A Computational Approach to Analyzing Language Change and Variation in the Constructed Language Toki Pona

Daniel Huang[1], Hyoun-A Joo[2]

[1]Georgia Institute of Technology – dzh@gatech.edu
[2]Georgia Institute of Technology – joo.hyouna@gatech.edu

Abstract

This study explores language change and variation in Toki Pona, a constructed language with approximately 120 core words. Taking a computational and corpus-based approach, the study examines features including fluid word classes and transitivity in order to examine (1) changes in preferences of content words for different syntactic positions over time and (2) variation in usage across different corpora. The results suggest that sociolinguistic factors influence Toki Pona in the same way as natural languages, and that even constructed linguistic systems naturally evolve as communities use them.

1 Introduction

Human language typically evolves through centuries of usage, adaptation, and cultural exchange, gradually developing vocabulary and changing grammar. Thus, language change is a natural process and studying it “helps us explain the features of language structure because it provides a window onto how those structures come into being and evolve” (Bybee, 2015, p. 1). In contrast with these natural languages, constructed languages are artificially created with an intentionally designed phonology, grammar, orthography, and vocabulary. Constructed languages enjoy continued fascination, evidenced by the invention of thousands of different artificial languages in the past few centuries (Peterson, 2015). These include languages like Esperanto intended for international communication and artistic creations like Elvish and Klingon that bring fictional worlds to life.

Constructed languages provide a unique environment to study language variation and change when adopted by a speaking community. When a language is deliberately designed with specific constraints or features, we can observe how these structures evolve over time and how speakers innovate within existing frameworks. Constructed languages, therefore, provide a sandbox to identify drivers of change.

Toki Pona, the language examined here, tests the extremes of what a language can be. It was created by Canadian linguist Sonja Lang in 2001, designed with approximately 120 words and a grammar with very few rules. Lang (2014) published *Toki Pona: The Language of Good*, documenting its grammar and core vocabulary. Today, Toki Pona is primarily spoken on online platforms such as Reddit, Discord, and Facebook, with tens of thousands of community members and thousands of conversationally proficient speakers, making it the second most spoken constructed language in the world after Esperanto (Meulen, 2021). Because of its small lexicon, easy-to-parse grammar, and large community, Toki Pona is an ideal subject for a computational study on real language use.

2 Toki Pona: fluid word classes and transitivity

Toki Pona is an isolating language and exhibits a strict SVO order. All words, except proper names, are written in lowercase. Nouns do not inflect for number or definiteness, and verbs do not inflect for tense or aspect. Content words in Toki Pona can be used as a noun, adjective, adverb, or verb, i.e., the word classes are fluid. To delimit phrases, Toki Pona uses several particle words. Consider the usage of *moku* ’to eat’ in (1).[1]

(1) a. *moku* as a transitive verb

jan	li	moku	e	kili
person	PRED	eat	TR	fruit

‘The person is eating fruit.’

b. *moku* as a noun

moku	ni	li	suwi
food	DEM	PRED	sweet

‘This food is sweet.’

The particle *li* in (1) marks the following phrase as the predicate, which can be a noun with zero copula, an adjective, or a verb. The particle *e* in (1a) marks the following phrase as the direct object of a verb. Thus, when *moku* is used after *li*, it takes on the lexical meaning ’to eat.’ However, when used at the beginning of the sentence in (1b), it must be a noun and therefore converts, now meaning ’that which is eaten, food.’

[1] All Toki Pona examples were glossed using the Leipzig Glossing Rules (Comrie et al., 2024).

This fluidity extends to transitivity. If supplied with a direct object marked by *e*, nouns and adjectives are converted into transitive verbs. When an adjective is converted to a transitive verb, it typically becomes causative, as *pona* ’good’ in (2b). In contrast, in (2a), *pona* is used as a predicative adjective without a direct object assigned, describing the subject.

