu  ǧūl
month.PL.DEF latest.DEF picture Gül
```

"And it was natural for Gül's picture to spread in recent months."

Verbs following *ʾan* always have present-subjunctive form; we surmise that the modal meaning arises not from the adjective alone, nor the subjunctive mood alone, but from their interaction, and also from the irrealis or modally-loaded function of *ʾan* itself (Ryding, 2005:611). In Sinclair's (2004) terms, we might call this an *extended unit of (modal) meaning*. Intuition suggests that Colloquial Arabic may, like English but unlike MSA, deploy a bare adjective within this pattern. There do exist in the data two cases of a bare modal term, *lābud*, 'inevitable, necessary', following *kāna*. This word's POS is debatable, but here it seems to have adverbial

[12] Glosses on examples do not include morpheme boundaries except for clitics; instead, marked grammatical categories are indicated per word with full stops, using the following labels: PAST = past; PRES = present; SUBJ = subjunctive; PASS = passive;1,2,3 = first, second, third person; SG,DL,PL = singular, dual, plural; M,F = masculine, feminine; (IN)DEF = (in)definite; DEM = demonstrative; ABL = ablative, COMP = complementiser; REL = relativiser, QAD = mood/aspect particle *qad*. All examples use DIN 31635 transliteration.

or adjectival function (thus, while not a true PP, it seemed best treated within ModO3) and expresses either strong epistemic or strong deontic force.

The other observed form of PP consists of *ʿala*, 'to', followed by a noun or clitic pronoun referring to a sentient being. *ʿala* is used metaphorically in the sense of an obligation being placed *upon* someone; thus, these PPs too express modality (deontic) regarding the subordinate clause's state-of-affairs. The informal English idiom *it was on (someone) to VERB* is comparable.

(4) ثم كان علينا أن نعود ثانية إلى بيشاور

ṯumma kāna **ʿalay=nā** ʾan naʿūda ṯānīyatan ʾilā bīšāwar

then be.PAST.3SGM **on=1PL** COMP return.PRES.SUBJ.1PL again to Peshawar

"Then we were obliged to return to Peshawar."

Beyond the largely systematic patterns tabulated above, a more complex multi-clausal structure is observable in six examples. One of these, given as (5), evidences copula *kāna* used with clausal complement to create a cleft construction.

(5) الطالبات مع يتحاور **ما** دائما كان

kāna dāʾiman **mā** yataḥāwar maʿa aṭṭālibāt

be.PAST.3SGM always **REL** converse.PRES.3SG with student.FPL.DEF

"It was always (the case) that he was talking with the female students."

With only one example little can be said about this use of *kāna*; investigating cleft structures would require targeted corpus searches to locate sufficient examples. In the remaining five instances, exemplified by (6), *kāna* is followed directly by relativiser *mā*, '(that) which', and a relative clause of which *mā* is either subject (four cases) or object (one case). The relative clause is the subject of *kāna*, and is followed by a subject complement. These examples could have been counted along with the explicit-subject patterns (A1, N1, O1) but with a clausal, rather than NP subject; however, the fixity of the bigram *kāna mā* inclines us to treat this as a distinct phraseology.

(6) الدور طعم هو يهمني **ما كان**

kāna mā yuhimu=nī huwa ṭaʿm addawr

be.PAST.3SGM REL matter.PRES.3SGM=1SG 3SGM taste role.DEF

"What mattered to me was the taste of the role."

[Context: an interview with an actor explaining his criteria for choosing parts]

The interpretation of the relative clause in this instance is past imperfect ("that which was mattering to me") despite its present-tense verb. The present-tense subordinate clause is interpreted through the main clause's past tense. This parallels the major auxiliary use of *kāna*, in which past-tense *kāna* and a present-tense verb form the past imperfect construction.

4.2.2. Auxiliary *kāna*

Instances of *kāna* as auxiliary are found in what appears at first sight a bewilderingly immense array of patterns. This is in partly because our method distinguishes the tenses of the auxiliary */kāna/*, a step not applied to the copula, in order to make each pattern specific to just one tense-aspect construction. These tense labels are, however, notional, because many patterns exhibit examples where *kāna* takes a different tense than that characteristic of the construction due to

some preceding element. For example, negativiser *lām* is followed by a present-tense verb, so a past imperfect after *lām* begins with present-tense, not past-tense, *kāna*, and this is still considered an instance of a pattern beginning with */kāna/*$_{past}$.

Within 387 concordance lines,[13] our initial analysis identified 54 distinct combinatory patterns of *kāna* tense, main verb tense, and number/order of clause-level units (mostly NPs and PPs). The frequencies of these 54 patterns have roughly Zipfian distribution: a few highly frequent types account for the overwhelming bulk of instances, while many patterns occur only once or twice. Table 6 lists all patterns of frequency 10+, collectively accounting for 73% of the data (just the top two account for 48%).

