Computational Complexity of Natural Morphology Revisited

Hajime Senuma[*1] and **Akiko Aizawa**[1,2]
[1]National Institute of Informatics, Tokyo, Japan
[2]The University of Tokyo, Tokyo, Japan
hajime.senuma@gmail.com, aizawa@nii.ac.jp

Abstract

This paper revisits a classical, yet fundamental, discussion of theoretical computational linguistics: the computational complexity of natural languages. Past studies have revealed that syntax, as observed in Swiss-German, is not weakly context-free. Concerning morphology, Culy (1985) employed a construction in Bambara to show that morphology is not weakly context-free; however, Manaster-Ramer (1988) pointed out that the Bambara case can be problematic because the wordhood of the construction is reliant on special tonal behaviors, and it is ambiguous whether the behaviors belong to the morphological domain. This raises doubts about whether the case can be considered a genuine morphological phenomenon. In this paper, we argue that Classical Ainu, a language we examine, also defies weak context-freeness at the morphological level. The construction we introduce is unambiguously morphological because this language's valency-sensitive structure and valency-changing operations, such as noun incorporation, preclude its grammatical interpretation as syntactic.

1 Introduction

1.1 On Generative Capacity

Noam Chomsky argued in his seminal works (Chomsky, 1956, 1959) that human languages follow a set of computational rules. This naturally leads to the question: What level of computational power is required to process human languages?

This has been a dominant issue in theoretical computational linguistics, and a wide range of arguments have often been published in natural language processing (NLP) journals such as *Computational Linguistics* (Manaster-Ramer, 1988; Radzinski, 1991; Kuhlmann et al., 2015). The complexity question has two strands of significance.

Cognitive Aspects. The central tenet of the Chomskyan school is that human beings have special brain components dedicated to language processing, and thus certain linguistic (meta-)rules are innate in nature (Hauser et al., 2002). The veracity of this claim has been hotly debated; some neuroscientific studies support the idea (Fedorenko and Blank, 2020; Malik-Moraleda et al., 2022), whereas others are skeptical (Tremblay and Dick, 2016). Studies on complexity and the nature of the encoded linguistic properties are directly relevant here (Pullum and Gazdar, 1982; Chesi and Moro, 2014; Fang et al., 2021).

Engineering Aspects. A classical topic is to investigate models and/or algorithms for parsing languages. Context-free problems are parsable using the Earley algorithm in $O(n^3)$ time, where n represents the length of a sentence (Earley, 1970). Generally, the strings of most natural morphology can be generated by 1-way finite state transducers (FSTs) (Roark and Sproat, 2007). However, in many cases, more expressive models such as 2-way FSTs are required to capture its *strong*[1] generative capacity (Dolatian et al., 2021). Recently, there has been growing interest in exploring the patterns and biases learned by neural networks (Weiss et al., 2018; Dolatian and Heinz, 2019; Torres and Futrell, 2023).

Several languages have been examined to assess the *weak* generative capacity of formal grammars for natural languages. It is important to note that a language is not *strongly* context-free if it can or cannot be generated by context-free grammars (CFGs); however, if it can be, there exist some parse trees that are not interpretable by linguistic theories. A language is not *weakly*

[*]Most of the work was done while the first author was affiliated with and financially supported by the University of Tokyo.

[1]See the next paragraph for the definition.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 649–663, 2024. https://doi.org/10.1162/tacl_a_00665
Action Editor: Alexander Clark. Submission batch: 10/2023; Revision batch: 12/2023; Published 5/2024.

context-free if it cannot be generated by *any* CFGs. Therefore, the denial of weak context-freeness is a more stringent condition than that of strong context-freeness.

1.2 Problems

Extensive research has been conducted on the topic of syntax (see Section 2.1). However, when it comes to morphology, there appears to be a dearth of substantial arguments beyond a paper discussing the vocabulary of Bambara (Culy, 1985). The paper resorted to the fact that a Bambara word construction ‘‘wulu(filèla)^h(nyinina)^i o wulu(filèla)^j(nyinina)^k’’ $(h, i, j, k \geq 1)$ forms a valid noun if and only if $h = j$ and $i = k$, and is otherwise ungrammatical. The Bambara noun roughly means ‘‘whoever watches who watches who watches . . . who watches who searches for who searches for . . . who searches for dog searchers.’’

However, the wordhood of these strings is entirely reliant on tonal behaviors. It was noted that Bambara exhibits two types of tonal rules: one for external sandhi (‘‘the interaction of adjacent lexical items,’’ that is, syntactic) and another for internal sandhi (‘‘the interaction of components of a compound,’’ that is, morphological). Culy further states that the *Noun o Noun* construction shown above has its own special tonal behavior, which is not syntactic. He thus concluded that this is morphological.

Manaster-Ramer (1988, pp. 101–102) questioned this point, arguing:

1. The paper does not provide specific tonal rules but instead refers to a Bambara textbook. The behavior may imply the exact opposite of Culy’s claim: It is a special behavior which is not morphological, and is thus syntactic.

2. Irregular behaviors of external sandhi are commonly seen in the world’s languages, and thus tonal behaviors are not suitable as an absolute criterion for word boundaries.

In reality, given that the boundary between morphology and syntax is not always clear, even in sophisticated linguistic theories, it can be problematic to rely on criteria such as ‘‘this phenomenon generally indicates wordhood,’’ whether phonological or lexical. How can one consciously establish that a construction is morphological rather than syntactic?

1.3 Solution

Our proposed solution is to create a construction that becomes explicitly ungrammatical when interpreted as a syntactic sequence.

In our research, we focus on Classical Ainu, an endangered language spoken by the Ainu, an indigenous people of the Island of Hokkaido in Japan who originally inhabited islands around the border of what is now Japan and Russia. The 17th edition of Ethnologue labeled Ainu as an `8b` or *nearly extinct* language (Lewis et al., 2013). Constant efforts are being made to revive this language (DeChicchis, 1995; Sato, 2012b; Ōno, 2022). No language has been confirmed to be genetically related to Ainu; thus, it is classified as a language isolate.

Classical Ainu, a language known for its highly polysynthetic nature, has a complex valency-computation system. The language can productively form new verbs by concatenating a variety of nouns (*noun incorporation*, NI) and affixes. In some cases, a single verb can express information that is almost equivalent to a sentence in other languages. *Valency*, or the number of arguments a verb can take, is recalculated along with this concatenation process. Several morphemes increase or decrease valency, whereas others constrain it. Using these properties, we devised a construction that was grammatical (but not weakly context-free) at the morphological level and ungrammatical at the syntactic level.

The rest of the paper is organized as follows. In Section 2, we summarize existing studies on the topic. Section 3 provides linguistic data to prepare our proof. Section 4 presents the construction and proof. In Section 5, we discuss the issue from various perspectives. Finally, in Section 6, we offer concluding remarks.

