The Toki Pona Language: an overview and some hacks

Renato Fabbri
renato.fabbri@gmail.com
University of São Paulo,
Institute of Mathematical and Computer Sciences
São Carlos, SP, Brazil

November 11, 2017

Abstract

1 Introduction

Toki Pona is a minimalist conlang (constructed language) with only 126 words.

1.1 Resources on the web

http://tokipona.net

1.2 Historical note

1.3 Natural and constructed languages

2 Overview of the language

2.1 Phonology

Words in Toki Pona are written using only 14 letters:

- Vowels a (open), e (mid front), o (mid back), i (close front), u (close back).
- Consonants j, k, l, m, n, p, s, t, w:
  - Nasal: m (labial), n (coronal).
  - Plosive: k (dorsal), p (labial), t (coronal).
  - Fricative: s (coronal).

– Approximant: j (dorsal), l (coronal), w (labial).

There are standard guidelines for pronunciation, but the language allows for considerable allophonic variation. For example, /p t k s l/ might be pronounced [p t k s l] or [b d g z ɾ].

Syllables are of the form (C)V(N): an optional consonant, a vowel and an optional nasal consonant. Non word-initial syllables must follow the pattern CV(N). The following sequences are forbidden: ji, wu, wo, ti, mn, nm, mm, nn.

2.2 Syntax

As in other natural languages, colloquial Toki Pona might have incomplete sentences and deviate from the norm. The basic structure of sentences are in the form: ¡subject¿ li ¡predicate¿ e ¡object¿. The li might be repeated to associate more than one predicate to the subject. The particle li is omitted if the subject is a simple mi (I or us) or sina (you). A discussion about problems with this rule and how I deal with them is in Appendix A.

The e might be repeated to associate more than one object to a predicate. Sentences might be related though la, 'sentence' la 'sentence', where the second sentence is the main sentence, and the first sentence is a condition to the first. Multiple la-s are not described in literature, but I assume that one might assume the last sentence being a conditional to the next, except in cases where the context strongly suggests otherwise.

Noun and verb phrases are built with the non-particle words. The first word is the noun and phrase and subsequent words qualify the noun or phrase. The pi particle might be used to separate sequences of words to be evaluated before the relations yield by pi: As pi is often ill understood and used, the following structures might be handy for newbies and as a reference:

- No pi, 'word word word': word ← (qualifies 1) word ← (qualifies 2) word.
- One pi, 'word pi word word': word ← (qualifies 2) [word ← (qualifies 1) word].
- Two pi-s: 'word pi word word word pi word word': word ←5 [word 2 word] 3 word ←4 word 1 word; or: word ←5 [word 1 word] 2 word ←4 word 3 word.

Notes on the usage of pi:

- In a sequence of words, without pi, the second word qualifies the first, the third word qualifies the phrase yield by the first two words, the fourth word qualifies the noun yield by the first three words and so on.
- It is redundant to use pi before the last word in a noun or verb phrase if there is no other pi, reason why it is most often omitted. Its use in this case is regarded as wrong [**?**, **?**], but, as one might notice, it does not introduce any ambiguity.

- The book by jan Pije [10] describes another use for pi: after li to mean possession, e.g. ‘soweli li pi sina’ (your pet). This employment of pi might be regarded as correct, but are promptly written as a noun phrase (e.g. ‘soweli sina’) and is not mentioned by the official book [**?**].

All the words except the structural particles (li, e, la, pi) are usable in noun and verb phrases. Notice that the phrase expresses a noun in a noun phrase (subject or object) or a verb (in the predicate).

At this point, the only missing syntax rule is related to the prepositions: kepeken, lon, sama, tan, tawa. They might appear at the end of noun phrases, should be followed by another noun phrase, and require no particle. E.g. ‘toki tan jan Pije li pana e sona tawa mi’.

Other particles are:

- a or kin, emphasis.
- o, vocative or imperative (’jan lukin sitelen o, li wawa’)
- taso, means however or ’only’ if adjective.
- anu, en: ’or’ and ’and’. Used for nouns in nous phrases. For repetition of verbs, repeat li. For object nouns, repeat e. If the noun is complementing a preposition (tawa, lon), one might repeat the preposition. As Toki Pona is a recent language, and is able to cope with variation due to its simplicity, I would advocate for using en and anu wherever there is no ambiguity.
- nanpa, denotes numbering.
- seme, for questions, used next to the thing being asked for. ’Why?’ might be expressed as ’seme la sina pana e moku lon sewi’.
- mu, for animal noises. For me it is not a particle, as in the official dictionary, but a noun. I also like to use it as a verb: ’mi pakala e luka. mu mute.