Pattern	**Freq.**
/kāna/$_{past}$ — V_{pres} — $Zero_{subj}$ —NP_{obj}	94
/kāna/$_{past}$ — V_{pres} — $Zero_{subj}$ — PP	92
/kāna/$_{past}$ — V_{pres} — $Zero_{subj}$	20
/kāna/$_{past}$ — V_{pres} — $Zero_{subj}$ — PP — NP_{obj}	20
/kāna/$_{past}$ — V_{pres} — $Clitic_{obj}$ — $Zero_{subj}$	20
/kāna/$_{past}$ — V_{pres}-modal — $Zero_{subj}$ — *ʾan* — Clause	20
/kāna/$_{past}$ — V_{pres} — $Clitic_{obj}$ — NP_{subj}	15
Total	***281***

Table 6. Patterns of auxiliary *kāna* with frequency >= 10 in the sample concordance

The seven major patterns involve present-tense main verbs; each represents a highly frequent pattern of verb valency in Arabic. The verb's agreement with the subject for person, number and gender means that a given-information subject is typically implicit, i.e. $Zero_{subj}$; only the seventh pattern has a full subject NP. The three common realisations of verbal argument nominals – object NPs, PPs, and enclitic object pronouns – generate different transitivity patterns: with no nominals, one nominal, or (rarely) two nominals. Thus, the most common use of auxiliary *kāna* is with the most simple and frequent verb complementation patterns, exactly as would be expected. The top two patterns having more-or-less equal frequencies is perhaps slightly remarkable. However, the second pattern's formal definition masks a distinction between PP as verb argument and PP as verb adjunct. Our analysis lacked scope to attempt to disambiguate PP argumenthood; we leave this as an avenue for future research.

The pattern of past imperfect with modal verb plus subordinate clause (sixth most frequent) is functionally idiosyncratic; we postpone it for separate consideration. We also exclude eight examples where the main verb carries passive inflection, each of which occurs within a different configuration of non-subject elements. On the basis of so little data, nothing solid can be said about passives within periphrastic tense-aspect constructions.

This leaves 359 examples spread over 46 patterns. Treating all 46 individually would be an exercise in futility. We sought therefore, as with copula *kāna*, to identify parameters on which the patterns might be categorised and, thus, understood. The most obvious such parameter is main verb tense, and we used this as our criterion of first division. Our literature review (§2) identified *kāna*'s combination with past, present and future main verbs as creating, respectively, the pluperfect/future perfect, past imperfect, and past counterfactual constructions. Table 7 gives the frequencies of each in our dataset. Present-tense main verbs are overwhelmingly most frequent. This is not merely a function of the overall frequency of

[13] Adding examples reclassified from the VV sample gave 393 instances of VX initially, but six others were reclassified as VV (having been mis-tagged due to a nearby relativiser) and have been dealt with above.

the tenses in the corpus, which are comparable[14] (49,947 past-tense verbs per million words and 52,947 present-tense verbs per million words).

Main verb tense	**Tense-aspect construction**	**Freq.**	**%**
Past (with past *kāna*)	Pluperfect	41	11.4%
Past (with present *kāna*)	Future perfect	1	0.3%
Present	Past imperfect	313	87.2%
Future	Past counterfactual	4	1.1%

Table 7. Frequency of main verb tense after auxiliary *kāna*

The tense of *kāna* is usually past. *Kāna* being present tense is motivated by certain items in the immediately preceding context, and seems not to change the tense-aspect meaning of the *kāna* construction. For instance, negative marker *lām* and complementiser *ʾan* cause a following *kāna* to take present tense (with jussive and subjunctive mood respectively). As this seems to be a general feature of these elements, without implications for how *kāna* relates to the rest of the clause, we did not divide the patterns according to *kāna*'s tense. The exception was the sole future perfect example, whose present-tense *kāna* is inherent to the construction.

Beyond tense, we found that the maze of patterns could be simplified by describing each in terms of two features: the number and relative ordering of clause-level elements other than the verb and subject (objects, PPs, complements, etc.); and the presence or absence, and if present the position, of an explicit subject NP. Subject NPs occur in three positions in this data: after the main verb, after *kāna*, and finally (i.e. after the sequence of elements captured by the other parameter). We label these SUBJ-1 to SUBJ-3 respectively, using SUBJ-0 for the implicit subject (see table 8 for definitions and frequencies). In some clause types, an implicit subject is realised as a pronoun enclitic on the *incipit* conjunction (e.g. in (7) below); as with the effect of prior *lām* or *mā*, this is best explained as a feature not of the behaviour of *kāna* itself, but of the conjunction; thus, such subjects are not here distinguished from wholly implicit subjects.