2 Related Work

2.1 On Natural Languages

Chomsky first formulated a hierarchy of computational complexity in terms of human languages and investigated whether it exceeded the level of context-free languages (Chomsky, 1956, 1959). Although several arguments have been made, including those on NI in Mohawk (Postal, 1964), almost all have been rejected by Pullum and

Gazdar (1982). After nearly 30 years of expeditions, linguists have finally discovered actual cases of non-context-free properties: the cross-serial dependency in the Swiss-German syntax (Huybregts, 1984; Shieber, 1985) and the reduplication of the Bambara morphology (Culy, 1985). Pullum (1986) provided a stimulating history of this period. Moreover, Swedish is not context-free because of the extraction phenomena and obligatory presence of resumptive pronouns (Miller, 1991).

After the discovery of non-context-freeness, the focus shifted to whether human languages were acceptable by *mildly context-sensitive grammars* (MCSGs) (see the next section). Attempts to prove the weak inadequacy of MCSGs include the following: crossing-dependency in Dutch (Manaster-Ramer, 1987; Groenink, 1997); the number-naming system in Mandarin (Radzinski, 1991); *Suffixaufnahme* in the Old Georgian syntax (Michaelis and Kracht, 1997), which was counter-argued by Bhatt and Joshi (2004); and recursive copying in the syntax of Yoruba and its kin languages (Kobele, 2006).

In the field of phonology, Heinz (2010) claimed that phonological patterns are subregular (later known as the Subregular Hypothesis).

2.2 On Formal Languages

Among the earliest studies on formal languages, Langendoen's (1977) work is most relevant to our study, because the technique that he developed has been used in many subsequent proofs (Shieber, 1985; Culy, 1985; Miller, 1991), including ours.

Joshi (1985, p. 225) hypothesized three universal and invariant properties that every natural language must have: 1. limited cross-serial dependencies, 2. constant growth, and 3. polynomial parsing. Joshi named the class of beyond context-free formal grammars that satisfy all these properties as MCSGs. This includes tree-adjoining grammars (TAGs), linear indexed grammars (LIGs), combinatory categorial grammars (CCGs), head grammars (HGs), multiple context-free grammars (MCFGs), and Stabler's (1996) formalism of the minimalist program. This class is significant because, from a cognitive perspective, it accounts for natural languages well, and from an engineering perspective, it is very efficient for parsing. Vijay-Shanker and Weir (1994) proved the weak equivalence of TAGs, LIGs, CCGs, and HGs. Kuhlmann et al. (2015), however, found modern versions of CCGs are in fact less powerful than TAGs. Furthermore, the work by Kanazawa and Salvati (2012) and Salvati (2015) challenges the coherence of MCSGs.

2.3 Statistical Complexity of Morphology

Whereas the complexity of a linguistic construction was traditionally studied in terms of formal languages, recent studies attempt to examine the complexities of languages from the perspective of statistical metrics. For instance, Cotterell et al. (2019) showed inflectional paradigms could not be both large and highly irregular at the same time.

Park et al. (2021) revealed that difficulty to obtain a good language model (LM) was positively correlated with several statistical complexity measures of the language's morphology. Morphologically simple languages are better modeled by general tokenizers such as BPE, while moderately complex languages benefit from linguistically motivated ones such as Morfessor, although they are outperformed by simple unigram for polysynthetic languages. Zevallos and Bel (2023) further investigated this nature and demonstrated a method to reduce the amount of language data to obtain good LMs for low-resource languages.

3 Data

This section provides real-world examples in a natural language that supports the validity of a set of formal rules used in Section 4.

Similar to the case of the defunct language Old Georgian, which Michaelis and Kracht (1997) used in their complexity study, today it is practically impossible to directly assess the grammaticality of Classical Ainu words with native speakers because this language is highly endangered (Sato, 2012b; Lewis et al., 2013). Therefore, we collected linguistic data to support the construction as far as possible.

3.1 Grammar Books and Dictionaries on Ainu

Refsing (1986) wrote the first modern grammar book on Ainu, which is also an essential source on the Shizunai dialect. The first part of a book by Shibatani (1990) is by far the most widely referenced Ainu grammar text. An article

by Tamura (1988), whose English translation is Tamura (2000), provides more detailed, precise accounts. It also includes a short survey on the linguistic studies of Ainu from the 18th century to the 1980s. Although written in Japanese, a textbook by Sato (2008) provides plenty of Ainu utterances (both colloquial and Classical) obtained directly from native speakers of the Chitose dialect. Each chapter of a handbook edited by Bugaeva (2022b) delineates linguistically interesting phenomena by an expert on the topic.

Among modern dictionaries, Tamura's (1996) Ainu-Japanese dictionary on the Saru dialect is the most comprehensive one. In addition, Nakagawa's (1995) Ainu-Japanese dictionary on the Chitose dialect provides insightful accounts of many complex phenomena in the language. Whereas Kayano's Ainu-Japanese dictionary (Kayano, 2002) on the Saru dialect lacks linguistic information, such as parts-of-speech and the valency of verbs, it is a valuable resource for Ainu, as the author himself was a native speaker. Kirikae's (2003) dictionary/commentary on *Ainu Shin'yoshu*, notable *kamuyyukar*-type poems, serves as a solid corpus for the Classical Ainu language. Bugaeva's (2011) colloquial Ainu dictionary has detailed English glossings and recorded utterances from a native speaker.

3.2 Notation

For interlinear glossing, unless otherwise noted, we follow the Leipzig Glossing Rules and Standard Abbreviations, rev. February 2008 (Comrie et al., 2008), with the following redefinitions and additions; INT for intensive, ITERA for iterative, MID for middle voice, MONO for monovalent, POLY for polyvalent, and POSS for the *possessed* case (not possessive).

3.3 Basics

Traditionally, the Ainu language had no writing systems. From the early 20th century, two types of systems have been co-developed: one with the modified Japanese katakana and the other with the Latin alphabet.

Linguistically, Ainu is characterized by agglutinativity, polysynthesis, and NI. Although the latter two properties are not so overt in colloquial Ainu, they are fully employed in Classical Ainu, a variant mainly used for epics, tales, poetry, songs, incantations, rituals, and other artistic or religious activities.

In Classical Ainu, with NI, one can rephrase a sentence ''I make an inaw (a wooden prayer symbol)'' as ''I inaw-make'' (Shibatani, 1990, p. 63; glossing follows him, with bracketed supplements):

(1) a. *Inaw a-ke*
[prayer] 1SG-make
'I make a wooden prayer symbol.'

b. *Inaw-ke-an*
[prayer-]make-1SG

Notice the difference in the pronominal affixes: the prefix *a-* is used in (a) and the suffix *-an* in (b). As mentioned previously, Ainu is sensitive to *valency*, or the number of arguments that can be taken by verbs. For a first person subject, one must use the suffix *-an* for monovalents (so-called intransitives) and the prefix *a-* for polyvalents (so-called transitives).[2] *ke* is a bivalent, so a polyvalent prefix *a-* is used. Conversely, *inawke* is a monovalent formed by NI; because the incorporated noun filled a direct object slot of the original verb *ke*, its valency decreased by one, so *inawke* takes a monovalent suffix *-an*. Neither **inaw kean*[3] nor **ainawke* is grammatical.