(2) a. *pona* as a predicative adjective

soweli	li	pona
land animal	PRED	good

‘The dog is good.’

b. *pona* as a transitive verb

lipu	li	pona	e	sona
document	PRED	make.good	TR	knowledge

‘Books make knowledge better.’ / ‘Books improve knowledge.’

This feature of fluidity presents an opportunity to analyze the distribution of content words in the different word classes and examine variation in their use. Because the particles that determine the word class of the content words take specific positions within the syntactic structure of Toki Pona as laid out above, the frequencies of different words in particular syntactic positions can be studied. The following research questions informed the study: (1) how do the frequencies of words in different syntactic positions change over time, and (2) to what extent does the usage of content words in different syntactic positions differ between informal and formal contexts?

3 Methods

3.1 Corpora

The data came from two corpora. The first, *ma pona pi toki pona* (“ma pona pi toki pona”, 2025), means “a good place for Toki Pona,” and is a Discord server with over 16,000 members. Discord is a realtime chat application that hosts large communities with several parallel chatrooms. As such, it contains primarily informal, conversational data. The downloaded channels contained 5.97 million total sentences, 1.23 million of which were scored as Toki Pona sentences with a total of 6.39 million tokens of Toki Pona, spanning from 2016 through January 2025.

The second corpus, *poki Lapo* (kala Asi et al., 2025), means “the collection named Lapo,” and is a monolingual Toki Pona corpus with long-form content like books, poetry, song lyrics, comics, and blog posts. It is continually updated with new published works and spans from 2002 to 2025. Because the two corpora represent different uses of Toki Pona (more conversational and informal versus more formal), a comparison between them will allow for insights regarding patterns of use and variation.

3.2 Filtering and tokenizing data

A significant majority of the messages in the informal corpus, *ma pona pi toki pona*, is in English and had to be filtered out.

The Python library `sona-toki` 'language knowledge' by Danielson (2025) takes a text as input, cleans the text to remove irregularities such as duplicated characters and punctuation, splits the text into sentences and further into tokens, and scores each of the sentences for whether or not they are Toki Pona. After splitting the text into tokens, it uses a variety of heuristics such as the ratio of Toki Pona words to non-Toki-Pona words and the phonotactic restrictions for proper names to generate a final output: a filtered list of sentences segmented into tokens.

Danielson (2024b) uses a specific configuration of `sona-toki` to create an n-gram corpus, that is, a dataset mapping Toki Pona phrases of various lengths to their frequencies. The same configuration was used to extract and filter sentences from both corpora in this research.

3.3 Parsing sentences

Before tagging parts of speech, each sentence had to be converted into a hierarchical structure in order to account for syntactic ambiguity (Section 3.4) and more easily address features like transitivity. For instance, the particle *e* that determines whether a verb is transitive could be separated from the verb by adverbs, but placing the phrases in a hierarchical structure makes the connection more easily identifiable. Therefore, the first author developed a parser for Toki Pona using the Earley (1970) algorithm with a specific implementation in the programming language JavaScript called *nearley* (Chandra & Radvan, 2020).

The Earley parsing algorithm defines a context-free grammar, a kind of phrase structure grammar (Chomsky, 1956), which is a set of rules in the form of $X \rightarrow Y...Z$. This allows the syntax of a language to be described by breaking down phrases into their constituents with statements like `PP -> P NP` (a prepositional phrase can be broken down into a preposition followed by a noun phrase). The first author created simplified phrase structure grammar based on the Toki Pona grammar rules, illustrated in Table 1.

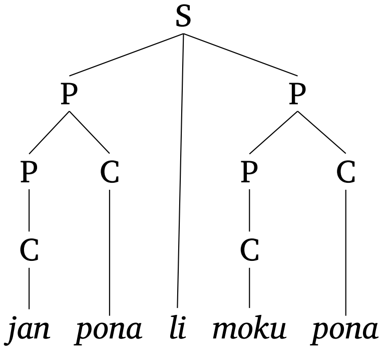
[2] The full grammar used to process the corpora can be found online at https://github.com/cubedhuang/ilo-nasin-sin/blob/main/src/lib/grammar.ne, with an interactive demo at https://nasin.nimi.li.