LABEL	**Subject position**	**Freq.**	**%**
SUBJ-0	No subject NP (implicit/zero subject (or clitic on conjunction))	304	84.7%
SUBJ-1	NP after main verb (or after main verb + clitic pronoun)	28	7.8%
SUBJ-2	NP after *kāna*, before main verb (mostly in pluperfects with *qad*; see below)	19	5.3%
SUBJ-3	NP final (i.e. after 1 or more non-subject post-verbal elements)	8	2.2%

Table 8. Presence and positioning of subject NPs

Table 9 lists the different values we found for different sequences of non-verb, non-subject elements (henceforth, *configurations*), and their frequencies. Two types of configuration emerged: simple, where a single non-subject follows the verb; and compound, where two non-subject elements follow the verb, and both their functions and their relative ordering must be specified. We assigned labels A to H to the simple configurations, and generated labels for the compound configurations based on what they combine (GB = G+B, etc.; J labels object

[14] These frequencies were ascertained via POS tag queries, using the following CQP syntax: for past, *[pos="V..P.*"]* ; for present, *[pos="V..I.*"]* .

complements, which do not occur alone). The most obvious finding is the utter dominance of two simple configurations: either a single PP or a single object NP. This surely reflects MSA's underlying verb-valency behaviour: it is most common for a verb to have one non-subject argument, and that argument is most commonly expressed as an object NP or a PP (again, however, we must note that this analysis has not distinguished argument and adjunct PPs). Yet, if we consider the predominance of configurations B and C alongside the even greater predominance of SUBJ-0, we identify a perhaps unexpected tendency: the typical Arabic clause involving a periphrastic tense-aspect construction has *only one explicit nominal* – a zero subject plus an NP or PP. This might have been predicted in a spoken corpus. Du Bois (1987:818) examines a small spoken corpus of Sacapultec Maya and observes that almost no clauses include more than one full NP (and full NPs are likely to be intransitive subjects, objects, or obliques, but *not* transitive subjects). Spoken discourse has been found to operate similarly in other languages (e.g. Nepali: Genetti and Crain, 2003). The relevance of this cross-linguistic research is that we have here evidence of similar behaviour in *written* MSA. This is a point which, we assert, ought to inform teaching of Arabic, as a first or foreign language.

This is the final authors' version of: Hardie, A and Ibrahim, W (in press) **Exploring and classifying the Arabic copula and auxiliary *kāna* via enhanced part-of-speech tagging**. *Corpora*. In case of any difference between this version and the journal's typeset version, the latter is to be considered definitive.

LABEL	**Configuration**	**N patterns**	**Freq.**	**%**
Simple patterns				
A	None	5	29	8.1%
B	— PP	8	118	32.9%
C	— NP_{obj}	6	103	28.7%
D	— *ʾan* — $Clause_{obj}$	1	2	0.6%
E	— $Clitic_{obj}$	5	41	11.4%
F	— $Comp_{subj}$	1	5	1.4%
G	— Adv	1	3	0.8%
H	— $Clause_{adv}$	1	1	0.3%
J	(— $Comp_{obj}$)			
Compound patterns with adverbs				
GB	— Adv — PP	1	1	0.3%
CG	— NP_{obj} — Adv	1	2	0.6%
EG	— $Clitic_{obj}$ — Adv	1	3	0.8%
Compound patterns with 2 non-adverb elements				
BB	— PP — PP	1	3	0.8%
BBv	— PP-shifted — (*verb*) — PP (variant of BB, with first PP shifted before V)	1	2	0.6%
CB	— NP_{obj} — PP	2	8	2.2%
BC	— PP — NP_{obj}	4	24	6.7%
EB	— $Clitic_{obj}$ — PP	3	4	1.1%
EC	— $Clitic_{obj}$ — NP_{obj}	1	1	0.3%
FB	— $Comp_{subj}$ — PP	1	1	0.3%
Compound patterns involving object complements				
CJ	— NP_{obj} — $Comp_{obj}$	1	5	1.4%
EJ	— $Clitic_{obj}$ — $Comp_{obj}$	1	3	0.8%
Total		46	359	

Table 9. Configurations of non-subject clause-level elements

Examples (7-10) illustrate all four subject-position types. Example (10) is pluperfect (*kāna* being present subjunctive, not past, due to preceding *ʾan*), the others past imperfect.

(7) القمر عن شيئا نعرف لا كنا أننا الآن اكتشفت

```
iktašaftu         alān ʾanna=nā kunnā       lā  naʿrif
discover.PAST.1SG now  that=1PL be.PAST.1PL NEG know.PRES.1PL
šayʾān      ʿan   alqamar
thing.INDEF about moon.DEF
```

"I now discovered that we didn't know anything about the moon."