In addition to NI, Ainu has another powerful tool called *applicative formation* (AF). Each time a verb is prefixed by an applicative, its valency increments by one and a newly formed verb accommodates an additional argument as a direct object. For instance (Shibatani, 1990, p. 65; glossing follows him):

(2) a. *A-kor kotan ta sirepa-an.*
1SG-have village to arrive-1SG
'I arrived at my village.'

b. *A-kor kotan a-e-sirepa.*
1SG-APPL-arrive

Again, notice the differences in the pronominal affixes. The monovalent *sirepa* (''to arrive'') requires the suffix *-an*. When it takes the allative applicative prefix *e-* and forms a new bivalent *esirepa* (''to arrive at X''), its valency increments by one. Thus, it requires the prefix *a-*.

[2]Note that these are for Classical Ainu. Colloquial Ainu uses a slightly different set of pronominal affixes.

[3]Do not be confused with *inawkean* (1b).

Type	Position	Word forms	Valency	Productivity
Pronominal	Circumfix	$\langle ku, \emptyset \rangle, \langle \emptyset, as \rangle, \langle \emptyset, \emptyset \rangle, \langle en, as \rangle, \ldots$	n/a	Obligatory
Applicative I	Suffix	ko-, e-	+	High
Applicative II	Suffix	o-	+	Middle
Causative I	Suffix	-e, -re, -te	+	High
Causative II	Suffix	-ka, -ke	+	Low
Reciprocal	Prefix	u-	−	Middle
Reflexive	Prefix	si-, yay-	−	Middle
Middle	Prefix	ci-	−	Middle
Antipassive	Prefix	i-	−	Middle
Anticausative	Either	si-, -ke	−	Low
Incorporation	Prefix	(nouns)	−	Middle
Aspectual	Suffix	-kosanu, -natara, -atki, . . .	0	Low
Metrical Expletive	Prefix	u-, kur-	0	n/a
Reduplication I	Suffix	(full)	0	High
Reduplication II	Suffix	CVC-VC	0	Low
Reduplication III	Suffix	CVCV-CV	0	Low
Qualification	Prefix	e- (‘‘head’’), o- (‘‘tail’’)	0	Middle

Table 1: List of verbal affixes (Tamura, 1988; Shibatani, 1990; Sato, 2008; Bugaeva, 2015; Bugaeva and Kobayashi, 2022). The list is neither exhaustive nor precise, but is intended to give readers an idea of the high degree of polysynthesis.

Its polysynthetic nature allows various affixes and affixal operations. Table 1 lists major verbal affixes. By combining NI, AF, and the other morphemes, it is possible to construct a complex (but still single) word, such as *aeyaykotuymasir-amsuypa* (Shibatani, 1990, p. 72; glossing follows him):

(3) *a-e-yay-ko-tuyma-si-ram-suy-pa.*
1SG-APPL-REFL-APPL-far-REFL-heart-sway-ITERA

‘I keep swaying my heart afar and toward myself over (something)’ = ‘I wonder about (something)’

Readers may question how it is possible to know that such a complex verb is a single word, rather than a phrase with multiple words. In the case of Ainu, the following two criteria are informative:

Phonological Criterion. Ainu is a pitch-accent language. In modern orthography, a high pitch syllable is marked by an acute on its vowel. The position of a high vowel is distinctive, e.g., *nísap* (‘‘suddenly’’) and *nisáp* (‘‘shin’’) (Tamura, 1996, p. 423). Additionally, the mechanism for determining the pitch position is fairly regular. For almost all words, the second vowel is accented if the first syllable is open (as in *kamúy*, ‘‘god’’), whereas the first vowel is accented if the first syllable is closed (as in *áynu*, ‘‘human’’). This mechanism is also applied when a new word is formed by affixes. For example, from *sapá* (‘‘head’’), *kusápaha* (‘‘my head’’) is formed (Tamura, 1988, p. 13).[4] In summary, a hearer can usually determine word boundaries by detecting the positions of high pitches.

Lexical Criterion. Several affixes, such as the reflexive prefix *yay-* and the aspectual suffix *-kosanu*, attach only to a verb. Therefore, the presence of these affixes suggests that the chunks are classified as a verb. Pronominal affixes are also useful markers for this purpose, because they always occur as the last elements of the verb formation process and directly indicates the boundaries of the words.

However, these criteria are not always applicable.

As for the first criterion, we seldom have recordings. In addition, in human language, pitch

[4]Because of this high consistency, in writing, the acute on the common position is often omitted.

positions can be changed by factors such as word length and individual habits.

As for the second criterion, it is not always possible to find salient pronominal affixal strings, because the most frequent one is the third-person pronoun, that is, the empty string ∅. Furthermore, several nouns can take pronominal affixes, in which case it is difficult to determine whether they are incorporated nouns or direct object nouns.

However, at least under some constructions, it is possible to unambiguously determine the wordhood, as seen in Section 2 (such as that used in Example 4.6, devised to satisfy the above lexical criterion).

3.4 Properties of NI and AF

3.4.1 Overlapping Semantic Roles of Applicatives

Ainu has three types of applicative prefixes: *e-*, *o-*, and *ko-*. Most of their functional differences (and similarities) are empirically well known (Tamura, 1996). Bugaeva (2010) derived various statistics on the usage of these prefixes, and uncovered that several semantic roles are shared by more than one applicatives. For example, she showed that both *e-* and *ko-* have the same semantic role of Cause/Purpose (6% of *e-* and 3% of *-ko* occurred in her corpus).

3.4.2 Number of Applicatives

Shibatani (1990, pp. 76–77) suggested the following ordering of verbal prefixes: 1. subject prefix, 2. object prefix, 3. applicative, 4. generalized object or reflexive or reciprocal, 5. applicative, and 6. verb.

This formulation limits the number of applicatives to two, and is in fact applicable to a large number of verbs. However, several verbs that were collected from informants exceed this limit, indicating that the formulation above is not imperative. For example, the bivalent *eyaykewutum-ekosanniyo* is formed from the monovalent *sanniyo* with three applicatives, a NI, and a reflexive. The following instances were taken from Tamura (1996):

(4) a. eyaykewtum-ekosanniyo (p. 153)

e-yay-kewtum-e-ko-sanniyo
APPL-REFL-heart-APPL-APPL-know

'to think in one's heart about something'

b. ewkoyaykopuntek (p. 146; kopuntek: p. 332)

e-u-ko-yay-ko-puntek
APPL-RECP-APPL-REFL-APPL-be.pleased

'(for several people) to be pleased with something together'

c. eyaykouwepeker (p. 155; uwepeker: p. 808)

e-yay-ko-u-e-peker
APPL-REFL-APPL-RECP-APPL-be.clear

'to think over one's troubles'

3.4.3 Redundant NI

Now, we examine the possibility of using the same nouns multiple times. As Classical Ainu poetry values periphrasis and verbosity, it is possible to form a tautological statement such as ''I am holding a red staff as a staff'', as in Kirikae (2003, p. 140):

(5) hure kuwa ekuwakor

húre kuwa ∅-e-kuwa-kor
be.red staff 3-APPL-staff-have

'She staff-held a red staff.'

