Liwei Gong* and Satoshi Uehara

Encoding of nominal predication constructions: a typological investigation in verb-initial languages

https://doi.org/10.1515/lingty-2023-0035
Received May 3, 2023; accepted September 15, 2023; published online November 6, 2023

Abstract: Encoding of nominal predication constructions (NPC) is an essential component in typological debates concerning lexical flexibility and parts of speech. This study investigates encoding strategies of NPCs in 65 verb-initial languages from 20 language families. The results indicate that the combinations of a zero strategy and certain other typological features are cross-linguistically disfavored due to non-iconicity. The varying degree of lexical flexibility observed among languages reflects a competition between economy and iconicity, as in many other aspects of linguistic diversity.

Keywords: nominal predication; verb-initial language; Construction Grammar; lexical flexibility; iconicity

1 Introduction

The noun-verb distinction in many languages is reflected by the pattern that nominal1 predication requires extra structural coding (for example, a copula or predicate marker) while verbal (action) predication does not. However, productive zero coding for nominal predication (and adjectival predication) is also observed in a wide range of languages.

1 “Nominal”, “verbal” and “adjectival” (predication) here and in the remaining parts of the article refer to prototypical semantic classes of noun, verb and adjective, respectively: object, action and property (Croft 2001; see also Stassen 1997: 13). They do not presuppose or refer to any word classes in specific languages.

Author contribution statement: Both authors contributed to the design of the research, the theoretical discussion and the data analysis. Liwei Gong is responsible for collecting the data, drafting the manuscript and revising it. Satoshi Uehara helped revise the manuscript and supervised the research.

*Corresponding author: Liwei Gong [liwei.ko@], Graduate School of International Cultural Studies, Tohoku University, Sendai, Japan, E-mail: gong.liwei.t2@dc.tohoku.ac.jp
Satoshi Uehara [sato@uehara], Graduate School of International Cultural Studies, Tohoku University, Sendai, Japan

Open Access. © 2023 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
In Ch’ol and Tagalog, for example, property predication (1b and 2b) and object predication (1c and 2c) do not require a copula morpheme, similar to action predication (1a and 2a). The zero coding for non-prototypical predication constructions here represents a type of lexical flexibility, i.e. “the possibility, in a particular language, to use one or more groups of lexemes in more than one function, without any morphosyntactic adaptations, and without semantic shift” (van Lier 2016: 197). This study addresses an undetermined question regarding typology of lexical flexibility: what types of languages have lexical flexibility (of predication), and what types do not? Or in other words, what typological features correlate with lexical flexibility (of predication)?

It is worth noting that there is a large overlap between verb-initial (V1) languages and languages with flexibility of predication (zero copula), although neither feature entails the other (Clemens and Polinsky 2017; Hengeveld et al. 2004). A majority of languages allowing extremely flexible predication have a V1 word order, such as many Malayo-Polynesian languages, Mayan languages, Salishan languages,
Wakashan languages, and others. However, there is also a considerable number of V1 languages that do require overt copula morpheme(s) for adjectival and/or nominal predicates, such as most Celtic languages.

Despite the prominent status of V1 languages in the topic of lexical flexibility, there have been no typological studies which focus on non-prototypical predication constructions in V1 languages. Cross-linguistic investigations of lexical flexibility are also scarcely conducted from a Construction Grammar perspective (apart from van Lier 2016) which provides a clear distinction between comparative concepts and language-specific categories (Croft 2001, 2022). Aiming to fill these gaps and contribute to general discussion of lexical flexibility and parts of speech in a cross-linguistic context, this study investigates the encoding of nominal predication constructions (NPC) in 65 V1 languages from a Radical Construction Grammar perspective, examining potential correlations between lexical (in)flexibility and other typological features.

We focus on NPCs because predication as an information packaging function is most likely to be flexible for multiple semantic/lexical classes in a language, compared with reference and modification (van Lier 2016). In addition, NPC represents the furthest deviation from the prototypical action predication in terms of semantics and frequency (Croft 1991, 2001; Givón 1984; Stassen 1997), and is thus most likely to differ from action predication in surface structures and require extra structural coding (e.g. copulas, predicate markers). By looking into NPCs, we can observe whether a language uses copula morphemes at all and how flexibility of predication interacts with other typological features.

This article is structured as follows. Section 2 briefly reviews previous hypotheses about correlations between flexibility of predication (usage of copula) and other typological features. Some of the hypotheses are reexamined in this study. Section 3 introduces the main theoretical framework of this study and relevant definitions of comparative concepts. Section 4 introduces the language sampling methodology. Section 5 presents results of this study and discusses their implications. Section 6 summarizes the conclusions of the article and briefly discusses future directions.

2 Previous studies

2.1 Typological features related to the usage of copula morphemes

TAM marking has long been associated with the usage of copula morphemes. One of the most influential theories addressing this correlation is the Dummy Hypothesis, which is well-established in Lyons (1968) and Dik (1989/1997). The hypothesis
assumes that a copula is semantically empty, and its only grammatical function is to carry verbal grammatical categories, especially TAM markers. Thus, a copula is predicted to be used only when TAM markers are morphologically overt, as shown in Table 1.

This is true in languages such as Russian and Modern Standard Arabic, where the copula is only used in sentences with marked TAM. However, Stassen (1997) critically reviews the Dummy Hypothesis with plenty of counter-examples, demonstrating that it is empirically untenable. He shows that while only two combinations of TAM marking and overtness of copulas are predicted to be possible by the Dummy Hypothesis, all four combinations are attested in a wide range of languages (Table 1). Nevertheless, TAM marking, or morphologically bound tense/aspect (T/A) marking in particular, is indeed less frequent in NPCs given the typical high time-stability of nominal predicates (Stassen 1997). This will be further examined in this study and discussed in Section 5.3.

A predicate-initial or predicate-final word order has also been related to the lack of copulas or predicate markers. Hengeveld et al. (2004) propose a functional explanation for this correlation based on Hengeveld’s parts of speech typology and “identifiability of predicate”. Hengeveld and colleagues argue that lexical flexibility may lead to functional ambiguity if language speakers cannot identify which constituent is the predicate and which is (are) the referential phrase(s) in a sentence. There are two ways to avoid such ambiguity in general: (1) a morphological method: to mark the predicate with overt structural coding (copula morpheme or predicate marker); and (2) a syntactic method: to fix the predicate in a uniquely identifiable position (sentence-initial or sentence-final position). If a language allows different lexical classes to function as predicate without overt structural coding, then it abandons the morphological method and can only resort to the syntactic one, resulting in either a predicate-initial or a predicate-final word order.

Although the hypothesis is generally supported by the results of Hengeveld and colleagues’ investigation, there are still several factors that the hypothesis fails to address. Empirically, on the one hand, there are predicate-medial languages in which copula morphemes are rarely used or omissible; and on the other hand, copula morphemes are not too uncommon in predicate-initial and predicate-final languages

<table>
<thead>
<tr>
<th>Copula</th>
<th>Overt TAM</th>
<th>Unmarked TAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>No copula</td>
<td>−</td>
<td>+</td>
</tr>
</tbody>
</table>

**Table 1:** Prediction of the Dummy Hypothesis (adapted from Stassen 1997: 67).
either. It appears that “identifiability of predicate” only exerts a weak influence on the flexibility of predication. And, theoretically, the hypothesis presupposes that all clausal constructions need to have a predicate and are realized with a “subject-predicate” structure. However, pragmatic/discoursal structures may overrule the grammatical structure of “subject-predicate” in topic-prominent languages, pragmatically marked constructions in any languages and most importantly NPCs that are encoded by “non-predicational” strategies developed from pragmatically marked constructions (Stassen 1997).

The hypothesis also does not take into consideration the role of behavioral potential when evaluating “identifiability of predicate”. In addition to structural coding and word order, the behavioral potential of a construction can also serve the purpose of indicating the syntactic role of a constituent. For example, Hawkins (2004: 87) notes that definite articles in many languages can signal a nominalization of some kind, such as the poor/rich referring to ‘the poor/rich (ones)’ in English. The definite marker here hints at the identification of a referential phrase. Similarly, the behavioral potential of predication such as TAM and agreement/indexation markers can imply the function of a constituent as the predicate. And this has been considered one of the main methods for distinguishing the functions of a phrase in languages in which the Noun-Verb distinction is argued to be blurred or nonexistent, such as Tongan (Oceanic, Tonga) (Broschart 1997), Samoan (Oceanic, Samoa) (Mosel and Hovdhaugen 1992) and Strait Salishan (Salish, Pacific Northwest) (Jelinek and Demers 1994), among others. The identification of the functions of ‘woman’ and ‘run’ in (3) and ‘sing’ and ‘noble/chief’ in (4) is largely reliant on the co-occurring behavioral potential markers.