[*/kāna/*$_{past}$ — V_{pres} — $Zero_{subj}$ — NP_{obj}— PP]

(8) الوقت طوال محمد يطلبه كان ما و هو

```
wa=huwa  mā  kāna         yaṭlubu=hu          muḥammad {ṭuwāl alwaqt}
and=3SGM REL be.PAST.3SGM call.PRES.3SGM=3SGM Muhammad {all the time}
```

"And that's what Muhammad was always calling for."[15]

[*/kāna/*$_{past}$ — V_{pres} — $Clitic_{obj}$ — NP_{subj}]

(9) الكريم القرآن آي يرتلان محمود علي والشيخ ندا أحمد الشيخ وكان

```
wa=kāna           aššayḫ     ʾaḥmad nadā wa=aššayḫ      ʿalī maḥmūd
and=be.PAST.3SGM sheikh.DEF Ahmad  Nada and=sheikh.DEF Ali  Mahmoud
yuratilānni       āy       alqurān   alkarīm
recite.PRES.3DLM verse.PL Quran.DEF noble.DEF
```

"And Sheikh Ahmad Nada and Sheikh Ali Mahmoud were reciting verses of the Holy Quran."

[*/kāna/*$_{past}$ — NP_{subj} — V_{pres} — NP_{obj}]

(10) عام آلاف عشرة عليها مر قد يكون أن بعد الأرض إلى العودة ثم

```
ṯumma alʿawdah   ʾilā alʾarḍi   baʿda ʾan  yakūna            qad
then  return.DEF to   earth.DEF after COMP be.PRES.SUBJ.3SGM QAD
marra          ʿalay=hā ʿašrat ālāf     ʿām
pass.PAST.3SGM on=3SGF  ten    thousand year
```

"Then, the return to Earth after ten thousand years had passed there."

[*/kāna/*$_{past}$ — *qad* — V_{past} — PP — NP_{subj}]

As the patterns appended to (7-10) show, these examples *also* illustrate four different non-subject-element configurations: CB, E, C, and B respectively. C and B are the two most common configurations, though their *combinations* with subject-position types in these specific patterns are not necessarily frequent.

Remaining to be explained is configuration BBv (see Table 9), 'variant of BB', where two PPs are separated by the main verb in the pattern [*/kāna/*$_{past}$ — $Zero_{subj}$ — PP — V_{pres} — PP] (2 examples). This sequence diverges from normal Arabic word order. In both examples, the pre-verbal PP is a time-adverbial, rather than an argument, and seems to have been moved forward in the clause for emphasis. The main-verb-then-argument sequence is preserved.

Introducing the past imperfect, Ryding (2005:446) asserts that "if there is a specific subject mentioned, it comes between the two parts of the verb", allowing only for SUBJ-0 and SUBJ-2. However, we found 26 examples of past imperfect with SUBJ-1, and six with SUBJ-3 (likewise, all four subject positions *do* co-occur with the pluperfect). The six SUBJ-3 examples all have a PP before the subject. In five, the clause-final position of the subject is explicable in terms of end-weight, as the subject is very long. In the remaining example, given above as (10), the subject *ʿašrat ālāf ʿām* is not especially long, but is longer than the PP *ʿalayhā*, 'on it'. End-weight is thus the likely explanation here as well.

One other subject position is observed in the data, but not listed in table 8 because its sole occurrence is with one of the set-aside passive examples. In this ordering, the subject occurs *between* two other post-verbal elements.

(11) الثانية ليون جامعة في **الأسبوعية الندوة** فيها تعقد كانت

```
kānat        tuʿqadu             fī=hā   annadwah      allusbūʿiyah
be.PAST.3SGF hold.PASS.PRES.3SGF in=3SGF symposium.DEF weekly.DEF
fī ğāmiʿat    līawn aṯṯāniyah
in university Lyon  second.DEF
```

"The weekly symposium was held there (*lit.* 'in it') at University of Lyon 2."

[15] {…} marks *ṭuwāl alwaqt* as a fixed idiom with non-compositional meaning, glossed with an English idiom.

In example (11), the two-word subject (emboldened) follows a short PP (preposition plus clitic pronoun) but precedes a four-word PP. (The English translation fails to make it clear, but the first PP would *not* be interpreted as coreferential with the second.) Although one example is too little to be certain, end-weight again seems a reasonable explanation.

Arabic reference grammars claim use of *qad* in the pluperfect to be optional. However, our data includes 15 patterns with *qad* (33 examples) and 5 without *qad* (8 examples). All but one of the examples without *qad* involve an implicit subject with one of the most common argument configurations (PP and/or NP_{obj} or none).Clearly, the pluperfect with *qad* (i) is strongly preferred and (ii) permits a greater variety of structures in the remainder of the clause. Our view is that reference and, even more, pedagogical grammars ought to reflect these strong tendencies when introducing the pluperfect.