If the object *húre kuwa* is incorporated, a new verb may be *húrekuwaekuwakor*. In fact, there is at least one case where the same noun is incorporated twice, as in *orupkorupus* (Sato, 2012a, p. 15).

(6) orupkorupus

o-rup-ko-rup-us
groin-ice-APPL-ice-put

'(for one's groin) to be freeze-frozen'

As we have seen, it is semantically tautological but still grammatical to incorporate the same noun more than once as a form of poetical rhetoric.

3.4.4 Reduplication

Ainu has a rich reduplication system. For verbs, Tamura (1988, pp. 65–66) identified three types of reduplication.

1. *The full reduplication of stems or roots* functions as an iterative and/or intensive aspect. For example, *kik* ''to hit once'' and *kikkik* ''to rain punches''.

2. *The reduplication of VC in CVC* functions as a progressive and incessant aspect. For example, *cirir* (cir-ir) ‘‘(for a stream) to trickle’’.

3. *The reduplication of the last CV in CVCV* functions as a progressive and intensive aspect. For example, *sikcupupu* (sik-cupu-pu) ‘‘to keep on narrowing one’s eyes’’.

Though the second and third types are limited to certain cases, Tamura (1988) claimed that the first one is productive. It is also possible to reduplicate a verb formed using NI (Kirikae, 2003, pp. 109, 265, 315)[5]:

(7) a. hokushokus

ho-kus∼hokus
tail-pass.through∼INT

(lit.) ‘to pass through one’s buttocks severely’ = ‘fall down, overturn’

b. chipokonannpe kohokushhokush

∅-cip-o	*kor*	*ne an pe*
3-boat-ride	and.then	it

∅-ko-hokushokus
3>3-APPL-overturn

‘They boarded a boat, and then they rowed it hard as if it overturns.’

Additionally, it is possible to reduplicate a verb formed by AF (Kirikae, 2003, pp. 107, 319):

(8) a. kosankosan

ko-san∼kosan
APPL.go.downstream∼INT.ITERA

b. chikor wenpuri unkosankosan

ci-kor	*wen*	*puri*
1SG-have	bad	habit

un-kosankosan
3>1SG-go.downstream.INT.ITERA

‘My evil thoughts flowed into me impulsively.’

Note that reduplication does not change the valency. Both *hokus* and *hokushokus* are monovalent (*kohokushokus* is bivalent because of AF), and both *kosan* and *kosankosan* are bivalent.

Example (8) also shows that its reduplication is morphological and not syntactic. If syntactic, then its form is *unkosan unkosan*. Because Classical Ainu also favors the syntactic reduplication of a verb, in some cases, it is ambiguous whether the process is morphological or syntactic. However, the presence of some verbal affixes uniquely determines the level of reduplication.

3.4.5 Periphrastic Affixal Idiom

The high polysynthesis of Classical Ainu reaches the point where the language provides various idioms of affixes. The previous examples demonstrated several of them: *yay-ko* (REFL-APPL = ‘‘by oneself’’ or ‘‘alone’’) and *u-ko* (RECP-APPL = ‘‘together’’).

These idioms can form circumfixes. For example, the periphrastic circumfix *ci-...-re* is the combination of the middle voice prefix *ci-*[6] and the causative suffix *-re*[7] (Tamura, 1996, p. 48). Semantically, the idiom is meaningless. It is a poetic device used only to embellish verbs in an elegant and graceful tone.

This idiom does not change valency, as the decrement by *ci-* and increment by *-re* cancel each other out. The idiom is also unique in that *ci-* alone affixes to polyvalents but not to monovalents, whereas *ci- -re* can be affixed to both monovalents and polyvalents. For example, the monovalent *cihopunire* is derived from the monovalent *hopuni* (Tamura, 1996, p. 48):

(9) cihopunire

ci-hopuni-re
MID-happen-CAUS

‘(for a serious matter) to happen’

In contrast, **cihopuni* is ungrammatical. Such idiomatic circumfixes are useful for identifying word boundaries, because they encircle a verb and prevent it from further concatenating with other morphemes.

[5]As Kirikae (2003, pp. 21, 26) himself does not see the case as NI (where *ho-* is not a free morpheme) unlike Tamura (2000, pp. 196–197), better NI examples that use free morphemes are *mawunmawun*, *nimunimu*, and *niyusniyus* (Tamura, 1996, pp. 383, 417, 430).

[6]There is another morpheme of the form *ci-*, the exclusive first-person plural pronoun (‘‘we, excluding you’’) for polyvalents, which also selects polyvalent verbs like the middle *ci-*.

[7]Or its allomorphs *-e* and *-te*.

3.4.6 Summary

To summarize, in Classical Ainu,

- applicatives have overlapping semantic roles (this property was used in Example 4.5),
- there is no limit on how many times applicatives can be affixed (used in Example 4.5),
- the same noun can be incorporated iteratively (used in Example 4.5),
- a verb can be fully reduplicated (used in Definition 4.3 and Example 4.6), and
- an affixal idiom may form a circumfix, a useful tool for identifying word boundaries (used in Example 4.6).

4 Proof

We define ad-hoc lexical models that are intended to capture the properties of the Classical Ainu lexicon. More refined approaches may model the lexical semantics of verb arguments (Comrie et al., 2015; Bugaeva, 2022a, pp. 43–44); however, it is beyond the scope of this study.

Definition 4.1 (Base Valency-Sensitive Lexicon)**.** First, we define a *base valency-sensitive lexicon* $L \subseteq \Sigma \times \mathsf{Z} \times I \times \tau$, where

- Σ is a set of strings, which roughly indicates a set of basic lexemes.
- Z is a set of integers, which corresponds to the base valency if the entry is a verb, otherwise a valency-increasing / decreasing.
- I is an integer range (e.g., [0, 1]), which corresponds to valency constraints. If an entry has no constraint, this term takes an empty set $\emptyset$.
- $\tau = \{\text{head}, \text{free}, \text{prefix}, \text{suffix}, \text{isolate}\}$ is a set of types, each of which specifies the type of affixation or incorporation. ‘‘head’’ represents a verb; ‘‘free’’ represents a free morpheme that can prefix to verbs (many nouns and a few adverbs are included in this group); ‘‘prefix’’ and ‘‘suffix’’ are bound morphemes which must occur as affixes to verbs; ‘‘isolate’’ means it does not affix to any verbs.