Table 1: An example of a simplified Toki Pona-like context-free grammar[2]

Grammar Rule	**Description**
S -> P_1 "li" P_2	A sentence (S) is composed of phrase (P_1), the particle *li*, and phrase (P_2), where P_1 is the subject and P_2 is the predicate.
P -> C	A phrase (P) can consist of a single content word (C).
P -> P C	A phrase (P) can alternatively consist of a combination of a phrase and a content word (C), whereby C is an adjective that must be attached to the phrase.

The *nearley* library will then use the grammar rules in order to generate a program that takes a sequence of tokens as input and outputs a resulting structure (or fails if the tokens do not conform to the grammar). Using the grammar above results in the hierarchical structure in Figure 1.

Figure 1: The resulting parse of *jan pona li moku pona* ‘the good person eats well’ generated by *nearley* with the grammar in Table 1.



This hierarchical structure allows for more advanced algorithm-based part-of-speech tagging by segmenting the sentence into phrases and revealing the structure of those phrases beforehand.

3.4 Resolving syntactic ambiguity

Toki Pona, like natural languages, has syntactic ambiguity. One instance of syntactic ambiguity is in prepositions. Toki Pona’s prepositions are a closed set of five words (*lon* ’at,’ *tawa* ’to,’ *tan* ’from, because of,’ *kepeken* ’by means of, with,’ *sama* ’like’), but all prepositions can also function as content words based on context. This is different from the case of fluid word classes laid out in Section 2. Examples in (3) illustrate the different uses of *tawa* ’to; moving.’

(3) a. *tawa* as a prepositional predicate

jan	li	tawa	sike
person	PRED	to	circle

‘The person goes to the ball.’

b. *tawa* as a non-prepositional content word

jan	li	tawa	sike
person	PRED	move	circle

‘The person moves circularly.’ / ‘The person spins.’

The Earley algorithm is particularly useful for such cases of ambiguity because it generates all possible syntactic interpretations of a string of tokens. However, only one parse should ultimately be chosen for each sentence.

To choose the most probable parse, the first author developed a heuristic scoring algorithm that favors specific patterns over others. For the preposition ambiguity illustrated in (3), it would prefer the prepositional interpretation (3a) when available.[3] This prioritization of interpretations reflects observed usage patterns in the Toki Pona community.

3.5 Tagging parts of speech

After sentences are represented hierarchically, a part-of-speech tagging module takes the parse structures as input and outputs a string of tokens tagged with parts of speech. The specific tags for Toki Pona are listed in Table 2.

Tokens tagged as IVerb or TVerb are collectively referred to as Verb, and the Noun, Mod, and Verb tags are referred to as Content.

When all tokens are tagged, the counts of each use of part of speech is aggregated by year, based on when the sentences were sent or published. For each year in each corpus, a distribution of words is created like in Figure 2.[6]

[3] The full implementation of the heuristic algorithm is found online at https://github.com/cubedhuang/ilo-nasin-sin/blob/main/src/lib/parser.ts.

[5] A complete implementation of the tagging algorithm is found online at https://github.com/cubedhuang/ilo-nasin-sin/blob/main/src/lib/tag.ts.

[5] The term ‘preverb’ originates from Lang (2014)’s description of Toki Pona’s grammar. Preverbs are a closed class of words that go between the predicate marker and the main predicate, such as *wile* ’to want to (be),’ *kama* ’to begin to; to become’ (e.g. *jan li wile moku* ’the person wants to eat’). Like prepositions, they double as content words, so their interpretation as a preverb or as the main verb is decided by context.

[6] A full copy of the aggregated data as well as a set of visualization tools can be found online at https://nasin.nimi.li/visualize.