We pass over the sole example of the future perfect, which proves only its rarity. The past counterfactual, also rare, appears largely in conditional sentences. As Sartori (2019) demonstrates, a central question for conditionals is the sequence of tenses between condition and main clauses, but we have insufficient data to address that issue.

Finally, we return to the modal pattern whose discussion we postponed: [*/kāna/*$_{past}$ — V_{pres}-modal — $Zero_{subj}$ — *ʾan* — Clause] (20 examples, not included in the frequencies in tables 7, 8 and 9). The fact that specifically *modal* verbs consistently occur with implicit subject plus *ʾan*-clause indicates this pattern to be a specific mechanism for verbal expression of modality. It parallels the structure noted in §4.2.1, where copula *kāna* accompanies *adjectival* expression of modality; just as there, this construction's *ʾan*-clause may be interpreted as the actual subject, but we treat it instead as extraposed and co-referential to a main-clause $Zero_{subj}$. The modal verbs observed in this pattern are: *yumkin*, 'be possible' (5 examples); *yurīd*, 'want' (4); *yanbaġī*, 'should, must' (3); *yaḥlum*, 'dream' (3); *yağib*, 'must' (2); *yastaṭīʿ*, 'can' (2); yaqṣid, 'intend' (1) (given as present-tense forms, since the usual past-tense citation form would not occur here). (12) exemplifies this evidently important phraseology.

(12) سرا القصة هذه تظل أن يمكن كان

```
kāna          yumkin              ʾan  taẓala                 haḏihi
be.PAST.3SGM possible.PRES.3SGM COMP remain.PRES.SUBJ.3SGF DEM.FSG
alqiṣatu  sirān
story.DEF secret.INDEF
```

"It was possible that this story would remain a secret."

5. Conclusion

Our goal in this research was to explore the uses of *kāna* in corpus data and present a description of its behaviour, which we hoped would extend or refine that in the literature – wherein especially auxiliary *kāna* is generally given limited treatment. For the copula, we described two major parameters along which its usage varies, and identified the more and less common structures in which it occurs. We also found some interesting but rare uses which clearly bear further investigation (e.g. cleft constructions).

For auxiliary *kāna*, we showed clearly that, of the four periphrastic tense-aspect constructions presented in the literature, one (the past imperfect) is vastly more frequent than the others. The structures around these constructions vary broadly, but we demonstrated that the majority of this variation can be characterised in terms of (a) subject position and (b) configuration of non-subject elements. The most frequent patterns appear to favour clauses with no more than one explicit nominal, a tendency previously observed in other languages for speech but here seen in *written* MSA. Some lower-frequency patterns seemed at first glance to

be 'peculiar' according to most accounts of MSA – for instance, an adverbial PP between the auxiliary and main verb – but for many of these oddities, we could propose explanations in terms of well-established principles such as forwarding for emphasis and, especially, end-weight.

Finally, we identified two common and consistent constructions for expression of modality (one with the copula and one with the auxiliary).

The overall picture with regards to the prior literature is that what is said about *kāna* is typically accurate but not adequate. The descriptions given of the four tense-aspect constructions are correct, but we would challenge *any* account of these structures that excludes the fact that the past imperfect is many times more common than the rest put together. Likewise, brief accounts of these structures can seem to rule out unusual patterns that are in fact observed – such as a time-adverbial between auxiliary and main verbs. It is, similarly, accurate to observe that use of *qad* in the pluperfect is optional – but it is misleading not then to add that it is *normally* present, and that most cases of its absence coincide with an implicit subject.

These findings have implications for teaching of Arabic as an L2 or, indeed, as an L1. Common fixed structures with specific functions, such as the two modal patterns or the *kāna mā…* pattern, ought in our opinion to be taught explicitly, with an eye to frequent configurations. Frequency patterns like the preference for clauses with at most one explicit nominal, or the high frequency of past imperfect versus pluperfect, ought to guide both classroom practice *and* the amount of attention different phenomena receive in pedagogical and reference grammars.

The contribution of this paper may, then, be stated as follows. First, we have demonstrated a means of researching copula/auxiliary verbs in MSA by using POS tagging to distinguish them. Second, we have added depth and breadth to prior descriptions of periphrastic tense-aspect constructions, a topic which remains underserved in the literature. Third, we have illustrated the potential impact of this research by outlining how, in our view, teaching of Arabic should be informed by the kinds of finding that emerge only from corpus study.