The vocabulary specifies morphosyntactic classes: nouns, adjectives, verbs, pre-verbs, adverbs, particles, prepositions, and numbers. I find that they might help the user and newcomer, but it might also suggest a deviation from what I understand and read: the words might be used indistinctly to be the nouns (subjects, the predicate when there is no object, objects, and preposition complements), the adjectives (anything that does not start the noun phrase or follows a pi), verbs (follows mi or sina or li or a preposition), adverbs (follows the verb). The pre-verbs (wile, ken, awen, kama, lukin, sona), might follow a verb, but might also be understood as the verb qualified with the next word, which carries a very similar if not identical meaning. The pre-verbs are all also defined as other morphosyntactic classes, such as adjective, noun, verb. The only exception is wile, which is only a pre-verb.

Thus, the classes given in the dictionary dictate little in practice: jan kala li lape lon ni. Where kala, lape and ni are in this phrase as adjective, verb and noun, and are in the dictionary as noun, adjective and adjective.

As far as I can see, one should regard the particles li, e, pi, and la and punctuation. The other tokens of the vocabulary might be used in any of the remaining positions. Detection:

- Noun: the first word in a noun phrase. After an 'e' and after a pi, The first word in the sentence if sentence does not start by the verb. Might be in the position of the verb if the sentence has no object.
- Adjective: second word on after an e and after Second word on in the subject phrase if present.
- Verb: after a li, mi or sina. If there is no object, the verb position is often a noun.
- After a preposition, there can be a noun phrase, a verb phrase or nothing.
- Notice that there is ambiguity in the structure introduced by the omission of li after mi and sina. Also, when there is no object, a noun or a verb or an adjective might be in the verb position if there is no object. The prepositional complement is also not defined. (mi moku tawa pali, tawa tomo). So, these are sources of syntactic ambiguities in Toki Pona. They might be solved or minimized by using the semantics of the words. One preliminary effort in this direction might be using the classes in the official dictionary to resolve ambiguities whenever possible. This solution is not optimal in correct POS tagging, and does not solve all possible ambiguities (there are words classified as nouns and adjectives, adjectives and verbs, nouns and verbs, particle and verb).

  Another source of ambiguity is the pre-verbs as described in the literature [7, **?**]. But I find it reasonable to understand them as verbs.

Also, the prepositional complement might be a noun or a verb. I could not come up with a sentence where it would be understood as an adjective.

2.3 Further notes

The only synonyms on Toki Pona are: a and kin; lukin and oko; sin or kamako; ale or ali.

In formations such as toki e ni:, wile e ni:, tan ni: etc. '(e) ni' can be omitted and : used alone.

Names are by default transliterated, but might not be, as described in Section A.

2.4 Main references for the language

- The official book is "Toki Pona: The Language of Good" and is authored by Sonja Lang, the creator of the language.
- The book "o kama sona e toki pona!", from jan Pije, is the other main reference for the language [10].

- [11].

3 Analysis and hacks for the language

4 Statistics of the vocabulary

In [**?**] there is statistics about Toki Pona corpus. This section focuses on the statistics of the vocabulary and syntactic rules: the letters, phonemes, word sizes, possible combinations for words and sentences. The statistics are in Appendix [**?**], and the next paragraph is an overview. Python scripts were used to obtain the measurements and are available at [6].

As described in Section [**?**], there is only 14 letters, and phonemes respect a few rules. There are 120 different words in the official vocabulary, 4 of them having synonyms. A total of 124 tokens. Not counting proper nouns (names) and punctuation. They include X nouns, Y verbs, Z adjectives, W prepositions, Y particles, K pre-verbs. Of which one might distinguish only between the particles and the other J words that might be used anywhere a particle will not.

Most often vowels, consonants, consonant vs vowel. Most often phonemes in general an specific positions.. Possible phonemes given the rules. Comparison of the occurring against the possible.

Syntactic structures: Possibilities on noun, verb and prepositional phrases and in sentences. Counting along the number of words.

5 Synthesis of text

The same package [6] has capabilities for synthesizing text. Noun, verb and prepositional phrases, sentences. It also aims at making larger scale texts by keeping a record of the used words and structures (context) and using stylistic outlines for poems and short narratives.

6 Syntax highlighting

The same package [6] has a Vim syntax highlighting plugin for Toki Pona. Instructions for installing and using the syntax highlighting is at [6].