Table 2: Token types and the contexts in which words are tagged by them[4]

Token Type	Context
NOUN	The first or only word in a subject, direct object, or prepositional object.
MOD	Modifiers, collectively adjectives and adverbs, that directly follow nouns and verbs.
IVERB	Intransitive verbs, typically the first word in the predicate when no direct object is supplied afterwards.
TVERB	Transitive verbs, typically the first word in the predicate when a direct object is supplied afterwards.
INTJ	The first word in a phrase when it is used without a complete sentence.
PREP	Prepositions, specifically when interpreted prepositionally.
PREVERB	Preverbs, lexical elements that precede a verb.[5]
PART	Grammatical particles and ordinal numbers.

4 Results and discussion

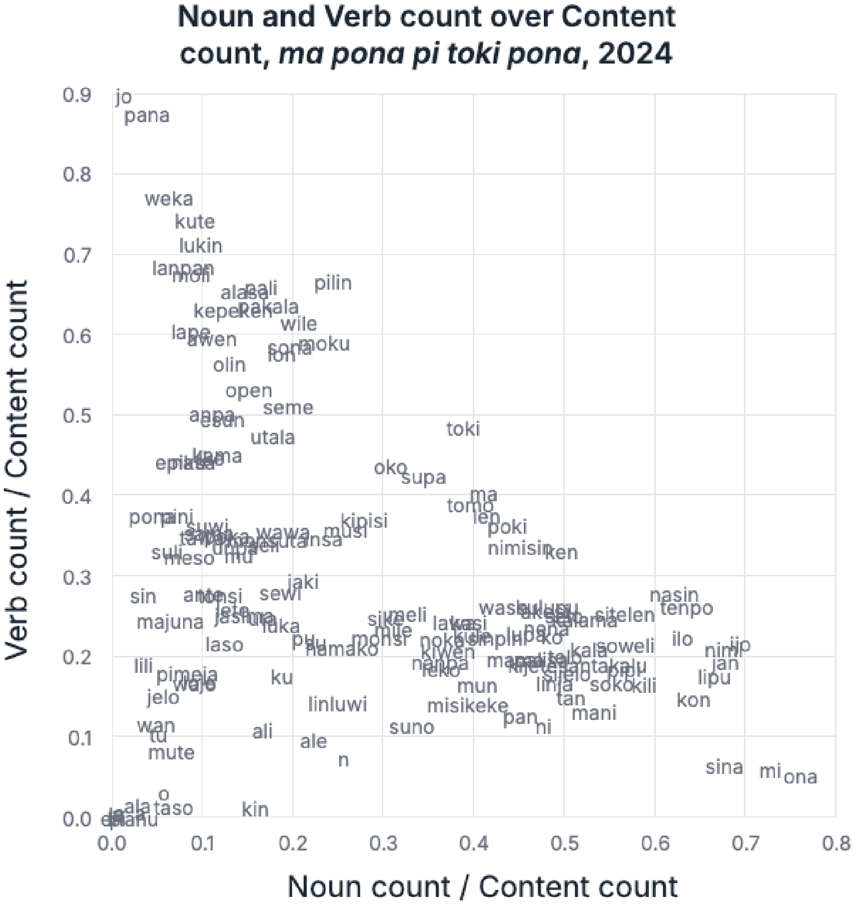
This study aimed to explore how content words' preferences for certain syntactic positions change over time and how the usage of content words in different syntactic positions differ between informal and formal contexts. The analysis revealed instances of diachronic variation in the use of several words and patterns that varied across two corpora.

Results will be presented in two sections: First, diachronic variation will be discussed, focusing on the changes in the use of body-part words as transitive verbs and the use of *pu* 'interacting with the official Toki Pona book' as a noun. Second, variation between the informal and formal corpora will be discussed, focusing on the use of interjections and the adoption of features across the two corpora.

4.1 Diachronic variation

The first research question aimed to examine how content words' preferences for certain syntactic positions change over time. In order to pursue the research questions, we chose words from the corpora that showed a high degree of change over time and whose change was referenced often in the Toki Pona community.