This paper represents only a first step in corpus-based analysis of the forms and functions of Arabic periphrastic constructions. Cleft constructions, passive verbs with *kāna*, and the argument/adjunct distinction for PPs were mentioned earlier as issues requiring additional investigation. We also aspire to extend the methodology exemplified here to other copula verbs, to assess to what extent they differ from *kāna*, and how they are used as auxiliaries. Moreover, we wish to investigate genre/register differences in use of periphrastic constructions; work on other languages suggests that such structures' frequency may vary widely across, and be a marker of, different registers (Biber, 1995). Ultimately, we hope to arrive at a 'catalogue' of Arabic syntactic features sufficiently detailed to enable multi-dimensional analysis following the model of Biber (1988). While recent research (Mohamed and Hardie, 2019) shows that multivariate analysis using only POS frequencies can differentiate major text-types, more nuanced results are attainable if features like the tense-aspect constructions described in this paper are incorporated into such analyses.

References

Abu-Chacra, F. 2007. *Arabic: An essential grammar*. London: Routledge.

Alhawary, M.T. 2011. *Modern Standard Arabic Grammar: a learner's guide*. Wiley-Blackwell.

Al-Sulaiti, L. and E. Atwell. 2006. 'The design of a corpus of Contemporary Arabic', *International Journal of Corpus Linguistics* 11(2), pp. 135-171.

Ammann, A. 2002. 'Arabic verbal inflection: an essay in de-exoticing', *Sprachtypologie und Universalienforschung (STUF)* 55(4), pp. 311-339.

Aoun, J.E., E. Benmamoun and L. Choueiri. 2010. *The Syntax of Arabic*. Cambridge University Press.

Benmamoun, E and L. Choueiri. 2013. 'The syntax of Arabic from a generative perspective' in J. Owens (ed.) *The Oxford Handbook of Arabic Linguistics*. Oxford University Press.

Bahloul, M. 2008. *Structure and Function of the Arabic Verb*. London: Routledge.

Biber, D. 1988. *Variation across speech and writing*. Cambridge University Press.

Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Buckwalter, T. 2004. *Buckwalter Arabic Morphological Analyzer Version 2.0*. Linguistic Data Consortium, catalog number LDC2004L02.

Cardona, G. 1976. *Pāṇini: A survey of research*. The Hague: Mouton.

Cuvalay, M. 1994. 'Auxiliary verbs in Arabic' in E. Engberg-Pedersen, L.F. Jakobsen and L.S. Rasmussen (eds.) *Function and expression in functional grammar*. Berlin: Mouton de Gruyter, pp. 265-284.

Cuvalay-Haak, M. 1997. *The Verb in Literary and Colloquial Arabic*. Berlin: Mouton de Gruyter.

Forbes, P.B.R. 1933. 'Greek pioneers in philology and grammar', *The Classical Review* 47(3), pp. 105-112.

Garside, R., G. Leech and G. Sampson. 1987. *The Computational Analysis of English: A Corpus-based Approach*. Harlow: Longman.

Genetti, C. and L.D. Crain. 2003. 'Beyond preferred argument structure: Sentences, pronouns and given referents in Nepali' in J.W. Du Bois, L.E. Kumpf, and W.J. Ashby (eds.) *Preferred argument structure: Grammar as architecture for function*, pp. 197-203. Amsterdam: Benjamins.

Habash, N. and O. Rambow. 2005. 'Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop', *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, pp 573-580.

Habash N., O. Rambow and R. Roth. 2009. '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*.

Habash N., O. Rambow and R. Roth. 2012. *MADA+TOKAN Manual*. Center for Computational Learning Systems technical report #CCLS-12-01. New York: Columbia University.

Hardie, A. 2012. 'CQPweb – combining power, flexibility and usability in a corpus analysis tool', *International Journal of Corpus Linguistics* 17(3), pp. 380-409.

Holes, C. 2004. *Modern Arabic: structures, functions and varieties*. Revised Edition. Washington, D.C.: Georgetown University Press.

Hunston, S. and G. Francis. 1999. *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Ibrahim, W.M.A. and A. Hardie. 2019. 'Accessible corpus annotation for Arabic' in T. McEnery, A. Hardie, and N. Younis (eds.) *Arabic Corpus Linguistics*. Edinburgh University Press, pp. 56-75.

McEnery, T. and A. Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Marmorstein, M. 2016. *Tense and Text in Classical Arabic*. Leiden: Brill.

Mohamed, G. and A. Hardie. 2019. 'Approaching text typology through cluster analysis in Arabic' in T. McEnery, A. Hardie, and N. Younis (eds.) *Arabic Corpus Linguistics*. Edinburgh University Press, pp. 201-228.

Ouali, H. 2018. 'The syntax of tense in Arabic' in E. Benmamoun and R. Bassiouney (eds) *The Routledge Handbook of Arabic Linguistics*. London: Routledge, pp. 89-103.