There may be several lexemes (entries) for one string. For instance, in English, the verb ‘‘give’’ has three entries $(\text{give}, 1, \emptyset, \text{head})$, $(\text{give}, 2, \emptyset, \text{head})$, $(\text{give}, 3, \emptyset, \text{head})$, which correspond to monovalent (intransitive), bivalent (monotransitive), and trivalent (ditransitive), respectively.

Definition 4.2 (Valency-Sensitive Lexicon)**.** Given a base valency-sensitive lexicon L, a *valency-sensitive lexicon* $L^* \subseteq \Sigma^* \times \mathsf{Z} \times I \times \tau$ (countably infinite) is recursively defined as follows:

- If $x \in L$, then $x \in L^*$.
- If $(x, a, \emptyset, \text{head}) \in L^*$, $(y, b, r, \text{free}) \in L^*$, and $a \in r$, then $(yx, a + b, \emptyset, \text{head}) \in L^*$.
- If $(x, a, \emptyset, \text{head}) \in L^*$, $(y, b, r, \text{prefix}) \in L^*$, and $a \in r$, then $(yx, a + b, \emptyset, \text{head}) \in L^*$.
- If $(x, a, \emptyset, \text{head}) \in L^*$, $(y, b, r, \text{suffix}) \in L^*$, and $a \in r$, then $(xy, a + b, \emptyset, \text{head}) \in L^*$.

For example, given a bivalent (monotransitive) verb $(\text{kar}, 2, \emptyset, \text{head})$ meaning ‘‘to make something’’, and a noun $(\text{cise}, -1, [2, \infty), \text{free})$ meaning ‘‘house’’, a new monovalent (intransitive) verb $(\text{cisekar}, 1, \emptyset, \text{head})$ meaning ‘‘to build a house (lit. to house-make)’’[8] can be formed using an operation called noun incorporation.

It is also possible to form an avalent verb; e.g., the concatenation of $(\text{sir}, -1, [1, \infty), \text{prefix})$ ‘‘environment’’ and $(\text{sések}, 1, \emptyset, \text{head})$ ‘‘to be hot’’ forms an avalent $(\text{sirsesek}, 0, \emptyset, \text{head})$ ‘‘it is hot’’.[9] In English, semantically avalent expressions (e.g., ‘‘it is hot’’ and ‘‘it rains’’) must take dummy subjects to satisfy syntactic constraints, but Ainu does not have such constraints. The maximum valency of Classical Ainu is not known, but Bugaeva (2015, p. 828) reported that there is at least one tetravalent verb $(\text{korere}, 4, \emptyset, \text{head})$ ‘‘to make someone give something to someone.’’

Definition 4.3 (Full Reduplicative Valency-Sensitive Lexicon)**.** Because Ainu verbs are productive for full reduplication (see Section 3.4.4), we also define a *full reduplicative valency-sensitive lexicon* $\tilde{L}^*$ such that

- If $l \in L^*$, then $l \in \tilde{L}^*$.

[8] Tamura (1996, p. 60).
[9] Tamura (1996, p. 658).

- If $(x, v, \emptyset, \text{head}) \in L^*$, then $(xx, v, \emptyset, \text{head}) \in \tilde{L}^*$.
- The three affixation rules (last three rules in Definition 4.2 with modifying L^* into $\tilde{L}^*$).

Definition 4.4 (Vocabulary)**.** We further define a *vocabulary*, or a set of strings produced by a lexicon, $V(L) = \{w \mid (w, \cdot, \cdot, \cdot) \in L\}$.

Example 4.5. The set of basic Classical Ainu lexemes L_{CA} contains the following items:

- (siknak, $1, \emptyset$, head) ''to be unable to see''[10]
- (núpe, $-1, [2, \infty)$, free) ''tears''[11]
- (e, $1, [0, \infty)$, prefix) (see Section 3.3)
- (ko, $1, [0, \infty)$, prefix) (see Section 3.3)
- (ci, $-1, [2, \infty)$, prefix) (see Section 3.4.5)
- (re, $1, [0, \infty)$, suffix) (see Section 3.4.5)

The valency-sensitive lexicon L^*_{CA} then has the following items, where $m, n \geq 1$:

- (esiknak, $2, \emptyset$, head) ''to be unable to see something''[12] or ''to be unable to see because of something''[13]
- (núpeesiknak, $1, \emptyset$, head) ''to be unable to see because of tears''[12]
- ((núpee)^{n}siknak, $1, \emptyset$, head)
- (ko(núpee)^{n}siknak, $2, \emptyset$, head)
- (núpeko(núpee)^{n}siknak, $1, \emptyset$, head)
- ((núpeko)^m(núpee)^{n}siknak, $1, \emptyset$, head)

The last verb should mean ''to be unable to see because of tears,'' which is the same as núpeesiknak. For the reasoning behind this derivation, refer to Sections 3.4.1, 3.4.2, and 3.4.3.

Example 4.6. The full reduplicative valency-sensitive Classical Ainu lexicon $\tilde{L}^*_{CA}$ has the following items, where $w_{m,n} =$ (núpeko)^m(núpee)^{n}siknak ($m, n \geq 1$), $x =$ ci, $y =$ re.

- ($w_{m,n}w_{m,n}, 1, \emptyset$, head) roughly, ''to be unable to see because of lots of tears'' (see Section 3.4.4 for reduplication)
- ($w_{m,n}w_{m,n}y, 2, \emptyset$, head) ''to let someone be unable to see because of lots of tears''
- ($xw_{m,n}w_{m,n}y, 1, \emptyset$, head) ''to be unable to see because of lots of tears (in a grave tone)'' (see Section 3.4.5 for ''ci-...-re'')

Example 4.7. The string sequence $w_{m,n}w_{m,n}$ has both morphological and syntactic interpretations, because syntactic reduplication is also productive in Ainu. If we consider the string $w_{m,n}w_{m,n} \in V(\tilde{L}^*_{CA})$, this sequence is interpreted as a single word. If we consider the string $w_{m,n} \in V(\tilde{L}^*_{CA})$, this sequence is interpreted as a phrase containing two identical words.

On the other hand, the string sequence $xw_{m,n}w_{m,n}y$ only has a morphological interpretation. Let us assume $xw_{m,n}w_{m,n}y$ can be divided into more than one words. Considering x (= ci) and y (= re) are bound morphemes that can only appear as affixes to verbs, the only possible way to divide this sequence is $xw_{m,n}$ and $w_{m,n}y$.

$w_{m,n}y$ is in $V(\tilde{L}^*_{CA})$, as $w_{m,n} = (w_{m,n}w_{m,n}, 1, \emptyset, \text{head}) \in \tilde{L}^*_{CA}$ and (re, $1, [0, \infty)$, suffix), and $1 \in [0, \infty)$.