Figure 2: Example of distribution of words across different content word classes for the year 2024 (*ma pona pi toki pona* corpus)



4.1.1 Nouns as transitive verbs

As discussed in Section 2, because Toki Pona makes no syntactic distinction between content words, all content words can also become transitive verbs when followed by a direct object that is introduced with the transitive marker *e*. The body-part words *luka* ’hand, limb, branch’ and *uta* ’mouth, lips’ are nouns but can be used as transitive verbs meaning ’to touch by hand’ and ’to touch by mouth’ respectively, as shown in (4) and (5).

(4) a. *luka* as a noun

waso	li	lukin	e	luka	mi
bird	PRED	see	TR	hand	my

‘The bird sees my hands.’

b. *luka* as a transitive verb

jan	li	luka	e	waso
person	PRED	hand	TR	bird

‘The person touches the bird (with their hand).’

(5) a. *uta* as a noun

uta	mi	li	pilin	ike
mouth	1.POSS	PRED	feel	bad

‘My mouth feels bad.’

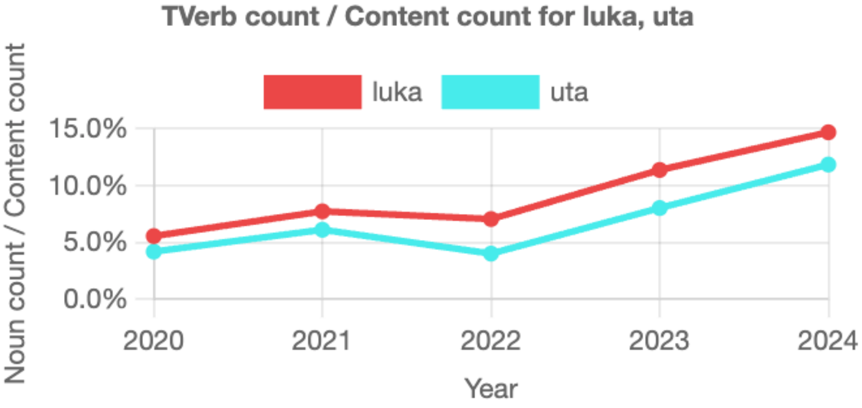
b. *uta* as a transitive verb

ona	li	uta	e	olin	ona
3	PRED	mouth	TR	love	3

‘They kissed the one they love.’

These body-part words stand out because their usage in the informal corpus *ma pona pi toki pona* has drastically changed over time. As Figure 3 shows, the nouns *luka* and *uta* are increasingly used as transitive verbs in informal use of Toki Pona. This usage still falls within the grammar written in *Toki Pona: The Language of Good*, as the meaning after conversion is a physical application of the noun to the direct object. However, only more recently has this usage proliferated in the community.

Figure 3: Change in usage of *luka* and *uta* as transitive verbs in *ma pona pi toki pona* from 2020 to 2024 as percentage of total content word usage.



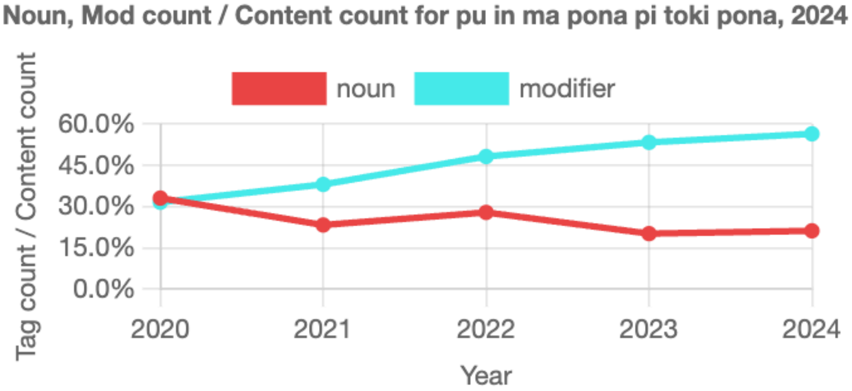
A possible explanation for this change is processing efficiency following Bybee’s (2015, p. 1) argumentation that “mental processes ... are ... main causes of change”: body-part terms can be used as nouns to express similar ideas with the verb *pilin* ’to feel,’ as in *jan li pilin e waso kepeken luka* ’the person feels the bird with their hand.’ However, this construction is longer and more complex than simply using the body-part term as a transitive verb, as shown in (4b). Because the transitive construction is shorter while communicating the same information to the listener, it is likely to be preferred in informal contexts where brevity is valued.