Owens, J. 1990. *Early Arabic Grammatical Theory: Heterogeneity and standardization*. Amsterdam: Benjamins.

Owens, J. 1997. 'The Arabic grammatical tradition' in R. Hetzron (ed.) *The Semitic Languages*, pp. 46-58. London: Routledge.

Ryding, K. 2005. *A reference grammar of modern standard Arabic*. Cambridge University Press.

Sartori, M. 2019. 'A relational approach to modern literary Arabic conditional clauses' in T. McEnery, A. Hardie and N. Younis (eds.) *Arabic Corpus Linguistics*. Edinburgh University Press, pp. 143-169.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Wickens, G. 1980. *Arabic grammar: A first workbook*. Cambridge University Press.

Du Bois, J.W. 1987. 'The discourse basis of ergativity', *Language* 63(4), pp. 805-55.

Appendix: complete list of forms of *kāna* which occur in the Leeds CCA, with frequencies

The following frequency table was generated using a CQPweb query for all tokens annotated with lemma *kāna*. Using CQPweb's "simple query" syntax, the search pattern is:

{كـا ن}

Because short vowel diacritics are generally omitted in Arabic writing, some forms are ambiguous, for instance *takun*, '(that) she be' and *takunna*, 'you (feminine plural) are'. Ambiguity of this kind is indicated in the frequency table by alternative transcriptions.

More systematic ambiguity exists among inflections distinguished only by short vowel suffixes, for instance indicative *yakūn-u*, 'he is', versus subjunctive *yakūn-a*, '(that) he be'. Unless these vowels are, exceptionally, written explicitly, such pairs will always be spelt the same, and moreover in modern pronunciation short vowel suffixes may be omitted entirely. This being the case, in the table below we do not list every possible reading for forms as systematically ambiguous as *yakūnu/yakūna/yakūn*.

Abbreviations used in the inflectional form labels: 1, 2, 3 = first person, second person, third person; sg, du, pl = singular, dual, plural; m., f., = masculine, feminine; tense/mood labels used are past, present, subjunctive, jussive, imperative, and future; cliticised forms are labelled with #*n*, where *n* references the row where the uncliticised form is explained; the cliticised elements themselves are noted last, and include (a) proclitic conj(unctions). such as *wa-*, 'and'; (b) the proclitic interrog(ative). marker *ʾa-*; and (c) enclitic pronouns, such as *-hā*, 'her'.

Rank	Form	Transliteration	Inflection	Frequency	% of *kāna* tokens
1	كان	kāna	3sg m. past	2711	30.62%
2	كانت	kānat	3sg f. past	1429	16.14%
3	يكون	yakūnu	3sg m. present	782	8.83%
4	وكان	wakāna	#1 + proclitic conj.	776	8.76%
5	تكون	takūnu	2sg m. or 3sg f. present	488	5.51%
6	كنت	kuntu	1sg past	389	4.39%
7	يكن	yakun	3sg m. jussive	372	4.20%
8	وكانت	wakānat	#2 + proclitic conj.	307	3.47%

9	كانوا	kānū	3pl m. past	236	2.67%
10	تكن	takun *or* takunna	3sg f. jussive 2pl f. present	154	1.74%
11	كنا	kunnā	1pl past	132	1.49%
12	فكان	fakāna	#1 + proclitic conj.	104	1.17%
13	وكنت	wakuntu	#6 + proclitic conj.	87	0.98%
14	سيكون	sayakūnu	3sg m. future	84	0.95%
15	ستكون	satakūnu	2sg m. or 3sg f. future	61	0.69%
16	فكانت	fakānat	#2 + proclitic conj.	60	0.68%
17	أكون	ʾakūnu	1sg present	53	0.60%
18	كأن	kaʾana	Tagger error, not a form of *kāna*	47	0.53%
19	ويكون	wayakūnu	#3 + proclitic conj.	43	0.49%
20	ليكون	liyakūna	#3 + proclitic conj.	35	0.40%
21	لتكون	litakūna	#5 + proclitic conj.	34	0.38%