(3) Tongan (Oceanic, Tonga)
   a. na’e lele e kau fefiné
      PST run SPEC PL.HUM woman.DEF
      ‘The women were running.’
   b. na’e fefine kotoa e kau lelé
      PST woman all SPEC PL.HUM run.DEF
      ‘The ones running were all female.’
      (Broschart 1997: 134)

(4) Strait Salishan (Salish, Pacific Northwest)
   a. t’ilem=la=sxʷ
      sing=PST=2SG.NOM
      ‘You sang.’
   b. cə t’ilem=la
      DET sing=PST
      ‘the (one who) sang’
A thorough discussion of “identifiability of predicate” goes beyond the scope of this article, but it is obvious that there are additional factors that can affect the flexibility of predication in languages. We will explore some of those potential factors and the motivation behind their interactions with lexical flexibility in this study.

Apart from TAM and word order, Hengeveld (2007) and Hengeveld and Valstar (2010) identify two typological features that interfere with lexical flexibility. They predict that a flexible lexical class, that is, one that can be used in more than one function without extra structural coding, will not exhibit stem-alternating morphology or internal lexical subclasses. The former refers to stem-alternating morphology that is not predictable based on semantics and phonology but internal to the lexemes, such as the irregular ablaut in Kisi (Bantu, Tanzania) shown in (5). And the latter refers to internal lexical subclasses that are again not predictable based on semantics and phonology, but can trigger morphological changes of the lexemes, such as the declension classes in Polish (Slavic, Poland) shown in Table 2.

Table 2: Polish declensional paradigms (adapted from Teslar 1953: 255, 266, 270, cited in Hengeveld and Valstar 2010: 13).

<table>
<thead>
<tr>
<th></th>
<th>Masculine</th>
<th>Neuter</th>
<th>Feminine</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>Nominative</td>
<td>genitive</td>
<td>dative</td>
</tr>
<tr>
<td></td>
<td>kraż ‘country’</td>
<td>kraj-u</td>
<td>kraż-owi</td>
</tr>
</tbody>
</table>

2 Lexical flexibility discussed in the two studies is defined based on Hengeveld’s (1992) parts of speech theory, which only considers propositional functions and structural coding, but not semantic shifts (cf. Croft 2001: 67–75). What they call “flexibility” covers a wider range of phenomena than discussed in this study. Thus, if a correlation with other typological features is attested in their “lexical flexibility”, we expect it to also hold in ours, but not vice versa.
In this study, we will limit our discussion to internal lexical (morphological) subclasses because the prevalence of stem-alternating morphology in a language is difficult to measure and is not documented in most of our data sources.

Hengeveld and Valstar’s hypothesis is based on the assumption that lexical subclassification is a feature of lexemes that are used in specific syntactic slots; for example, declension classes are a feature of lexemes used as the head of a referring phrase. Thus, the combination in question places a heavy burden on language production, requiring speakers to use different subclassifications for the same lexeme depending on the function in which it is used (Hengeveld and Valstar 2010: 9). However, it is not made clear why declensions (for case, number or definiteness) cannot occur in a function other than reference. For example, a nominal lexeme may retain partial declensional categories even when being used as a predicate. Although the hypothesis is supported by the results of Hengeveld and Valstar’s (2010) investigation of 50 languages in which no counter-examples are reported, it has not been further examined to the best of our knowledge. We hope to further test this hypothesis in our sample languages.

2.2 Typological discussion of word classes and lexical flexibility

There has also been a huge volume of discussion on classification of word classes and/or lexical flexibility in specific languages, which we will not list or review in detail here. We observe inadequacy in at least two aspects. The first is that the distinction between universal comparative concepts and language-specific concepts is often either absent, presumed or vaguely defined (van Lier 2016 being an exception), which makes the results difficult to interpret in a cross-linguistic context. Lack of such distinction can lead to contradictory views on whether one language has a distinction between certain word classes and lexical flexibility at all (cf. Croft 2001, 2020; Croft and van Lier 2012; Haspelmath 2012). Second, a myriad of theoretical approaches and frameworks is employed in previous discussions on lexical flexibility, many of which are not compatible with others at all. This in itself creates formidable obstacles for integration and comparison of previous results, regardless of the linguistic phenomena per se, since the observations are based on disparate...
assumptions of comparative concepts, parts of speech and lexical flexibility. We hope to resolve these two aspects of inadequacy by analyzing lexical flexibility with Radical Construction Grammar, which will be briefly introduced in the next section.

3 Theoretical framework

This study follows Croft’s (2001, 2016, 2022) Radical Construction Grammar (RCG) in terms of the theory of parts of speech and comparative concepts (both construction and strategy). Parts of speech are defined as prototypical combinations of semantic meaning and pragmatic functions, as in Table 3.

The (non-)prototypicality of constructions is represented by three aspects of typological markedness: structural coding, behavioral potential and frequency. Structural coding refers to dedicated morphemes that encode the pragmatic function of a lexeme. In this study, a copula or predicate marker refers to the structural coding of NPCs; that is, a morpheme that signifies the predication function of a nominal (object-denoting) lexeme. Behavioral potential refers to markers which express categories associated to a certain pragmatic function, but do not mark it as such, such as TAM marking in predication constructions.

It is predicted that non-prototypical function-meaning combinations will show at least as much structural coding as prototypical combinations, and conversely, at most as much behavioral potential as prototypical combinations.

Croft (2016: 380) further defines two comparative concepts: construction and strategy.

<table>
<thead>
<tr>
<th>MEANING</th>
<th>Reference</th>
<th>Predication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>Prototypical noun</td>
<td></td>
</tr>
<tr>
<td>Action</td>
<td>Prototypical verb</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Prototypical constructions of parts of speech (adapted from Croft and van Lier 2012: 62).

3 “Pragmatic functions” are also referred to as “information-packaging functions” or “propositional functions” in the literature.
4 Frequency is not discussed in this study due to lack of commensurate corpus data of many sample languages.
CONSTRUCTION: a construction (or any construction) in a language (or any language) used to express a particular combination of semantic structure and information packaging function.

STRATEGY: a construction in a language (or any language), used to express a particular combination of semantic structure and information packaging function, that is further distinguished by certain characteristics of grammatical form that can be defined in a cross-linguistically consistent fashion.

The two concepts pertain to this study in that nominal predication construction is identified as a type of construction; that is, the combination of a pragmatic function (predication) and a semantic class (object/entity-denoting lexeme). Whether a language employs an overt copula morpheme or not represents the strategy of encoding nominal predication in the language. By defining parts of speech and comparative concepts as above, RCG provides a clear distinction between language-specific concepts and universal comparative concepts, and thus a consistent ground for the cross-linguistic discussion of lexical flexibility.

4 Language sampling and data collection

4.1 Language sampling

The sample consists of 65 languages covering all the major language families featuring V1 word order. In order to reflect the overall properties of V1 languages, we controlled the composition of language families so that their proportions resemble those among the world’s V1 languages, as presented in Table 4 (see the Appendix for the full list of sample languages).

Some language families are under-represented in this sample either because we do not have access to descriptions of relevant languages that provide enough information for this investigation, or the relevant languages are described in reference grammars as having a non-V1 word order in contradiction with the classification in the World Atlas of Language Structures (WALS).

We focus on V1 languages in this study for two main reasons. First, there is a large overlap between V1 languages and flexibility of predication (zero copula) as mentioned in Section 1. We aim to incorporate as many languages as possible in our sample that demonstrate a high degree of lexical flexibility. Second, Hengeveld et al. (2004) have addressed such an overlap from the perspective of disambiguation. By restricting our sample languages to V1 languages, we would like to explore how typological features other than word order may influence lexical flexibility.