4.1.2 Adjectives as nouns

The adjective *pu* was defined in 2014 to mean “interacting with the official Toki Pona book” (Danielson, 2024a; Lang, 2014). The typical way to refer to the book *Toki Pona: The Language of Good* in the language is with the phrase *lipu pu* ’the book that is the official Toki Pona book.’ The word *pu* can be used in a variety of contexts, such as the sentence *ona li pu* ’they are reading the official Toki Pona book’ and the phrase *nasin pu* ’the ways of the official Toki Pona book.’

Many speakers instead opt to refer to the book simply with *pu* as a standalone noun without *lipu* ’document’ preceding it. However, more recently, this noun usage referring to the book has decreased, with speakers switching back to referring to the book using the full phrase *lipu pu*, as seen in Figure 4.

Figure 4: Change in the use of *pu* as a modifier and noun in *ma pona pi toki pona*.



This is likely due to pressure to conform with existing name constructions in the language, applying known patterns to new contexts and thereby reinforcing these patterns (cf. Bybee, 2015, p. 10). In Toki Pona, names must be adjectives and cannot be nouns: a person named *Tani* is referred to as *jan Tani* ’Tani the person,’ and a country named *Kanata* is referred to as *ma Kanata* ’Kanata the place.’ This construction is similar to the typical way of referring to the book, *lipu pu*. Toki Pona speakers may encounter the construction *lipu pu* enough that they analyze *pu* more as a proper name attached to *lipu* rather than a typical content word. As such, Toki Pona users become more likely to match the form of these other constructions.

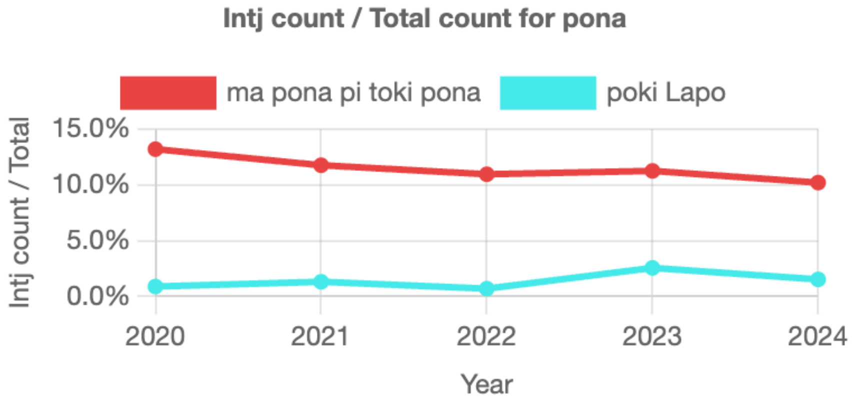
4.2 Variation across corpora

The second research question aimed to examine how the usage of words in different syntactic positions vary between informal and formal contexts.

4.2.1 Use of interjections

In conversational text, speakers will often use words as standalone interjections rather than full sentences. For example, saying *pona* ’good’ alone is often a sign of acknowledgement, like English ’okay.’ Interjections are also used to answer polar questions by either repeating the verb or its negation.

Figure 5: Usage of *pona* as a standalone interjection between the two corpora.



As seen in Figure 5, there is a consistent difference in the use of *pona* as an interjection between the corpora, and the ratio remains relatively stable over time.