22	يكونوا	yakūnū	3pl m. present	32	0.36%
23	وكأن	wakaʾana	Tagger error, not a form of *kāna*	30	0.34%
24	وكانوا	wakānū	#9 + proclitic conj.	29	0.33%
25	أكن	ʾakun	1sg jussive	28	0.32%
26	وكنا	wakunnā	#11 + proclitic conj.	27	0.30%
27	كانا	kānā	3du m. past	25	0.28%
28	نكون	nakūnu	1pl present	25	0.28%
29	لكان	lakāna	#1 + proclitic conj.	21	0.24%
30	فتكون	fatakūn	#5 + proclitic conj.	17	0.19%
31	وتكون	watakūn	#5 + proclitic conj.	16	0.18%
32	كنتم	kuntum	2pl m. past	15	0.17%
33	وسيكون	wasayakūnu	#14 + proclitic conj.	15	0.17%
34	يكونون	yakūnūn	3pl m. present	13	0.15%
35	كن	kunna *or* kun	3pl f. past 2sg m. imperative	12	0.14%
36	نكن	nakun	1pl jussive	12	0.14%
37	فسيكون	fasayakūnu	#14 + proclitic conj.	8	0.09%
38	فيكون	fayakūn	#3 + proclitic conj.	8	0.09%
39	كانتا	kānatā	3du f. past	8	0.09%
40	وستكون	wasatakūn	#15 + proclitic conj.	8	0.09%
41	سأكون	saʾakūnu	1sg future	7	0.08%
42	لكانت	lakānat	#2 + proclitic conj.	6	0.07%
43	وكانا	wakānā	#27 + proclitic conj.	6	0.07%
44	يكونا	yakūnā	3du m. present	6	0.07%
45	أكان	ʾakāna	#1 + proclitic interrog. marker	5	0.06%
46	أكانت	ʾakānat	#2 + proclitic interrog. marker	5	0.06%
47	فكنت	fakuntu	#6 + proclitic conj.	5	0.06%
48	فكانوا	fakānū	#9 + proclitic conj.	4	0.05%
49	لأكون	liʾakūna	#17 + proclitic conj.	4	0.05%
50	ليكن	liyakun	#7 + proclitic conj.	4	0.05%
51	ليكونوا	liyakūnū	#22+ proclitic conj.	4	0.05%
52	تكونها	takūnahā	#5 + enclitic pron.	3	0.03%
53	تكونوا	takūnū	2pl m. subjunctive	3	0.03%
54	فستكون	fasatakūn	#15 + proclitic conj.	3	0.03%
55	لكنت	lakuntu	#6 + proclitic conj.	3	0.03%
56	وكونوا	wakūnū	2pl m. imperative + proclitic conj.	3	0.03%
57	ولتكن	walitakun *or* walitakunna	#10 + double proclitic conj. 2pl f. present + double proclitic conj.	3	0.03%

58	وليكن	waliyakun	#7 + double proclitic conj.	3	0.03%
59	تكونا	takūnā	2du or 3du f. subjunctive	2	0.02%
60	سيكونان	sayakūnān	2du m. future	2	0.02%
61	سيكونون	sayakūnūn	3pl m. future	2	0.02%
62	فلتكن	falitakun *or* falitakunna	#10 + double proclitic conj.	2	0.02%
63	فلتكونوا	falitakūnū	#53 + double proclitic conj.	2	0.02%
64	لتكونوا	litakūnū	#53 + proclitic conj.	2	0.02%
65	لنكون	linakūna	#28 + proclitic conj.	2	0.02%
66	وأكون	waʾakūnu	#17 + proclitic conj.	2	0.02%
67	وكن	wakunna *or* wakun	#35 + proclitic conj.	2	0.02%
68	وكنتم	wakuntum	#32 + proclitic conj.	2	0.02%
69	وليكون	waliyakūnu	#3 + double proclitic conj.	2	0.02%
70	ونكون	wanakūnu	#28 + proclitic conj.	2	0.02%
71	يكونها	yakūnahā	#3 + enclitic pronoun	2	0.02%
72	أكانوا	ʾakānū	#9 + proclitic interrog. marker	1	0.01%
73	أكونه	ʾakūnahu	#17 + enclitic pronoun	1	0.01%
74	اكان	akāna	Spelling error for #45	1	0.01%
75	ستكونين	satakūnīna	2sg f. future	1	0.01%
76	سنكون	sanakūnu	1pl future	1	0.01%
77	فسنكون	fasanakūnu	#76 + proclitic conj.	1	0.01%
78	فليكن	faliyakun	#7 + double proclitic conj.	1	0.01%
79	فنكون	fanakūnu	#28 + proclitic conj.	1	0.01%
80	كونوا	kūnū	2pl m. imperative	1	0.01%
81	لايكون	lāyakūnu	Spelling error for *lā yakūn*, cf. #3	1	0.01%
82	لتكن	litakun *or* litakunna	#10 + proclitic conj.	1	0.01%
83	لكأنه	lakaʾanahu	Tagger error, not a form of *kāna*	1	0.01%
84	لكانوا	lakānū	#9 + proclitic conj.	1	0.01%
85	لكنك	lakinnaka	Tagger error, not a form of *kāna*	1	0.01%
86	وسنكون	wasanakūnu	#76 + proclitic conj.	1	0.01%
87	وكأنه	wakaʾanahu	Tagger error, not a form of *kāna*	1	0.01%
88	ولتكون	walitakūna	#5 + double proclitic conj.	1	0.01%
89	يكونان	yakūnān	3du m. present	1	0.01%