We do not distinguish between ‘verb-initial languages’ and ‘predicate-initial languages’ in this study. We use both terms to refer to languages in which a
<table>
<thead>
<tr>
<th>Language family</th>
<th>Number</th>
<th>Percentage</th>
<th>Intended number in a 65-language sample</th>
<th>Actual number in this sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austronesianb</td>
<td>53</td>
<td>30.5 %</td>
<td>19.80</td>
<td>20</td>
</tr>
<tr>
<td>Oto-Manguean</td>
<td>19</td>
<td>10.9 %</td>
<td>7.10</td>
<td>7</td>
</tr>
<tr>
<td>Eastern Sudanic (Nilotic/Surmic/Kuliak)c</td>
<td>13</td>
<td>7.5 %</td>
<td>4.86</td>
<td>6</td>
</tr>
<tr>
<td>Afro-Asiatic</td>
<td>11</td>
<td>6.3 %</td>
<td>4.11</td>
<td>5</td>
</tr>
<tr>
<td>Salishan</td>
<td>10</td>
<td>5.7 %</td>
<td>3.74</td>
<td>4</td>
</tr>
<tr>
<td>Mayan</td>
<td>7</td>
<td>4.0 %</td>
<td>2.61</td>
<td>3</td>
</tr>
<tr>
<td>Uto-Aztecian</td>
<td>6</td>
<td>3.4 %</td>
<td>2.24</td>
<td>3</td>
</tr>
<tr>
<td>Arawakan</td>
<td>5</td>
<td>2.9 %</td>
<td>1.87</td>
<td>3</td>
</tr>
<tr>
<td>Indo-European</td>
<td>5</td>
<td>2.9 %</td>
<td>1.87</td>
<td>3</td>
</tr>
<tr>
<td>Penutian</td>
<td>5</td>
<td>2.9 %</td>
<td>1.87</td>
<td>0</td>
</tr>
<tr>
<td>Pama-Nyungan</td>
<td>4</td>
<td>2.3 %</td>
<td>1.49</td>
<td>1</td>
</tr>
<tr>
<td>Wakashan</td>
<td>4</td>
<td>2.3 %</td>
<td>1.49</td>
<td>2</td>
</tr>
<tr>
<td>Austro-Asiatic</td>
<td>2</td>
<td>1.1 %</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>Iroquoian</td>
<td>2</td>
<td>1.1 %</td>
<td>0.75</td>
<td>0</td>
</tr>
<tr>
<td>Mixe-Zoque</td>
<td>2</td>
<td>1.1 %</td>
<td>0.75</td>
<td>0</td>
</tr>
<tr>
<td>Oregon Coast</td>
<td>2</td>
<td>1.1 %</td>
<td>0.75</td>
<td>0</td>
</tr>
<tr>
<td>Tsimshianic</td>
<td>2</td>
<td>1.1 %</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>Hokan (Tequislatecan)</td>
<td>1</td>
<td>0.6 %</td>
<td>0.37</td>
<td>1</td>
</tr>
<tr>
<td>Peba-Yagua</td>
<td>1</td>
<td>0.6 %</td>
<td>0.37</td>
<td>1</td>
</tr>
<tr>
<td>Totonacan</td>
<td>1</td>
<td>0.6 %</td>
<td>0.37</td>
<td>1</td>
</tr>
<tr>
<td>Kuot (isolate)</td>
<td>1</td>
<td>0.6 %</td>
<td>0.37</td>
<td>1</td>
</tr>
<tr>
<td>Movima (isolate)</td>
<td>1</td>
<td>0.6 %</td>
<td>0.37</td>
<td>0</td>
</tr>
<tr>
<td>… (another 17 families with 1 V1 language)</td>
<td>1</td>
<td>0.6 %</td>
<td>0.37</td>
<td>0</td>
</tr>
</tbody>
</table>

Total 174 100 % 65 65

The 174 languages are filtered out in the following way. We first narrow the list down to 194 languages with a dominant “Verb-Subject” order according to “Feature 82A: Order of Subject and Verb” (Dryer 2013). Then, we excluded 20 languages that have a dominant non-V1 word order according to “Feature 81A: Order of Subject, Object and Verb” (Dryer 2013), such as languages with a VSO or OVS word order. A reviewer expressed concerns regarding the potential over-representation of Austronesian languages, which constitute around 30% of the sample languages. In response, we conducted a secondary analysis, excluding the Austronesian languages, to supplement the primary analysis results to be discussed in Section 5. Results of the secondary analysis, which are presented in the Appendix, show that the tendencies to be examined in Section 5 may be somewhat diluted by the exclusion of Austronesian languages, but they remain clearly observable. We use Glottolog (Hammarström et al. 2022) as our major reference for language family classification. The “Eastern Sudanic” family in WALS is not recognized in Glottolog so the six “Eastern Sudanic” languages in our sample are represented as Nilotic, Surmic, or Kuliak language(s). So is the “Hokan” family and the one “Hokan” language in our sample is represented as a Tequislatecan language.
prototypical action-denoting predicate appears at the sentence-initial position and precedes its main arguments in a pragmatically unmarked sentence. It does not exclude languages in which a non-prototypical predicate, such as a nominal predicate, follows its subject.

### 4.2 Data collection

Both classifying nominal predication and identity predication are included as subconstructions of NPCs in this investigation. The former typically classifies the subject into a more general category denoted by the predicate. The latter ascribes information of specific identity to the subject, or in other words, equates the subject with another entity.

We investigate two main aspects of encoding strategies of NPCs in sample languages, structural coding and availability of a part of behavioral potential (tense/aspect [T/A] marking).

Encoding strategies are classified into three main types according to their structural coding: zero strategy, copula strategy and mixed strategy. The zero strategy refers to nominal predication that is not encoded by overt structural coding, regardless of the availability of behavioral potential. In contrast, the copula strategy refers to nominal predication that requires some overt structural coding such as a copula or predicate marker. We will refer to the overt structural coding of NPCs as a copula (morpheme), regardless of its morphological status as an inflectional word, a non-inflectional particle, a clitic or an affix. Besides the two simple strategies, languages can also employ a mixed strategy; that is, a zero strategy and a copula strategy split by grammatical contexts (e.g., present vs. past tense, classifying nominal predication vs. identity predication, etc.) or alternate according to discourse needs.

Then we look into the morphological status of T/A markers in our sample languages and their availability in NPCs, in order to explore the interaction between T/A marking and encoding strategies of NPCs. Specifically, we test the possibility of bound T/A markers being attached to nominal predicates, which is predicted to be low in previous studies.

We expand the scope of discussion from tense in Stassen’s theory to both tense and aspect, but do not include mood or modality for several reasons. The hypothesis that bound tense markers will not occur on nominal predicates is grounded in time stability and iconicity (Stassen 1997; Givón 1984): adjectival and nominal predicates are generally more time-stable than verbal predicates, so tense is irrelevant to the
first two types of predicates and should not be expressed by morphologically bound markers on them. Given this, grammatical aspect which expresses “different ways of viewing the internal temporal constituency of a situation” (Bybee 1985: 21) should also be considered non-essential for nominal predicates, which are highly time-stable or even “time-less” or “a-temporal” in the case of identity predication (Stassen 1997: 109). By contrast, mood and modality generally carry semantic meaning that is not directly related to time stability. Their definition and scope are also far more undetermined than tense/aspect in a cross-linguistic context. Thus, we do not include mood and modality in this investigation. We do not distinguish between tense and aspect because the two categories are often interrelated by nature and are represented by mixed or fused morphemes in many of our sample languages. It is difficult to separate the two categories with cross-linguistically valid and consistent criteria.

In terms of morphological status, we classify T/A markers in our sample languages into separable and inseparable, according to their morphological separability from the head of a predicate: if a T/A morpheme is inseparable from the head of a predicate and/or occurs in a fixed order contiguous to the head of a predicate, it is considered inseparable (adapted from Bybee 1985: 27). Inseparable morphemes include stem-alternation and affixes on the head of a predicate. Separable morphemes include clitics, particles, words and affixes that are hosted by elements other than the head of a predicate (e.g., an auxiliary). We also refer to other corroborating evidence, especially for the distinction between clitics and affixes, such as whether a morpheme triggers stem alternation, whether a morpheme participates in morphophonological processes and whether a morpheme has allomorphs (Haspelmath and Sims 2010: 155). The data used for classification are available in the Supplementary Materials.

Furthermore, we investigate whether the sample languages have intrinsic noun classes that cannot be predicted based on semantics and phonology, in order to examine Hengeveld and Valstar’s (2010) hypothesis that such a feature is not compatible with lexical flexibility due to the consequential heavy burden on language production.

---

5 Another reason for integrating aspect into the discussion is that previous typological investigations on the relevant hierarchies of verbal categories (e.g., Bybee 1985; Malchukov 2004; Noonan 1985) have demonstrated that tense is cross-linguistically external to aspect and is thus “more readily lost/acquired in the process of category changing” (corresponding to non-prototypical constructions in our Construction Grammar framework) (Malchukov 2004: 26). If we hypothesize that tense will not be acquired in certain nominal predication constructions, then we should also be able to predict that aspect will not be acquired in the same constructions.
5 Discussion

5.1 General strategies

General results of encoding strategies of NPCs are presented in Table 5.

In line with previous observations, the results show that V1 languages exhibit a noticeable preference for zero encoding of NPC. As many as 47.7% of the sample languages only employ the zero strategy for NPCs and 76.9% of the sample languages (“zero” + “mixed”) allow zero coding of NPC. In Stassen’s (1997) investigation of encoding of intransitive predication among 410 languages, 33.9% only employ zero strategy for NPCs and 60.2% allow zero coding to some degree. Stassen did not control or discuss word order in his investigation. Nevertheless, given the large size (410 languages) and broad coverage of his sample, we consider his results as a general quantified impression of the encoding strategies of NPCs across the world’s languages.