Basically, it distinguishes the words among the morphosyntactic classes according to the official dictionary. As a word often belongs to more than one class, The precedence of them might be set by the user. Also, some classes might be further refined or joined, such as by distinguishing only particles and the rest, or maybe particles and prepositions and the rest. The colors are also promptly changed according to [**?**] and exemplified in the package documentation.

Currently, the Python package synthesizes the syntax file. The user has control of class precedence and merging. The choice of precise coloring schemes might involve hacking the colorscheme being used in Vim (such as 'blue', 'elflord' and 'gruvbox'), and Vim's highlighting schemes as described in [**?**]. In summary,

the usage of the package and plugin might be performed through the following actions:

- Installation of the plugin.
- Tweak of the syntax file by hand.
- Running the Python script to generate a new syntax file according to other settings.
- Write a file inside Vim using Toki Pona and save the file with the .tokipona extension. Reload the highlighting scheme whenever you change the syntax file by hand or through the Python script.
- Access the used highlighted groups with :syntax, Access all the highlighting groups with :so $VIMRUNTIME/syntax/hitest.vim. Change the coloring of a set of terms by associating a used group (e.g. tokiponaADJECTIVE) to an existing group (e.g. Visual): :highlight link tokiponaADJECTIVE Visual. The plugin comes with the :TokiStation command, which opens a window with the files: ijositelen.tokipona, makevimSyntax.py, Highlight Test (created by hitest.vim above), syntax/tokipona.vim. Another tab with shells: python to run over the makeVimSyntax.py and make new syntax/tokipona.vim files. Another with Readme, and PDF documentation. Another with an IPython. Imported tokipona. It resets completely upon command :TokiClose, closing all created windows.

6.1 Hacks from other people

The tokipona.net has a number of tools, just as to transliterate names into Toki Pona phonemes, and search in corpus.

7 Conclusions and further work

- Relate Toki Pona to Wordnet: should one Toki Pona word be related to more then one synset of the English language?
- Understand how the corpus is gathered in tokipona.net.
- Know about previously existing words that were used for Toki Pona (e.g. suno and suwi might come from sun and sweet), and about the reasons that lead Sonja (and maybe other people) to choose the 14 letters and the syllable structure. This might require a dedicated communication with the speaker community and the documentation authors.
- Corpus-based analysis.
- Publication of original texts and translations.

- Make an article written in Toki Pona. I wrote Section **??**, and believe that a summary in both English and Toki Pona for facilitating the acquisition of context, a reasonable article on some scientific topics is possible. I first conceived something around complexity, statistics, physics, or computer science. But one possibility is to write about linguistics, philosophy, literature or psychology with the partners I write in English. I can start a draft, they might learn the language in a few hours (with or without my help), to contribute, and we can write a short paper.

- Enhance the synthesis of text to yield better contextualized text and stylistic traces, such as for poems and short stories. Give the user the ability to choose the sentences (generate randomly according to previously written or given text, some rules input by the user, the package and the language guidelines, outputs to the screen and asks to keep and discard).

- Enhance the syntax highlighting implementation described in Section **??**. It uses only syntactic cues to choose POS tags, which might be overcome by using n-grams and further techniques from Natural Language Processing [**?**].

Acknowledgments

FAPESP (project 2017/05838-3); Vim developers and documentation maintainers; Vim user community.

A My usage of Toki Pona

I use the standard sounds, but often use [z] for s. I often translate texts to Toki Pona (e.g. biblical excepts) and create new texts as poems and short stories. Most of them are in [**?**]. I omit the li particle after subjects sina and mi, in accordance with the norm, but sometimes I use them when there are many predicates. E.g. sina li wawa li pimeja li lukin pona li moku e kasi mute. In such cases, the first li is sometimes omitted. Also, sometimes I use li before mi and sina where I find that there is unwanted ambiguity, e.g. sina moku pona e jan (might be sina li moku pona e jan or sina moku li pona e jan).

Names are by default transliterated, but I advocate that, as in other languages, names might be used as they are in the correspondent mother tongue. E.g. the name Erdös is used in English and Portuguese although the standard alphabet does not contain ö in such languages. I also tend to legitimate the use of English (or German) words in Toki Pona texts if it is the case, as happens often in scientific writing (kernel is a German word used in English, webpage is an English word used in Portuguese).