However, $xw_{m,n}$ is not in $V(\tilde{L}^*_{CA})$, as $(w_{m,n}w_{m,n}, 1, \emptyset, \text{head}) \in \tilde{L}^*_{CA}$ and (ci, $-1, [2, \infty)$, prefix), but $1 \notin [2, \infty)$.[14]

This result contradicts the assumption; therefore, $xw_{m,n}w_{m,n}y$ has only a morphological interpretation.

The unambiguously morphological construction $xw_{m,n}w_{m,n}y$ exhibits a symptom of non-context-freeness. Nevertheless, we do not directly prove that this is the case. Even if a subset of a language is not context-free, it is not always true that the entire language is not context-free. To prove the non-context-freeness of the vocabulary of Classical Ainu, we used the fact that context-free languages are closed under

[10] Nakagawa (1995, p.208) and Tamura (1996, p. 630).
[11] Tamura (1996, p. 442).
[12] Tamura (1996, p. 124).
[13] Bugaeva (2010, p. 767).
[14] Note that the base verb *siknak* has only monovalent usage (unlike the English *give*), if we believe Nakagawa (1995, p. 208) and Tamura (1996, p. 630). This is relevant, because if the verb had a bivalent entry, it could be possible to construct a lexeme ($w_{m,n}w_{m,n}$, 2, $\emptyset$, head).

intersection with regular languages. This trick is often attributed to Bar-Hillel et al. (1961), but the mathematically rigorous form was developed by Langendoen (1977) as an improvement upon the work of Bar-Hillel et al. It was used in subsequent studies, including Culy (1985).

Theorem 4.8. *$V_{CA} = V(\tilde{L}^*_{CA})$, the vocabulary of Classical Ainu, is not context-free.*

Proof. Assume V_{CA} is context-free.

Consider the following regular language:

$$F = \{xw_{h,i}w_{j,k}y \mid h, i, j, k \geq 1\},$$

where w, x, y are as defined in Example 4.6.

Let $N = V_{CA} \cap F$. Then,

$$N = \{xw_{m,n}w_{m,n}y \mid m, n \geq 1\}.$$

Intersecting a context-free language with a regular language yields another context-free language (Hopcroft et al., 2007, Theorem 7.27 [p. 291]). Therefore, if V_{CA} is context-free, so is N. Now, if we construct a string homomorphism h such that $h(\text{núpeko}) = 0, h(\text{núpee}) = 1, h(\text{ci}) = \epsilon, h(\text{re}) = \epsilon, h(\text{siknak}) = \epsilon$, then

$$\begin{aligned} h(N) &= \{0^m1^n0^m1^n \mid m, n \geq 1\} \\ &= \{xx \mid x \in F, F = \{0^m1^n \mid m, n \geq 1\}\}. \end{aligned}$$

Because context-free languages are closed under homomorphism (Hopcroft et al., 2007, Theorem 7.24 [p. 290]), $h(N)$ shall also be context-free. This $h(N)$ is, however, a class of non-context-free languages called *copying languages*; a copying language $L = \{xx \mid x \in F\}$, where F is a regular language, is not context-free unless there is a finite string r and finitely many finite strings q and s such that $F = \{qr^ns \mid n \geq 0\}$ (Langendoen, 1977, Theorem 7.c). By contradiction, V_{CA} is not context-free. □

5 Discussion

5.1 Mental Lexicon and NI

After Mithun (1984) analyzed NI, the related literature has become ‘‘a microcosm of linguistic theory’’ (Massam, 2009) because NI lies at the interface of morphology and syntax.

Mithun stated that NI is ‘‘perhaps the most nearly syntactic of all morphological processes’’ (Mithun, 1984, p. 847), and went on to say that ‘‘Formally, it is a morphological process, not a syntactic one; and it shares all the characteristics unique to such process’’ (Mithun, 1984, p. 891). The fact that syntax is not context-free suggests that morphological processes with NI are also not context-free. We demonstrated that this is true.

She also reported an interesting story about the mental lexicon. Mohawk speakers remember NI that has actually occurred, and feel excited when hearing a new (and grammatical) NI that they had never heard before, because complex NI is considered an elegant art (Mithun, 1984, pp. 872, 889).

However, as a literary register, the products of Classical Ainu are usually memorized and recited, rather than freely produced. Therefore, it may raise the question whether it is a valid tool to test the linguistic capacity of humans.

Concerning this issue, we point out folklore collected by Nakagawa (2006), which has a unique style that combines *uwepeker* (prose literature) and *kamuyyukar* (a type of poetry). The storyteller Nabe Shirasawa explained that she felt *sakehe* (refrains) used in the first part of the original poem were rustic and boring, and therefore she reframed that part in the prose form.

This suggests that though the Classical register is often used in works whose forms are fixed, native Ainu poets can freely tell tales in Classical Ainu if the situation demands it.

5.2 The Lexicality of NI

The lexicality of NI is a subject of classical debate (Kroeber, 1910; Sapir, 1911; Kroeber, 1911). Even today, some linguists who advocate transformational types of grammar (such as the government and binding theory) claim that NI is syntactic rather than morphological (Baker, 1988, 1996). This argument is important in the literature, because head-movement, one of the major syntactic operations in their theory, is built upon the syntacticity of NI (Anderson, 2000). Anderson (2000) criticized this analysis as being reasonable for their theory, but not in terms of accounting for NI.

Moreover, their claim comes at the expense of the *Lexicalist Hypothesis* (Anderson, 2000). This foundational hypothesis, attributed to Chomsky (1970), states that words are the basic and atomic units of syntax. It is also formulated as ‘‘The syntax neither manipulates nor has access to the internal form of words’’ (Anderson, 1992, p. 84). As the Lexicalist Hypothesis is presupposed in

various theories, including modern NLP frameworks such as Universal Dependencies (Nivre, 2015, p. 6), interpreting NI as syntactic operations abolishes not only our argument but also many of the existing models in computational linguistics. Thus, it should be reasonable to follow the classic hypothesis and maintain that NI (and the construction we have devised) is morphological.[15]

6 Conclusion

This study demonstrates that Classical Ainu is not weakly context-free unambiguously at the morphological level because of its valency-sensitive structure and valency-changing operators. Thus, we reconfirmed the weak inadequacy of CFGs for natural morphology.

However, the properties of valency, polysynthesis, NI, and AF have not yet been fully explored in computational linguistics. Further examinations may provide more insight into natural languages and aid cognitive and engineering studies.

Acknowledgments

The authors express their sincere appreciation to the action editor and the three anonymous reviewers for their extensive and insightful feedback. We also extend deep gratitude to Makoto Kanazawa, who provided valuable comments about formal language studies.