The following examples will exemplify each type of the encoding strategies. Ch’ol and Tagalog represent sample languages that use zero strategy, as shown in (1) and (2) and reproduced here as (6) and (7). NPCs in the two languages do not require any overt structural coding, being parallel to an action (and property) predication construction.

(6) Ch’ol (Mayan, Mexico)
   a. tyi majl-i jiňi wiňik
      PFV go-ITV DET man
      ‘The man went.’
   b. chaň jiňi wiňik
      tall DET man
      ‘The man is tall.’
   c. maystraj jiňi wiňik
      teacher DET man
      ‘The man is a teacher.’
      (Coon 2014: 79)

Table 5: Encoding strategies of NPCs in the sample languages.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>31</td>
<td>47.7 %</td>
</tr>
<tr>
<td>Mixed (split/alternating)</td>
<td>19</td>
<td>29.2 %</td>
</tr>
<tr>
<td>Copula</td>
<td>15</td>
<td>23.1 %</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td>100.0 %</td>
</tr>
</tbody>
</table>
(7) Tagalog (Central Philippine, Philippines)
a. nag-aaral ako
   IMPF.NOM-study 1SUBJ
   ‘I’m studying.’
b. maganda ako
   beautiful 1SUBJ
   ‘I’m beautiful.’
c. doktor ako
docto
   doctor 1SUBJ
   ‘I’m a doctor.’

In contrast, the Ik (Kuliak, Uganda) NPC presented in (8) requires an overt copula morpheme mit and represents the copula strategy.

(8) Ik (Kuliak, Uganda)
mit-íá nka bábó
be-1SG 1SG:NOM father.your.OBL
‘I am your father.’
   (adapted from Schrock 2017: 574)

Besides the two simple strategies, sample languages may employ mixed strategies. Encoding strategy of NPCs in Modern Standard Arabic (Semitic, Middle East) splits by tense/mood. No copula morpheme is used in the present indicative, while a copula is used in NPCs with overtly marked TAM.

(9) Modern Standard Arabic (Semitic, Middle East)
a. zawjat-ii tabiibatu-un
   wife-my doctor-NOM
   ‘My wife [is] a doctor.’
b. kaan-a jaasuus-an
   be-PST.3SG.M spy-ACC
   ‘He was a spy.’
   (Ryding 2005: 60, 635)

Makah (Wakashan, Pacific Northwest) also employs a split strategy, which is sensitive to the distinction between classifying nominal predication and identity predication. No extra structural coding is used in classifying nominal predication (10a), while a copula morpheme derived from the deictic pronoun ʔux is required in identity predication (10b).
Makah (Wakashan, Pacific Northwest)

a. \textit{wikwi:ya:kʷ=i}
   \textit{boy=INDIC.3SG}
   ‘He is a boy.’

b. \textit{ʔux-\textit{u}=i}
   \textit{so.and.so-APPEN\textit{=INDIC.3SG}}
   \textit{Bill know.-CAUS-PFV-AGENT\textit{=ART}}
   ‘Bill is the teacher.’

(Davidson 2002: 132)

A split strategy is distinguished from another type of mixed strategy, the alternating strategy. For example, a zero strategy (11a) and copula strategy (11b) are both available in Yagua (Peba-Yagua, Peru). The choice between the two strategies is conditioned not by grammatical contexts but by discoursal needs and intentions. Payne (1985: 58) notes that the copula strategy is chosen over the zero strategy if “the speaker wishes to indicate tense or stipulate certain aspectual conditions”, but the copula strategy can also be used when T/A morphology is not overt.

Yagua (Peba-Yagua, Peru)

a. \textit{Machíturu-numaa-(níí) Antonio}
   \textit{teacher-now-3SG Antonio}
   ‘Antonio is now a teacher.’

b. \textit{Riy-curáca sa-vicha-núúy-jnu}
   \textit{3PL-chief 3SG-COP-IMPF-PST3}
   ‘He was their chief.’

(Payne 1985: 58)

The general preference for a zero strategy corroborates the previously observed tendency that V1 or predicate-initial languages tend to be flexible in predication, using no copula morphemes (Clemens and Polinsky 2017; Hengeveld et al. 2004). Nevertheless, 52.31 % of the sample languages (mixed + copula strategy) still use a copula morpheme to some degree and 23.1 % only employ the copula strategy, disallowing zero marked NPCs. This pattern may result from various factors beyond word order, given the various possible forms and origins of a copula morpheme. In the following sections, we will further explore the relations between lexical flexibility and specific typological features, and the possible motivations behind those connections.

5.2 Flexibility and tense/ aspect marking

Theories represented by the Dummy Hypothesis suggest that the usage of a copula is motivated by overt TAM morphology. This is severely criticized by Stassen (1997)
with plenty of counter-examples, and he demonstrates that an overt copula in fact does not correlate with overt TAM morphology. Nevertheless, Stassen recognizes the connection between TAM marking and encoding of predication in an alternative way and proposes the Tensedness Universal to capture typological patterns of encoding strategy of adjectival (property-denoting) predication.

(12) The Tensedness Universal of adjectival encoding (Stassen 1997: 357)

a. If a language is TENSED, it will have NOUNY adjectives.
   If a language has NOUNY adjectives, it will be TENSED.

b. If a language is NON-TENSED, it will have VERBY adjectives.
   If a language has VERBY adjectives, it will be NON-TENSED.

Stassen attempts to explain the motivations behind this universal based on time stability (Givon 1984), semantic relevance (Bybee 1985) and iconicity (Haiman 1980, 1983). Specifically, morphologically bound tense markers are not likely to occur on adjectival predicates because the latter are generally more time-stable than prototypical action-denoting predicates, and it is therefore non-iconic or even anti-iconic to have the semantically irrelevant and morphologically bound tense markers expressed on adjectival predicates. Consequently, if a language has obligatory and bound tense morphology, adjectival predicates will not be encoded by a verbal strategy which otherwise results in tense morphology being bound to adjectival predicates.

Stassen's explanation of the Tensedness Universal has several implications for T/A marking in NPCs. First, nominal predicates are generally even more time-stable than adjectival predicates and thus should resist both tense and aspect marking expressed by bound morphemes (See Section 4.2 for more details regarding the inclusion of aspect). Second, the interactions between T/A marking and encoding of predication only pertain to the non-iconicity and thus improbability of bound T/A markers being attached to a time-stable entity-denoting predicate. Thus, if T/A markers in a language are realized as free morphemes, or as bound morphemes appearing on elements other than the predicate, then they should not be impossible even for non-prototypical predications. If T/A markers are realized as bound morphemes on the predicate, employing a copula to carry them is simply one way out of others to avoid the non-iconic combination. There is at least one more option available for this purpose, which is to neutralize T/A morphology in NPCs, as observed in plenty of languages which exhibit overt bound T/A marking on action predicates but have them neutralized in zero-marked NPCs (Stassen 1997: 68–70).

A more accurate interpretation of the relationship between T/A markers and encoding of NPCs would thus be the following:
Hypothesis on the relationship between T/A markers and NPC encoding:

a. Morphologically separable T/A markers may or may not occur with nominal predicates.

b. Morphologically inseparable T/A markers are unlikely to occur with nominal predicates due to non-iconicity. There are two options to avoid this non-iconic combination:
   1) to introduce a copula morpheme to carry the inseparable T/A markers;
   2) or to neutralize the T/A morphology in zero-marked NPCs.

The results of this study support the hypothesis above in two aspects. First, in languages allowing zero NPCs to some degree (“zero” and “mixed” type), morphologically inseparable T/A markers are indeed unlikely to occur in zero NPCs, while separable T/A markers may or may not be available. The morphological status of T/A markers and their availability in zero NPCs are presented in Table 6.

One data point here represents the group of separable/inseparable T/A markers in one language. If a language has both morphologically inseparable and separable T/A markers, then it constitutes two data points in the table. If a language allows at least some inseparable/separable T/A markers in NPCs, we consider the T/A markers to be ‘available’. Only when no T/A markers are allowed or attested in NPCs do we classify T/A markers as ‘unavailable’. A Pearson's chi-square test yields $\chi^2 = 7.9193$, $p = 0.004891$. We find a significantly low value on the cell representing inseparable T/A markers being available in zero-marked NPCs, which corroborates the predicted low probability of this combination.

Second, the availability of T/A markers in languages employing mixed strategies also supports the hypothesis presented in (13).

NPCs encoded with copulas are more likely to allow T/A marking than zero-marked NPCs in general. Such a preference is especially conspicuous for inseparable T/A markers, as shown by the bolded values in Table 7. Within the sample languages using mixed strategies, a copula NPC never exhibits less T/A

<table>
<thead>
<tr>
<th>Availability in ø NPC</th>
<th>Morphological status of T/A markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Separable</td>
</tr>
<tr>
<td>Available</td>
<td>20</td>
</tr>
<tr>
<td>Unavailable</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
</tr>
</tbody>
</table>
availability than a zero NPC in the same language. The former either allows richer T/A marking than the latter, or the same range of T/A marking as the latter.