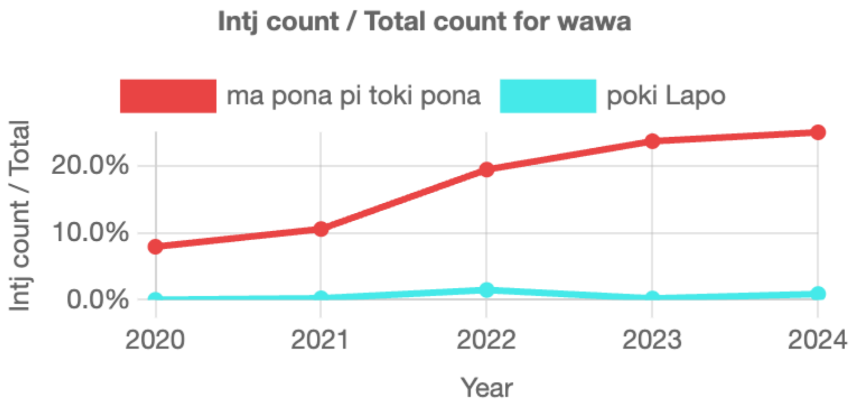
The usage of *wawa* ’strong, powerful’ as an interjection has drastically increased in conversational data as seen in Figure 6. Despite this increase, an increase in usage has not been observed in the formal corpus.

This pattern can be attributed to several factors. First, interjections may inherently appear less in non-conversational data because simpler statements like acknowledgements are less necessary. Second, formal writing and longer prose tends to be more impersonal and disembodied (McCulloch, 2020). In such contexts, interjections that express emotions and exclamations are less likely to be included.

4.2.2 Adoption of features

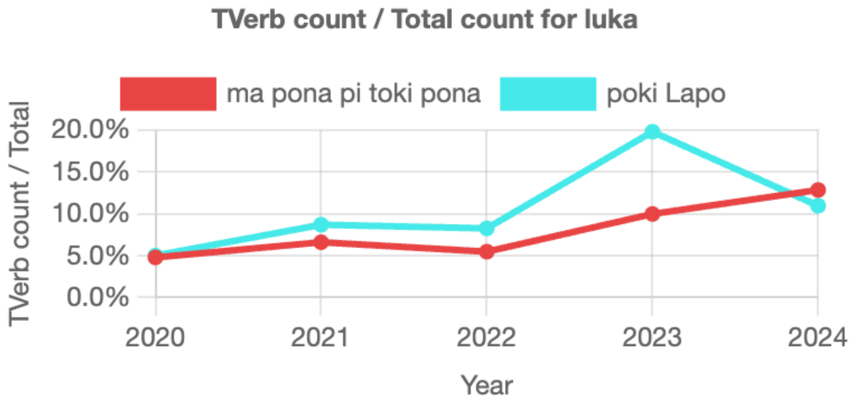
Because language change typically occurs slowly, adoption of linguistic innovations was expected to be delayed in more formally written texts where a more ‘traditional’ grammar might be preferred in order to prioritize accessibility to a

Figure 6: Usage of *wawa* as a standalone interjection between the two corpora.



wider audience. However, the adoption of features is not delayed in the formal corpus. Instances of diachronic variation observed in the informal corpus were observed in the formal corpus at the same time, as in Figure 7.

Figure 7: Usage of *luka* as a transitive verb between the two corpora.



One explanation for the lack of an apparent difference between the two corpora is the homogeneity of the Toki Pona community. A significant majority of authors with published works in Toki Pona are members of the *ma pona pi toki pona* Discord or are heavily connected to it; the language has a tightly knit community of speakers. Additionally, though increases in body parts as transitive verbs and *pu* as a modifier are observed, these changes still fit within the grammar and derivation patterns described in *Toki Pona: The Language of Good*, so writers may feel that these usages are still understandable to a reading audience. Continued repetition then may have led to the adoption of features in Toki Pona in general, formal and informal uses (Bybee, 2015, cf.). Adoption of features may also be understood as an act of identity (Labov, 2010, p. 193). In order to identify with the Toki Pona community, one may choose expressions that reflect its communicative convention, reinforcing the feature in turn.