References

Stephen R. Anderson. 1992. *A-Morphous Morphology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511586262

Stephen R. Anderson. 2000. Lexicalism, incorporated (or incorporation, lexicalized). In *Proceedings of the Chicago Linguistic Society*, volume 36.

Mark C. Baker. 1988. *Incorporation*. Chicago University Press.

Mark C. Baker. 1996. *The Polysynthesis Parameter*. Oxford University Press. https://doi.org/10.1093/oso/9780195093070.001.0001

Y. Bar-Hillel, M. Perles, and E. Shamir. 1961. On formal properties of simple phrase structure grammars. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 14:143–172. https://doi.org/10.1524/stuf.1961.14.14.143

Rajesh Bhatt and Aravind Joshi. 2004. Semilinearity is a syntactic invariant: A reply to Michaelis and Kracht 1997. *Linguistic Inquiry*, 35(4):683–692. https://doi.org/10.1162/ling.2004.35.4.683

Benjamin Bruening. 2018. The lexicalist hypothesis: Both wrong and superfluous. *Language*, 94(1):1–42. https://doi.org/10.1353/lan.2018.0000

Anna Bugaeva. 2010. Ainu applicatives in typological perspective. *Studies in Language*, 34(4):749–801. https://doi.org/10.1075/sl.34.4.01bug

Anna Bugaeva. 2011. Internet applications for endangered languages: A talking dictionary of Ainu. *Waseda Institute for Advanced Study Research Bulletin*, 3:73–81.

Anna Bugaeva. 2015. Valency classes in Ainu. In Bernard Comrie and Andrey Malchukov, editors, *Introducing the Framework, and Case Studies from Africa and Eurasia*, volume 1, pages 807–854. De Gruyter Mouton. https://doi.org/10.1515/9783110338812-025

Anna Bugaeva. 2022a. 1 Ainu: A head-marking language of the Pacific Rim. In Anna Bugaeva, editor, *Handbook of the Ainu Language*, pages 21–56, De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9781501502859-002

Anna Bugaeva, editor. 2022b. *Handbook of the Ainu Language*. De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9781501502859

Anna Bugaeva and Miki Kobayashi. 2022. 15 Verbal valency. In Anna Bugaeva, editor, *Handbook of the Ainu Language*, pages 515–548. De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9781501502859-016

[15] We are aware that there are debates in linguistics whether the Lexicalist Hypothesis is invalid and morphology ought to be subsumed under syntax, notably between Bruening (2018) and Müller (2018); however, the detail is beyond the scope of this paper. We also refer to Nedergaard et al. (2020) and Satō (2022) for recent researches on polysynthetic languages that offer insights into the interface between morphology and syntax.

Cristiano Chesi and Andrea Moro. 2014. Computational complexity in the brain. In *Measuring Grammatical Complexity*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685301.003.0013

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124. https://doi.org/10.1109/TIT.1956.1056813

Noam Chomsky. 1959. On certain formal properties of grammars. *Information and Control*, 2(2):137–167. https://doi.org/10.1016/S0019-9958(59)90362-6

Noam Chomsky. 1970. Remarks on nominalization. In R. Jacobs and P. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 184–221. Ginn, Waltham, MA.

Bernard Comrie, Iren Hartmann, Martin Haspelmath, Andrej Malchukov, and Søren Wichmann. 2015. Introduction. In Bernard Comrie and Andrey Malchukov, editors, *Introducing the Framework, and Case Studies fro Africa and Eurasia*, volume 1, pages 3–26. De Gruyter Mouton. https://doi.org/10.1515/9783110338812-004

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig glossing rules. Technical report, Max Planck Institute for Evolutionary Anthropology and University of Leipzig.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342. https://doi.org/10.1162/tacl_a_00271

Christopher Culy. 1985. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy*, 8(3):345–351. https://doi.org/10.1007/BF00630918

Joseph DeChicchis. 1995. The current state of the Ainu language. *Journal of Multilingual and Multicultural Development*, 16(1–2):103–124. https://doi.org/10.1080/01434632.1995.9994595

Hossep Dolatian and Jeffrey Heinz. 2019. Learning reduplication with 2-way finite-state transducers. *Proceedings of Machine Learning Research*, 93:67–80. https://doi.org/10.18653/v1/W18-5807

Hossep Dolatian, Jonathan Rawski, and Jeffrey Heinz. 2021. Strong generative capacity of morphological processes. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 228–243. Association for Computational Linguistics.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102. https://doi.org/10.1145/362007.362035

Hongjian Fang, Yi Zeng, and Feifei Zhao. 2021. Brain inspired sequences production by spiking neural networks with reward-modulated STDP. *Frontiers in Computational Neuroscience*, 15(February):1–13. https://doi.org/10.3389/fncom.2021.612041, PubMed: 33664661

Evelina Fedorenko and Idan A. Blank. 2020. Broca's area is not a natural kind. *Trends in Cognitive Sciences*, 24(4):270–284. https://doi.org/10.1016/j.tics.2020.01.001, PubMed: 32160565

Annius V. Groenink. 1997. Mild context-sensitivity and tuple-based generalizations of context-grammar. *Linguistics and Philosophy*, 20(6):607–636. https://doi.org/10.1023/A:1005376413354

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579. https://doi.org/10.1126/science.298.5598.1569, PubMed: 12446899

Jeffrey Heinz. 2010. Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4):623–661. https://doi.org/10.1162/LING_a_00015

John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2007. *Introduction to Automata Theory, Languages, and Computation*. Pearson Education, Inc.

Riny Huybregts. 1984. The weak inadequacy of context-free phrase structure grammars.*Van*

Periferie Naar Kern, pages 81–99. Foris Publications.

Aravind K. Joshi. 1985. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 206–250. Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.007

Makoto Kanazawa and Sylvain Salvati. 2012. MIX is not a tree-adjoining language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 666–674.

Shigeru Kayano. 2002. *Kayano Shigeru no Ainugo Jiten: Zouho-ban [The Ainu Dictionary by Shigeru Kayano: Expanded Edition]*. Sanseidō.

Hideo Kirikae. 2003. *Ainu Shin'yōshū Jiten [Lexicon to Ainu Shin'yōshū]*. Daigaku Shorin.

Gregory Michael Kobele. 2006. *Generating Copies: An Investigation into Structural Identity in Language and Grammar*. Ph.D. thesis, University of California, Los Angeles.

A. L. Kroeber. 1910. Noun incorporation in American languages. In *Verhandlungen Der XVI Internationaler Amerikanisten-Kongress [Reprinted in The Collected Works of Edward Sapir, Vol. 5, 1990]*, pages 569–576. Hartleben.