For example, NPCs in Yagua can be encoded by either a zero strategy or a copula strategy. Payne (1985: 57) notes that both morphologically bound and free T/A markers in Yagua can only occur in copula NPCs, but not in zero NPCs which generally express a current state of affairs, as in (11).

It is worth noting that examples against the general tendency are found in seven sample languages as shown in Table 6. For example, nominal predicates encoded by a zero strategy allow some bound T/A marking in Nandi (Nilotic, Kenya), Baure (Arawakan, Bolivia) and Chamorro (Oceanic, Guam).

(14) Nandi (Nilotic, Kenya)
   a. ná:nti:intèt kípe:t  
      Nandi Kibet  
      ‘Kibet is a Nandi.’
   b. ki:-ná:nti:intèt kípe:t  
      PST-Nandi Kibet  
      ‘Kibet was a Nandi’
   (adapted from Creider and Creider 1989: 121, 122)

(15) Baure (Arawakan, Bolivia)
   a. ver howe-wape=ri  
      PERF dolphin-COS=3SG.F  
      ‘She changed into a dolphin.’
   b. te ni=šir moestar-a-pa=ro  
      DEM.1M 1SG=son teacher-LK-GO=3SG.M  
      ‘My son is going to be a teacher.’
   (adapted from Danielsen 2007: 195, 196)
(16) Chamorro (Oceanic, Guam)

a. ma’estrun Juan
teacher.LK Juan
‘He is Juan’s teacher.’

b. ma’estrun Juajuan ha’
teacher.LK Juan.PROG EMP
‘He is still Juan’s teacher.’
(adapted from Chung 2020: 13)

The Nandi nominal predicate in (14) can bear a past tense prefix ki:- and the Baure predicates in (15) allow aspects suffixes such as the change of state marker -wape ‘cos’ and the future/intentional marker -pa ‘go’. The Chamorro example in (16) is more surprising in that the nominal predicate can undergo partial reduplication to express the progressive aspect in the same way as action-denoting predicates in the language (Chung 2020: 11–13). Additional instances of zero-marked nominal predicates co-occurring with inseparable T/A markers have been observed in some other languages, including Salishan languages and Wakashan languages in our sample, and Oceanic languages such as Mwotlap (Oceanic, Vanuatu) (François 2005), which is not part of our sample.

Nevertheless, we believe that the hypothesis presented in (13) is not undermined by the possible counter-examples. The combination of zero-marked nominal predicates and inseparable T/A markers is highly limited both across and within languages. Cross-linguistically, its probability is significantly low as indicated by the results in Tables 6 and 7. Within languages that allow this combination at all, it is also limited both paradigmatically and quantitatively: inseparable T/A morphology is infrequent in nominal predicates and possible T/A variations are few. For example, François (2005: 131) observes that nominal predicates with T/A marking in Mwotlap are statistically limited and most NPCs are constructed via juxtaposition without T/A marking.

The patterns observed in this study concerning the morphological status of T/A markers and their occurrence in NPCs also account for the empirical failure of the Dummy Hypothesis: it does not consider the morphological status of T/A markers and overestimates the correlation between overt T/A marking and an overt copula. On the one hand, morphologically separable T/A markers are more acceptable than inseparable T/A markers in NPCs and may not affect the encoding strategies. On the other hand, languages with inseparable T/A morphology can also allow zero coding of NPCs while neutralizing the T/A marking in NPCs. Finally, there are statistically minor exceptions where bound T/A marking is acceptable directly on nominal predicates. And usages of copula morphemes may be motivated by factors other than
T/A morphology, for example the requirement for more efficient processing, which will be discussed in the following sections.

In summary, T/A marking does interact with the encoding strategy of NPCs but in a much less decisive way than predicted by the Dummy Hypothesis. It is only morphologically inseparable T/A markers that are unlikely to occur on nominal predicates cross-linguistically. Languages with inseparable T/A markers may either employ a copula morpheme to carry them as predicted by the Dummy Hypothesis, or alternatively, they may neutralize T/A marking in NPCs, in which case a copula morpheme may or may not be used.

5.3 Flexibility and internal morphological subclasses

As mentioned in Section 3, Hengeveld and Valstar (2010) predict that lexical flexibility will not co-occur with internal morphological subclasses of a group of lexemes.

(17) General hypothesis (Hengeveld and Valstar 2010: 9):
The higher the degree of morphological unity (i.e. the absence of intrinsic subclasses triggering specific morphological processes) of a lexical class is, the higher its degree of applicability in various syntactic slots is. Intrinsic lexical subclasses are therefore not expected to occur in flexible languages.

The results of this investigation partially confirm this prediction, as presented in Table 8.

A Pearson’s chi-square test yields \( \chi^2 = 7.6425, p = 0.0219 \), indicating a significant correlation between the two grammatical features in the sample languages. Specifically, the standardized residuals show significant correlations between ‘morphological classes’ and ‘copula strategy’ and between ‘no morphological classes’ and ‘zero strategy’. No correlation is observed between languages using a mixed strategy and the existence of morphological subclasses within them.

<table>
<thead>
<tr>
<th>Morphological subclasses</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes:</td>
</tr>
<tr>
<td>Only zero</td>
<td>6</td>
</tr>
<tr>
<td>Mixed (split/alternate)</td>
<td>6</td>
</tr>
<tr>
<td>Only copula</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
</tr>
</tbody>
</table>
Despite this tendency, we observe several counter-examples to the predication. For example, Kuot (Isolate, Papua New Guinea) employs no copula morphemes for NPCs, as in (18).

(18) Kuot (Isolate, Papua New Guinea)

\textit{kuraibun u-sik makabun}

spirit.woman 3F-DEM woman

‘that woman (was) a spirit woman.’

(Lindström 2002: 12)

However, the language has a gender system in which the gender assignment is largely unpredictable according to semantics or phonology (Lindström 2002: 176–177). Gender of a common noun is reflected in agreement, index and cross-reference morphology, but is not overtly marked on the noun itself.

The combination of internal morphological classes and zero encoding of NPC is also observed in two Mayan languages, Mam (Mayan, Guatemala and Mexico) and K’iche’ (Mayan, Guatemala). Both these languages employ a zero strategy for NPC encoding and the internal morphological classes are reflected in possessive morphology. Common nouns undergo differentiated morphological processes when being possessed, and the variation of inflections is largely unpredictable based on phonology or semantics. For example, the possessive inflectional classes of Mam are presented in (19).

(19) Mam possessive inflectional classes

a. Invariable:

\textit{k’ooj ‘mask’, n-k’ooj=a (1SG-mask=1SG) ‘my mask’}^{7}

b. Vowel-changing:

\textit{xaq ‘rock’, n-xaaq=a ‘my rock’; tz’alom ‘plank’, n-tz’aalma=a ‘my plank’}

c. Suffix-adding:

\textit{chiky ‘blood’, n-chiky-eel=a ‘my blood’; b’aaq ‘bone’, n-b’aaq-al=a ‘my bone’}

d. Suffix-dropping:

\textit{qam-b’aj ‘foot’, n-qan=a ‘my foot’; aam-j ‘skirt’, n-aam=a ‘my skirt’}


---

6 The Kuot case may arise from a historical reason. Lindström (2002: 193–194) notes that the opaque gender system in Kuot might once have been semantically transparent, but the semantic principles for assignment of gender gradually declined due to semantic drifts. If we accept this hypothesis, then it is possible that Kuot has lost the semantic transparency of its gender system, but has not yet developed a copula morpheme or predicate marker.

7 The prefix \textit{n-} and the enclitic \textit{=a} together mark the first person singular possessor, which applies to all other examples in (19).
The other counter-examples include Baure (Arawakan, Bolivia), Nicrobarese Car (Austroasiatic, India) and Nandi (Nilotic, Kenya).

As discussed in Section 2, Hengeveld and Valstar’s (2010) hypothesis is based on the assumption that lexical subclassification is a feature of lexemes that are used in specific syntactic slots. For example, declension classes are a feature of lexemes used as the head of a referring phrase. Thus, the combination of internal subclasses and lexical flexibility places a heavy burden on language production, requiring speakers to use different subclassifications for the same lexeme depending on the function in which it is used (Hengeveld and Valstar 2010: 9). However, morphological subclasses may also function in non-prototypical syntactic slots or constructions. For example, a nominal lexeme may retain a part of its declensions (e.g., case, number, gender, etc.) even when being used as a predicate.

Another point not addressed by Hengeveld and Valstar is whether internal lexical subclasses can co-occur with split or alternating encoding strategies. The results of this study do not show significant correlations between mixed encoding of NPCs and the (non)existence of internal morphological classes. It appears that internal noun classes are acceptable as long as a copula strategy is available.