5 Conclusion

This study used a computational and corpus-based approach to examine language variation over time and across genres of the constructed language Toki Pona. The results of the parsing and part-of-speech tagging analysis revealed that the word classes of certain lexical items are associated with change over time, indicating that changes noticed by speakers of Toki Pona community were true. For instance, nouns depicting body-parts are increasingly used as transitive verbs, and proper nouns may be used more as adjectives.

Although the study contributes valuable insights to the field of language change and variation, the results should be taken with care as the study has a few limitations. The available time period of Toki Pona is relatively short (2020-2025) compared to studies of natural languages. Furthermore, because both corpora contained written data, generalizations about spoken variations cannot be made.

Future research could determine whether semantic features found in natural languages extend to Toki Pona. For instance, Glass (2024) found that in English, pragmatic factors make subjective adjectives more likely predicative and objective adjectives more likely attributive, and Dyer et al. (2020) found that the subjectivity rating of adjectives in a language is a strong predictor of distance from the noun. These methods could be replicated in Toki Pona by scoring subjectivity for Toki Pona's content words and evaluating the prediction power in the same corpora.

To conclude, this study has shown language change and variation in a constructed language, Toki Pona. The linguistic innovations all occur within the existing grammatical framework rather than expanding on it. Toki Pona's fluid word class system allows for unique usage without losing intelligibility. Thus, constructed languages are influenced by cognitive and sociolinguistic factors such as processing efficiency, conforming of known structural patterns to new contexts, and enactment of identity as member of the Toki Pona community just like natural languages (cf. Bybee, 2015; Labov, 2010). The findings suggest that language communities naturally balance innovation and stability, maintaining mutual intelligibility while finding new ways to communicate. Thus, innovation within constraints may be a fundamental property of language use.

References

Bybee, J. (2015). *Language Change*. Cambridge University Press.

Chandra, K., & Radvan, T. (2020, June). *nearley: A parsing toolkit for JavaScript*. https://doi.org/10.5281/zenodo.3897993

Chomsky, N. (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

Comrie, B., Haspelmath, M., & Bickel, B. (2024, January). *Leipzig Glossing Rules*. https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Danielson, G., III. (2024a, March). *When was pu added to toki pona?* Retrieved March 1, 2024, from https://web.archive.org/web/20240301100156/https://mun.la/lipu/pu.html

Danielson, G., III. (2024b, August). *ilo Muni*. https://ilo.muni.la

Danielson, G., III. (2025, April). *sona-toki* [GitHub repository]. https://github.com/gregdan3/sona-toki

Dyer, W., Futrell, R., Liu, Z., & Scontras, G. (2020). *Predicting cross-linguistic adjective order with information gain*. arXiv: 2012.15263 [cs.CL]. https://arxiv.org/abs/2012.15263

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

Glass, L. (2024). The red dress is cute: Why subjective adjectives are more often predicative. *Corpus Linguistics and Lingustic Theory*. https://doi.org/10.1515/cllt-2024-0044

kala Asi, ijo vivi, jan Juwan, & jan Kita. (2025, April). *poki Lapo* [GitHub repository]. https://github.com/kulupu-lapo/poki

Labov, W. (2010). *Principles of linguistic change*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444327496.ch9

Lang, S. (2014). *Toki Pona: The Language of Good*. Tawhid.

ma pona pi toki pona [Discord server]. (2025, January). https://discord.gg/mapona

McCulloch, G. (2020). *Because Internet: Understanding how Language is Changing*. Vintage.

Meulen, S. v. d. (2021, October). *Request for New Language Code Element in ISO 639-3*. https://iso639-3.sil.org/sites/iso639-3/files/change_requests/2021/2021-043_tok.pdf

Peterson, D. J. (2015). *The Art of Language Invention: From Horse-Lords to Dark Elves to Sand Worms, the Words Behind World-Building*. Penguin Books.