A. L. Kroeber. 1911. Incorporation as a linguistic process. *American Anthropologist*, 13(4):577–584. https://doi.org/10.1525/aa.1911.13.4.02a00070

Marco Kuhlmann, Alexander Koller, and Giorgio Satta. 2015. Lexicalization and generative power in CCG. *Computational Linguistics*, 41(2):187–220. https://doi.org/10.1162/COLI_a_00219

D. Terence Langendoen. 1977. On the inadequcy of type-3 and type-2 grammars for human languages. *Studies in Descriptive and Historical Linguistics: Festschrift for Winfred P. Lehmann*, pages 159–171. https://doi.org/10.1075/cilt.4.12lan

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2013. *Ethnologue: Languages of the World, Seventeenth Edition*. SIL International, Dallas, Texas, US.

Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. 2022. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25(8):1014–1019. https://doi.org/10.1038/s41593-022-01114-5, PubMed: 35856094

Alexis Manaster-Ramer. 1987. Dutch as a formal language. *Linguistics and Philosophy*, 10(2):221–246. https://doi.org/10.1007/BF00584319

Alexis Manaster-Ramer. 1988. Book reviews: The formal complexity of natural language. *Computational Linguistics*, 14(4):98–103.

Diane Massam. 2009. Noun incorporation: Essentials and extensions. *Linguistics and Language Compass*, 3(4):1076–1096. https://doi.org/10.1111/j.1749-818X.2009.00140.x

Jens Michaelis and Marcus Kracht. 1997. Semilinearity as a syntactic invariant. *Logical Aspects of Computational Linguistics*, pages 329–345. Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0052165

Philip H. Miller. 1991. Scandinavian extraction phenomena revisited: Weak and strong generative capacity. *Linguistics and Philosophy*, 14(1):101–113. https://doi.org/10.1007/BF00628305

Marianne Mithun. 1984. The Evolution of Noun Incorporation. *Language*, 60(4):847–894. https://doi.org/10.1353/lan.1984.0038

Stefan Müller. 2018. The end of lexicalism as we know it? *Language*, 94(1):e54–e66. https://doi.org/10.1353/lan.2018.0014

Hiroshi Nakagawa. 1995. *Ainugo Chitose Hōgen Jiten [The Ainu{Japanese Dictionary: Chitose Dialect]*. Sōfūkan.

Hiroshi Nakagawa. 2006. Ainu Kōshō Bungei Tekisuto Shū 7: Shirasawa Nabe Kōjutsu, ''Ōkami ga Ningen no Hahaoya ni Gyakutai Sareta'' [Ainu oral literature text collection 7: ''A Wolf was Abused by a Human Mother,'' narrated by Nabe Shirasawa]. *Journal of Chiba University Eurasian Society*, 9:219–256.

Johanne S. K. Nedergaard, Silvia Martínez-Ferreiro, Michael D. Fortescue, and Kasper

Boye. 2020. Non-fluent aphasia in a polysynthetic language: Five case studies. *Aphasiology*, 34(6):675–694. https://doi.org/10.1080/02687038.2019.1643000

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14–20, 2015, Proceedings, Part I*, Springer International Publishing, pages 3–16. https://doi.org/10.1007/978-3-319-18111-0_1

Tetsuhito Ōno. 2022. 12 The history and current status of the Ainu language revival movement. Anna Bugaeva, editor, *Handbook of the Ainu Language*, pages 405–442. De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9781501502859-013

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276. https://doi.org/10.1162/tacl_a_00365

Paul M. Postal. 1964. Limitations of phrase structure grammars. In Jerry A. Fodor and Jerrold J. Katz, editors, *The Structure of Language: Readings in the Philosophy of Language*, pages 137–151. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Geoffrey K. Pullum. 1986. Footloose and context-free. *Natural Language & Linguistic Theory*, 4(3):409–414. https://doi.org/10.1007/BF00133376

Geoffrey K. Pullum and Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*, 4:471–504. https://doi.org/10.1007/BF00360802

Daniel Radzinski. 1991. Chinese number-names, tree adjoining languages, and mild context-sensitivity. *Computational Linguistics*, 17(3):277–300.

Kirsten Refsing. 1986. *The Ainu Language: The Morphology and Syntax of the Shizunai Dialect*. Aarhus University Press.

Brian Roark and Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. 4. Oxford University Press, Great Clarendon Street, Oxford, UK.

Sylvain Salvati. 2015. MIX is a 2-MCFL and the word problem in Z2 is captured by the IO and the OI hierarchies. *Journal of Computer and System Sciences*, 81(7):1252–1277. https://doi.org/10.1016/j.jcss.2015.03.004

Edward Sapir. 1911. The problem of noun incorporation in American languages. *American Anthropologist*, 13:250–282. https://doi.org/10.1525/aa.1911.13.2.02a00060

Tomomi Sato. 2008. *Ainugo Bunpō no Kiso [The Basics of the Ainu Grammar]*. Daigaku Shorin.

Tomomi Sato. 2012a. Ainugo Chitose Hōgen ni okeru Meishi Hōgō: Sono Shurui to Kanren Shokisoku (Noun incorporation in the Chitose dialect of Ainu: Its types and related rules). *Bulletin of the Hokkaido Ainu Culture Research Center*, 18:1–31.

Tomomi Sato. 2012b. Ainugo no Genjō to Fukkō [The present situation of the Ainu language and its revitalization]. *GENGO KENKYU (Journal of the Linguistic Society of Japan)*, 142:29–44. https://doi.org/10.11435/gengo.142.0_29

Tomomi Satō. 2022. 16 Noun incorporation in Ainu. In Anna Bugaeva, editor, *Handbook of the Ainu Language*, pages 549–572. De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9781501502859-017

Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge University Press.

Stuart M. Shieber. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343. https://doi.org/10.1007/BF00630917

Edward P. Stabler. 1996. Derivational minimalism. In *Logical Aspects of Computational Linguistics: First International Conference, LACL '96 Nancy, France, September 23–25, 1996 Selected Papers*, Springer Berlin Heidelberg, pages 68–95. https://doi.org/10.1007/BFb0052152

Suzuko Tamura. 1988. Ainugo [The Ainu language]. In *Gengogaku Daijiten*, volume 1, pages 6–94. Sanseidō.

Suzuko Tamura. 1996. *Ainugo Saru Hōgen Jiten [The Ainu{Japanese Dictionary: Saru Dialect]*. Sōfūkan.

Suzuko Tamura. 2000. *The Ainu Language*. Sanseidō.

Charles J. Torres and Richard Futrell. 2023. L0-regularization induces subregular biases in LSTMs. In *Proceedings of the Society for Computation in Linguistics*, volume 6. University of Massachusetts Amherst. https://doi.org/10.7275/SS3-D749

Pascale Tremblay and Anthony Steven Dick. 2016. Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language*, 162:60–71. https://doi.org/10.1016/j.bandl.2016.08.004, PubMed: 27584714

K. Vijay-Shanker and D. J. Weir. 1994. The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory*, 27(6):511–546. https://doi.org/10.1007/BF01191624

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision RNNs for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 740–745, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2117

Rodolfo Zevallos and Nuria Bel. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.699