We would like to propose an alternative explanation of the interaction between internal morphological classes and lexical flexibility based on typological markedness and prototype effect.\(^8\)

As introduced in Section 1, lexical (in)flexibility is defined as a choice between overt versus zero forms in this study. Another aspect of this choice is whether to encode an NPC with the same structural coding as a prototypical predication construction, or to encode the NPC with an overt and unique structural coding. Stassen (1997: 112) notes that there is a competition between iconicity and economy in effect here. The same zero structure may be used to encode NPCs and other predication constructions in a language. This is in favor of both syntagmatic and paradigmatic economy since no extra structural coding and a minimum number of patterns in total are used for the predication function. Alternatively, a unique structure can be used to encode NPCs, which is in favor of iconicity given the semantic/functional differences between NPCs (especially identity statements) and prototypical action predications. Based on this study’s results and the observations made in Hengeveld (2007) and Hengeveld and Valstar (2010), we propose that internal lexical subclasses and stem-alternating morphology signify a salient distinction between the two prototypes – noun and verb – in a language, and thus prompt an overt and unique encoding of NPCs.

\(^8\) This explanation should also apply to interaction between lexical flexibility and stem-alternating morphology as proposed in Hengeveld (2007), though it is not examined in this study.
As introduced in Section 3, our theoretical framework of RCG defines universal parts of speech as prototypical meaning-function combinations (Table 9). The prototype effect is manifested by typological markedness patterns of these constructions: core members of a category tend to have less structural coding, greater behavioral potential and higher frequency than peripheral members.

Lexical flexibility pertains to the structural coding of the non-prototypical constructions, while internal lexical subclasses (declensional classes) and stem-alternating morphology are a part of the behavioral potential of the category of ‘noun’ in relevant languages. What the two types of behavioral potential in question have in common is unpredictability or irregularity: the classification of internal lexical subclasses and morphological changes involved in stem-alternation are not predictable according to semantic or phonological rules but are internal to the relevant lexemes. Such irregularity represents greater behavioral potential and more salient properties of the category, say, compared with regular morphological classes assigned based on semantic or phonological rules, or agglutinative and periphrastic morphological changes.

- greater allomorphy or morphological irregularity of any type, not just suppletion, is evidence for the greater inflectional potential of the category in question. (Croft 2003: 97)

In other words, internal declensional classes and prevalent stem alternations serve as distinctive features that contribute to defining the category ‘noun’ and set its members apart from the adjacent, complementary prototype in the same sphere: ‘verb’. We suspect that when the behavioral potential of ‘prototypical noun’ (object reference construction) is so salient and distinct to the point that a majority of the members exhibit unpredictable morphological subclasses and/or irregular stem alternations, the cognitive contrast between ‘noun’ and ‘verb’ will lead to a contrast in linguistic forms; that is, a prototypical parts of speech construction will not share the same encoding with a non-prototypical construction. Thus, an NPC as a non-prototypical construction will be encoded by a unique and overt strategy that is different from a prototypical action predication construction.

<table>
<thead>
<tr>
<th>MEANING</th>
<th>FUNCTION</th>
<th>Reference</th>
<th>Predication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>Prototypical noun</td>
<td>Non-prototypical construction (NPC)</td>
<td></td>
</tr>
<tr>
<td>Action</td>
<td>Non-prototypical construction</td>
<td>Prototypical verb</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Prototypes of parts of speech constructions (adapted from Croft and van Lier 2012).
Besides the repeated tendency observed in this study and in Hengeveld and Valstar (2010), the hypothesis is also supported by the observation that most of the languages reported to have high lexical flexibility exhibit mainly isolating or agglutinative morphological systems, and no (or less prominent) internal declensional classes, for example Malayo-Polynesian languages, Salishan languages, Wakashan languages, Mandarin and Archaic Chinese, among others. We would also expect to see the same tendency in other non-prototypical parts of speech constructions. For example, if action-denoting lexemes can be used for reference without structural coding in a given language (e.g., Mandarin Chinese), then there should be no internal conjugational classes in that language. We hope to extend and further examine this hypothesis in future studies.

In summary, the results of this study confirm Hengeveld and Valstar’s hypothesis as a general tendency but are not without counter-examples: lexical flexibility tends not to co-occur with internal lexical subclasses. And we believe that the interactions here can be explained by the prototype effect of parts of speech constructions: unpredictable internal subclasses (and stem-alternation) represent great behavioral potential of a category, signify a salient cognitive distinction between the two prototypes ‘noun’ and ‘verb’, and thus lead to differentiated encoding between prototypical and non-prototypical parts of speech constructions.

5.4 A general motivation for a copula strategy and against lexical flexibility

We have discussed several factors that can affect lexical flexibility in previous sections; namely, basic word order, tense/aspect marking and internal morphological classes. This list is obviously not exhaustive since we observe copula morphemes used in a group of Polynesian languages which do not have any of the features that may hamper lexical flexibility. The Polynesian languages in question all have a predicate-initial word order, mostly morphologically separable tense/aspect markers, no internal noun classes, and rare stem alternation in their nominal morphology. This suggests some other motivation(s) for developing and using overt structural coding for non-prototypical parts of speech constructions such as NPCs.

The morpheme ko and its cognates (o, ’o, or go) are shared by many Polynesian languages as predicate markers for nominal predicates (Bauer 1993; Brown and Koch 2016; Clark 1976; Kieviet 2017; Massam et al. 2006). The functional scope of these predicate markers varies across Polynesian languages. In languages such as Niuean (Oceanic, Niue) and Samoan, ko and its cognates are used to encode both classifying nominal predication (20a and 21a) and identity predication (20b and 21b).
(20) Niuean (Oceanic, Niue)

a. *ko e kamuta a au*
   PRED ABS carpenter ABS I
   ‘I’m a carpenter.’

b. *ko e takitaki gahua e fifine ia*
   PRED ABS boss work ABS woman that
   ‘That woman is the boss.’

(Seiter 1979: 53)

(21) Samoan (Oceanic, Samoa)

a. *o se fale Samoa t=o=u aiga fou*
   PRED ART (NSP.SG) house Samoa ART=POSS=1SG home new
   ‘My new home was a Samoan house.’
   (adapted from Mosel and Hovdhaugen 1992: 502)

b. *’o le Aso Gafua le aso muamua o le vaiaso*
   PRED ART Monday ART day first POSS ART week
   ‘Monday is the first day of the week.’
   (adapted from Mosel and Hovdhaugen 1992: 508)

In contrast, *ko* is used only to encode identity predication in other Polynesian languages such as Maori (Oceanic, New Zealand) (22a) and Rapa Nui (Oceanic, Chile) (23a). Classifying nominal predication (22b and 23b) in the two languages does not require extra structural coding.

(22) Maori (Oceanic, New Zealand)

a. *ko te rooia teenei*
   PRED the lawyer this
   ‘This is the lawyer.’

b. *he maahita ia*
   INDF teacher 3SG
   ‘He is a teacher.’
   (adapted from Bauer 1993: 79, 80)

(23) Rapa Nui (Oceanic, Chile)

a. *pero ko au te suerekao o te hora nei*
   but PRED 1SG ART governor of ART time PROX
   ‘But I am the governor now (or: the governor now is me).’
   (adapted from Kieviet 2017: 454)

b. *a Thor Heyerdahl he científico e tahi*
   PROP Thor Heyerdahl NTR scientist NUM one
   ‘Thor Heyerdahl was a scientist’
   (adapted from Kieviet 2017: 452)
Clark (1976: 38) observes in his reconstruction of Proto-Polynesian (PPN) that “the most plausible reconstruction of PPN is that *ko was required with definite NP predicates, but optional with indefinites. The extension of its use to indefinites is surely a natural syntactic generalization”. Given such extension from definite nominal predicates (typical identity predicates) to indefinite ones (typical classifying nominal predicates) and the other major function of ko and cognates as a topic/focus marker, we believe that the morpheme underwent a grammaticalization process where it developed from a grammatical element (an information structure marker) to a “further grammatical form” (a copula morpheme) (Narrog and Heine 2021: 1). This specific route is discussed in detail by Stassen (1997: 100–120) as Identity Takeover. He observes that overt information structure markers (“discourse functional elements” in his terms) are often recruited in identity predications and gradually grammaticalized into predicate markers, which may further extend to classifying nominal predications. The synchronic usage of ko and cognates in Polynesian languages aligns well with this pattern, indicating that the copula morpheme (predicate marker) has developed in these languages even though they exhibit none of the previously discussed typological features that may hamper lexical flexibility.