Proposed notations for numbers seem numerous. I tend to think that one might indicate is two numbers are multiplying (pi) or are in different scales (such as in decimal or binary notations). For example, I take luka two to mean 52. Or it might not be taken to such level of strictidness, for 52 is a reasonable

notation for a simple language. mi jo e jan sama nanpa 12. Or even: ona li lon e soweli 27.

Remove the 'li'. I've been avoiding e ni: and using only :. 've been omitting the subject if it is the same as in the last sentence. I've been avoiding also the li sometimes, and starting only a predicate + object phrase, or a whole phrase altogether.

Table 1: "POS tags incident and chosen. The official dictionary often relates tokens to more than one POS tag. For the text highlighting Plugin, for example, a token has to have an established tag to have a defined color. On the Chosen column, the tokens were regarded only once by choosing the first classification in the dictionary in ['PRE', 'VERB', 'PREPOSITION', 'PARTICLE', 'ADJECTIVE', 'NOUN', 'NUMBER'].

POS	All	Chosen
NOUN	58	49
ADJECTIVE	40	35
VERB	18	13
PARTICLE	12	12
PRE	5	5
PREPOSITION	5	5
NUMBER	4	1
total	142	120

B Final words in Toki Pona

toki li nasin e lawa. li nasin tawa (pi) toki insa.

toki pona li pona e nasin tan ni: ona li jo e sitelen mute lili. ona li pona. pona kepeken weka 'p' li ona, a.

o taso la toki pona li kalama li lukin pona. sitelen en nasin li open e sitelen suli[7, **?**, **?**, **?**, **?**, 11]. li sona. li nasin e toki e sona e lawa e lon.

toki ni li wawa tawa jan mute nasin en toki. wawa tawa toki pi jan sona. pi ilo nanpa en nanpa nasin. taso tawa toki sona a. sitelen sona, sitelen musi.

ilo lon linha 3 li pana e sitelen e sona tan sitelen, en nasin. linha 2 li pana e nasin pi toki pona. e sitelen tawa kama sona.

mi wile pali e sitelen lon toki pona. sitelen lon nasin, sitelen sona, sitelen musi. ante e toki pona la la ona li nasa. taso nasa li pona mute. li pona mute tawa lawa, tawa kama sona, tawa sitelen e toki. ante la toki pona o. toki e toki pona tawa sina. kama sona e toki pona sina.

o pona tawa jan pi toki pona. jan Sonja, Birns-Sprage, Kipo, Pije, Siwejo, Malija, jan kulupu mute.

References

[1] Anthropological physics and social psychology in the critical research of networks. Complex Networks Digital Campus (CS-DC'15). Available at https://youtu.be/oeOKYc3-nbM

[2] Fabbri, R. What are you and I? [anthropological physics fundamentals], 2015. Available at https://www.academia.edu/10356773/What_are_you_and_I_anthropological_physics_fundamentals_

[3] Ouroboros. (2017, November 2). In Wikipedia, The Free Encyclopedia. Retrieved 22:19, November 9, 2017, from https://en.wikipedia.org/w/index.php?title=Ouroboros&oldid=808392809

[4] Fabbri, R. (2017). A reasonable vimrc file. Available at https://raw.githubusercontent.com/ttm/vim/master/vimrc

[5] Ex (text editor). (2017, March 22). In Wikipedia, The Free Encyclopedia. Retrieved 22:22, November 9, 2017, from https://en.wikipedia.org/w/index.php?title=Ex_(text_editor)&oldid=771621020

[6] Fabbri, R. (2017). A Toki Pona Python Package and Vim Syntax Highlighting. Available at https://github.com/ttm/tokipona

[7] Lang, S. (2014). Toki Pona: the language of good. Tawhid Publishing. ISBN-10: 0978292308, ISBN-13: 978-0978292300.

[8] Fabbri, R., Fabbri, R., Vieira, V., Penalva, D., Shiga, D., Mendonça, M., Negrao, A., Zambianchi, L., & Thumé, G. (2013). AA: The Algorithmic Autoregulation (Distributed Software Development) Methodology. RESI. From https://arxiv.org/abs/1604.08255

[9] Fabbri, R. (2017). The Algorithmic-Autoregulation (AA) Methodology and Software: a collective focus on self-transparency. ENMC2017. From https://github.com/ttm/ensaaio/raw/master/emc/article.pdf

[10] Knight, B. (2017) o kama sona e toki pona! Available at: http://tokipona.net/tp/janpije/okamasona.php

[11] Toki Pona community (2017). Wikipesija. Available at: http://tokipona.wikia.com