Stassen (1997) proposes a compelling explanation for the obligatory usage of discourse-motivated elements in identity statements, which is identified as the initial step of the grammaticalization path “Identity Takeover”. He argues that the differentiation between the two NPs in an identity statement, which refer to the same entity, is not sensitive to grammatical roles but is solely concerned with pragmatic-functional categories such as focus, topic/comment, or background/foreground. In prototypical predication constructions, the subject and predicate coincide with the discourse topic and focus by default, respectively. Therefore, overt information structure marking is only necessary when the pragmatically unmarked situation is overridden. On the other hand, in identity statements where the distinction between the subject and predicate is absent or elusive, discourse-motivated notions such as topic or focus have to be made explicit. And if a language has overt markers for these notions, such markers will be frequently or obligatorily used in identity statement and may eventually grammaticalize into the structural coding of the construction.

This explanation, however, does not address the extension of discourse-motivated elements towards the structural coding of classifying nominal predication, despite the semantic and cognitive affinity between an identity statement and a classifying nominal predication. The latter has a clear distinction between the subject and predicate, and discourse-motivated elements should not be necessary in pragmatically unmarked situations. On the basis of Stassen’s explanation, we would like to argue for a more general motivation of developing a copula morpheme for both types of NPCs: the need for more efficient parsing of constituents and processing of the construction.
Hawkins (2004) proposes the Maximal Online-processing Principle with which he attempts to address choices between competing language structures, including overt versus zero structural coding of a construction. One of the relevant examples he discusses is the usage of an overt complementizer for a complement clause, which corresponds to structural coding of a (clausal) action reference construction from an RCG perspective. In English, the complementizer *that* of a complement clause can often be omitted when the clause is not functioning as the subject. To omit the complementizer is undoubtedly a more economical strategy to encode the construction, but the zero strategy may lead to more “unassignment or misassignment of syntactic properties” and thus more efforts in processing, especially when the subject of the complement clause is non-case-marked and relatively long (Hawkins 2004: 58). In such cases, the overt complementizer is preferred because it helps to resolve the potential ambiguity earlier and renders more efficient processing. This is supported by Rohdenburg’s corpus investigation of the complement clause construction with the matrix verb *realize* (Rohdenburg 1999: 102, cited in Hawkins 2004: 59) (Table 10).

As the subject of the complement clause becomes longer and loses case marking (from pronouns to longer full NPs), there is an increasing preference for the complementizer *that*. Similar results are also observed in Shank et al. (2014) for English (*I think that* vs. *I think ∅*) and Boye et al. (2012) for Danish. While a complement clause with a zero complementizer is grammatical and comprehensible in both English and Danish, one with an overt complementizer could be more precise and less demanding in terms of parsing and processing, especially when the structure of the subordinate clause is rather complicated (for example, containing a complex subject).

The copula morpheme, as an overt structural coding of an NPC, can also serve the same purpose. An easier parsing of complex phrases or clauses motivates an overt copula in languages where the copula is otherwise optional.

Sneddon et al. (2012) note that the Indonesian copulas *adalah* and *ialah* are only used when the two constituents in an NPC are relatively complex and the construction may be difficult to parse without the copula. Compare (24a) with (24b) and (24c).

### Table 10: Relation between ∅ versus *that* complement and subject of the complement clause (adapted from Rohdenburg 1999: 102, cited in Hawkins 2004: 59).

<table>
<thead>
<tr>
<th>Subject of the complement clause</th>
<th>Zero complement</th>
<th><em>That</em> complement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal pronoun</td>
<td>48 % (127)</td>
<td>52 % (37)</td>
</tr>
<tr>
<td>1–2 word full NP</td>
<td>26 % (32)</td>
<td>74 % (89)</td>
</tr>
<tr>
<td>3+ word full NP</td>
<td>10 % (15)</td>
<td>90 % (130)</td>
</tr>
</tbody>
</table>
Indonesian (Malayic, Indonesia)

a. Ayah guru
   father teacher
   ‘Father is a teacher’

b. Ayah Tomo adalah pegawai Bank Indonesia
   father Tomo COP employee Bank of Indonesia
   ‘Tomo’s father is an employee of the Bank of Indonesia.’

c. Kain kebaya ialah pakaian wanita Jawa
   kain kebaya COP clothing woman Jawa
   ‘The kain kebaya is the clothing of Javanese women.’

(Sneddon et al. 2012: 247)

Pustet’s (2003) typological investigation of copulas also finds that if a language has a distinction between a formal style and a colloquial style in terms of copula usage, a copula may be omitted in the colloquial style, but is always required in the formal style, where economy is usually overruled by preciseness or iconicity.

Even for predicate-initial languages in which the predicate is relatively easy to identify (Hengeveld et al. 2004), a well-grammaticalized predicate marker can still provide an even more unambiguous and efficient identification of the predicate: in the case of the Niuean NPC, the predicate marker ko precedes the first constituent in a clause and explicitly marks it as the predicate, indicating its syntactic role and avoiding potential misassignment.\textsuperscript{9} Overt structural coding provides more explicit morphological cues for the assignment of syntactic/pragmatic roles within a construction, compared with zero structural coding.

The question of lexical (in)flexibility, as defined in Section 1, is ultimately a question of overt versus zero structural coding (of non-prototypical parts of speech constructions). We believe that behind this choice is a trade-off between (formal and paradigmatic) economy and iconicity (or processing economy). If a language is highly flexible in terms of parts of speech, and more non-prototypical meaning-function combinations are encoded without overt structural coding in the same way as prototypical ones, then language users will have less morphosyntactic cues for parsing constituents and assigning syntactic roles, and thus need more contexts and efforts to correctly process a sentence. But such a system is economical both syntagmatically and paradigmatically: there are fewer forms to articulate and process, and fewer paradigms to acquire and carry. For example, the same zero encoding is used for all types of predications in omnipredicative languages.

\textsuperscript{9} For example, see Frajzyngier et al. (2002) for potential ambiguity between NP-NP predication and NP-NP modification.
Conversely, if a language is less flexible in this regard and exhibits a clearer distinction of word classes, non-prototypical parts of speech constructions will be encoded by more overt structural coding. As a result, language users will have access to more morphosyntactic cues (overt structural coding) available for parsing of constituents and assignment of syntactic roles, but at the same time more elements to articulate and process, and more paradigms to acquire and carry.

Factors affecting lexical flexibility in languages, as discussed in previous studies, interact with this competition by either reinforcing or weakening the motivations for a particular side. For example, a dominant predicate-initial word order renders high identifiability of the predicate, thereby reducing the need for overt structural coding of predicates (Hengeveld et al. 2004). Predicate-initial languages are thus more likely to adopt a zero strategy for NPCs, although a copula strategy is not ruled out. In contrast, internal lexical subclasses (and stem-alternating morphology) deter a zero strategy due to iconicity: distinct categories tend to be encoded by different strategies, leading to a preference for a copula strategy, as discussed in Section 5.3. The influence of various typological features on this competition has not been exhausted in this investigation and we look forward to further exploration in future studies.

6 Conclusion

Within the framework of RCG, this study has investigated the encoding strategies of nominal predication constructions (NPCs) among V1 languages, which were reported to be highly flexible in the propositional function of predication. Based on the typology of encoding strategies, we examined potential correlations between NPC encoding and other typological features, namely, tense/aspect (T/A) marking and internal lexical subclasses.

The results show that V1 languages generally exhibit high flexibility in predication. Among the sample languages, 76.9% (50) allow entity-denoting lexemes to function as predicates without extra structural coding and 49.2% (32) only use zero encoding for NPCs.

Following Stassen's (1997) detailed critique of the Dummy Hypothesis, we found that tense/aspect marking has a limited impact on the encoding of NPCs. It is only the morphological combination of bound T/A markers and nominal predicates that is unlikely to occur cross-linguistically. Besides introducing a copula to carry the T/A markers, a language can also neutralize T/A morphology in NPCs to avoid the non-iconic combination.

In terms of internal lexical subclasses, the results support Hengeveld and Valstar's (2010) hypothesis as a tendency but not an absolute constraint: languages with internal noun classes prefer a copula strategy for encoding NPCs. We proposed a
tentative explanation for this pattern based on prototype effects. We suspect that internal lexical subclasses (and stem-alternating morphology) represent a salient cognitive distinction between ‘noun’ and ‘verb’ as two prototypical parts of speech, and thus lead to differentiated structural coding between prototypical and non-prototypical parts of speech constructions, for example NPC (non-prototypical) and action predication construction (prototypical).

Along the classification of encoding strategies, we observed the development of ko as a predicate marker in Polynesian languages, which is not motivated by factors that hinder lexical flexibility. We attempted to involve processing economy as a general motivation against lexical flexibility and for developing a copula morpheme. Lexical flexibility is ultimately a choice between zero versus overt structural coding, and a competition between formal/paradigmatic economy and processing economy or iconicity. Other grammatical features participate in this competition by reinforcing or weakening motivations for a particular side.

We await future studies to expand the investigation of lexical flexibility towards other non-prototypical parts of speech constructions (e.g., the action reference construction) and a wider scope of sample languages. Typological factors that are not yet related to this topic may also interact with the iconicity–economy competition of lexical (in)flexibility and remain to be explored.

**Research funding:** This work was supported by the Japan Science and Technology Agency (1st author) & Japan Society for the Promotion of Science (2nd author) and Pioneering Research Initiated by the Next Generation [J210002435] (1st author) & Grant-in-Aid for Scientific Research [21K00496] (2nd author).

**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2/3</td>
<td>1st/2nd/3rd person</td>
</tr>
<tr>
<td>ABS</td>
<td>absolutive</td>
</tr>
<tr>
<td>ACC</td>
<td>accusative</td>
</tr>
<tr>
<td>AGENT</td>
<td>agent nominalization</td>
</tr>
<tr>
<td>APPEN</td>
<td>appended vowel</td>
</tr>
<tr>
<td>ART</td>
<td>article</td>
</tr>
<tr>
<td>CAUS</td>
<td>causative</td>
</tr>
<tr>
<td>COP</td>
<td>copula</td>
</tr>
<tr>
<td>COS</td>
<td>change of state</td>
</tr>
<tr>
<td>DEF</td>
<td>definite</td>
</tr>
<tr>
<td>DEM</td>
<td>demonstrative</td>
</tr>
<tr>
<td>DET</td>
<td>determiner</td>
</tr>
<tr>
<td>EMP</td>
<td>emphatic particle</td>
</tr>
<tr>
<td>F</td>
<td>feminine</td>
</tr>
<tr>
<td>Language</td>
<td>Family ( &gt; Sub-family)</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Tamasheq</td>
<td>Afro-Asiatic &gt; Berber</td>
</tr>
<tr>
<td>Sukur</td>
<td>Afro-Asiatic &gt; Biu-Mandara</td>
</tr>
<tr>
<td>Standard Arabic</td>
<td>Afro-Asiatic &gt; Semitic</td>
</tr>
<tr>
<td>Garifuna</td>
<td>Arawakan &gt; Antillean Arawakan</td>
</tr>
<tr>
<td>Language</td>
<td>Family ( &gt; Sub-family)</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Baure</td>
<td>Arawakan &gt; Bolivian Arawakan</td>
</tr>
<tr>
<td>Ashéninka Perén</td>
<td>Arawakan &gt; Kampa-Amuesha</td>
</tr>
<tr>
<td>Nicobarese Car</td>
<td>Austroasiatic &gt; Nicobarian</td>
</tr>
<tr>
<td>Tboli</td>
<td>Austronesian &gt; Blic</td>
</tr>
<tr>
<td>Tukang Besi</td>
<td>Austronesian &gt; Celebic</td>
</tr>
<tr>
<td>Tagalog</td>
<td>Austronesian &gt; Central Philippine</td>
</tr>
<tr>
<td>Chamorro</td>
<td>Austronesian &gt; Chamorro</td>
</tr>
<tr>
<td>Nias</td>
<td>Austronesian &gt; Northwest</td>
</tr>
<tr>
<td>Anityum</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Gilbertese/Kiribati</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Hoava</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Kokota</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Maori</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Niufo'ou</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Niuean</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Rapa Nui</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Roviana</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Samoan</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Longgu</td>
<td>Austronesian &gt; Oceanic</td>
</tr>
<tr>
<td>Timugon Murut</td>
<td>Austronesian &gt; Southwest</td>
</tr>
<tr>
<td>Scottish Gaelic</td>
<td>Indo-European &gt; Celtic</td>
</tr>
<tr>
<td>Welsh</td>
<td>Indo-European &gt; Celtic</td>
</tr>
<tr>
<td>Wari'</td>
<td>Chapacuran</td>
</tr>
<tr>
<td>Lowland Oaxaca</td>
<td>Hokan/Tequistlatecan</td>
</tr>
<tr>
<td>Chontal</td>
<td>Isolate</td>
</tr>
<tr>
<td>Kuot</td>
<td>Isolate</td>
</tr>
<tr>
<td>Movima</td>
<td>Isolate</td>
</tr>
<tr>
<td>Language</td>
<td>Family ( &gt; Sub-family)</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Chol</td>
<td>Mayan &gt; Cholan</td>
</tr>
<tr>
<td>K’iche’</td>
<td>Mayan &gt; Quichean</td>
</tr>
<tr>
<td>Nandi</td>
<td>Nilotic</td>
</tr>
<tr>
<td>Pókoot</td>
<td>Nilotic</td>
</tr>
<tr>
<td>Turkana</td>
<td>Nilotic</td>
</tr>
<tr>
<td>Copala Triqui</td>
<td>Oto-Manguean &gt; Trique</td>
</tr>
<tr>
<td>Wemba Wemba</td>
<td>Pama–Nyungan &gt; Western</td>
</tr>
<tr>
<td>Yagua</td>
<td>Peba-Yagua</td>
</tr>
<tr>
<td>Didinga</td>
<td>Surmic</td>
</tr>
<tr>
<td>Majang</td>
<td>Surmic</td>
</tr>
<tr>
<td>Huehueta Tepehua</td>
<td>Totonacan</td>
</tr>
<tr>
<td>Nisga’a</td>
<td>Tsimshian</td>
</tr>
<tr>
<td>Pipil</td>
<td>Uto-aztecan &gt; Aztec</td>
</tr>
<tr>
<td>O’odham/Papago</td>
<td>Uto-aztecan &gt; Piman</td>
</tr>
<tr>
<td>Southern Tepehuan</td>
<td>Uto-aztecan &gt; Tepehuan</td>
</tr>
<tr>
<td>Makah</td>
<td>Wakashan</td>
</tr>
</tbody>
</table>

Table 11: Encoding strategies of NPCs in the sample languages (excluding Austronesian languages).
A secondary analysis excluding the Austronesian sample languages

In response to a reviewer’s concern regarding the over-representation of Austronesian languages in the sample, we conducted a secondary analysis excluding the Austronesian languages and the results are presented below.

As shown in Table 11, in terms of encoding strategies of NPCs, the prominence of zero strategies is weakened in comparison to the primary analysis presented in Section 5.1, which is expected as the majority of Austronesian sample languages employ a zero strategy for NPCs. However, the percentage of languages allowing a zero strategy to some degree (zero + mixed) (73.3%) is still higher than that in Stassen’s large-scale investigation (60.3%), reflecting a greater tolerance for zero strategies among V1 languages.

The correlation between internal morphological classes and lexical flexibility (zero strategy) is diluted and becomes statistically insignificant after removing the Austronesian languages which were contributors to the correlation (Table 12). Nevertheless, the trend is still observable in which V1 languages with internal morphological classes prefer a copula strategy for NPC while those without internal morphological classes tend to opt for a zero strategy.

**Table 12:** Internal morphological subclasses and encoding strategy of NPCs (excluding Austronesian languages).

<table>
<thead>
<tr>
<th>Morphological subclasses</th>
<th>Yes:</th>
<th>No:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Only zero</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Mixed (split/alternate)</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Only copula</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>24</td>
</tr>
</tbody>
</table>

**Table 13:** Morphological status of T/A markers and their availability in zero-marked NPCs (excluding Austronesian languages).

<table>
<thead>
<tr>
<th>Availability in ø NPC</th>
<th>Morphological status of T/A markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Separable</td>
</tr>
<tr>
<td>Available</td>
<td>15</td>
</tr>
<tr>
<td>Unavailable</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
</tr>
</tbody>
</table>
Within the sample languages allowing a zero strategy to some degree, the tendency that bound T/A markers do not occur on zero-marked nominal predicates remains salient and statistically significant even after removing the Austronesian sample languages (Table 13). A Pearson’s Chi-square test yields $\chi^2 = 7.8625$, $p = 0.005047$. The tendency is also conspicuous within the sample languages using mixed NPC strategies, as shown in Table 14.

Copula NPCs generally exhibit more T/A marking behavioral potential compared to zero NPCs within languages using mixed NPC strategies, especially regarding the inseparable T/A markers as indicated by the bolded values in Table 14. In summary, Austronesian is undoubtedly the largest language family featuring a V1 word order, taking up around 30 % of the world’s V1 languages. Nevertheless, the secondary analysis has showed that the tendencies examined in this study are not solely driven by the Austronesian family. After removing the Austronesian languages, the correlation between morphological status of T/A markers and their availability in zero NPC strategies remains salient and statistically significant. The correlation between NPC strategies (lexical flexibility) and internal morphological classes, and the general preference for a zero NPC strategy among V1 languages are weakened but still clearly observable.

### References


Chung, Sandra. 2020. Chamorro grammar. Available at: https://escholarship.org/uc/item/2sx7w4h5.


**Supplementary Material:** This article contains supplementary material (https://doi.org/10.1515/lingty-2023-0035).