tomo suli pi kama sona

NIMI PI TOMO NI

tsbohc

TOKI PONA: APPLICATION OF SEMANTIC VECTOR SPACE IN VOCABULARY ANALYSIS

(tenpo ni la, lipu ni li pona ala)

ma pona 2022

CONTENTS

INTRODUCTION

Toki Pona is the second most spoken constructed language in the world. Its core vocabulary consists of only 120-140 words, not including words that are rare and/or considered non-standard by the majority of speakers. Despite the small vocabulary size, Toki Pona can be effectively used to convey a wide range of ideas of varying complexity.

This research aims to perform the semantic analysis and classification of the vocabulary of Toki Pona.

- **Subject.** Semantic analysis and classification of vocabulary.
- **Object.** Toki Pona, a constructed language.
- **Goal.** Perform the semantic analysis and classification of the vocabulary of the language.
  - **Problem.** Vocabulary cannot be analysed based on the available resources as they do not contain sufficient information for said analysis.
  - **Solution.** Use natural language processing techniques to construct a semantic model of the language and base the analysis on it.
- **Methodology.** Distributional semantics and natural language processing, namely language modelling (word embedding).

Objectives

1. Define and classify constructed languages.
2. Describe toki pona, its philosophy, history, and unique features.

3. Define distributional semantics.
4. Define modern approaches to Natural Language Processing applicable to the research.
5. Obtain the necessary corpora.
6. Construct a vector space model of the language.
7. Make observations on the model.
8. Classify the vocabulary based on the observed semantic relationships between the words of the vocabulary.

Relevance

With the rise of the internet, constructed languages now have a place where they can live and thrive. Constructed languages are rapidly gaining popularity. Despite this, the only constructed language that has seen much representation in scientific writing is Esperanto.

The existing dictionaries of Toki Pona can greatly benefit from the findings of this research. This data can also be used as an aid in teaching the language to new speakers.

The Vector Space Model of Toki Pona constructed in the course of this research can find further use in information retrieval, topic modelling, text prediction, sentiment analysis, and many other areas.

DISTRIBUTIONAL SEMANTICS AND CONSTRUCTED LANGUAGES

1 Natural language processing

“Linguistics is concerned not only with language per se, but must also deal with how humans model the world. The study of semantics, for example, must relate language expressions to their meanings, which reside in the mental models possessed by humans. <…> Whereas computational linguistics, as a subfield of linguistics, is concerned with the formal or computational description of rules that languages follow” [Tsujii 2021].

The aim of this research is to bridge the gap between the two disciplines, to use computational linguistics to build a semantic model of a constructed language. This model can then be used to explore the nuances of how humans speak the said language.

In turn, “Natural Language Processing is a field at the intersection of computer science, artificial intelligence, and linguistics” [Vajjala, Majumder 2020, p. 7]. “Natural language processing includes a range of algorithms, tasks, and problems that take human-produced text as an input and produce some useful information, such as labels, semantic representations, and so on, as an output” [Hagiwara 2021, p. 4].

2 Distributional semantics

The core idea behind distributional semantics has roots in American structuralism (Harris) and British lexicology (Firth) and is known as the distributional hypothesis. In its simplest form, it states that “similarity in

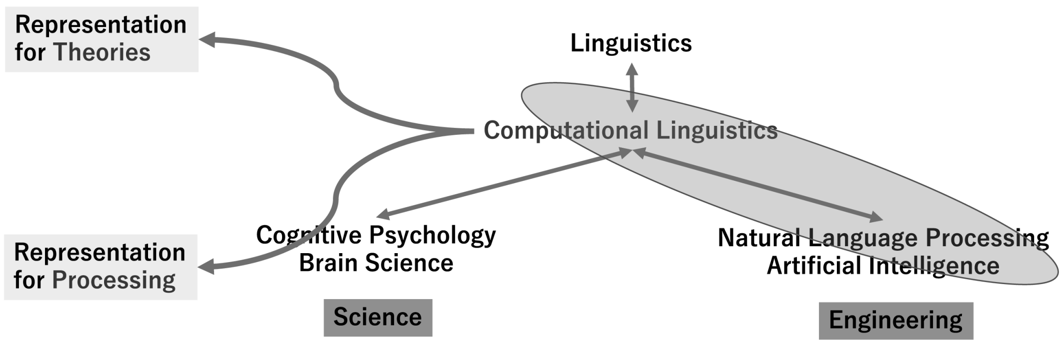

Figure .1: Language-related disciplines [Tsujii 2021]

meaning results in similarity of linguistic distribution" [Harris 1954].

The reverse of this statement is also true. Meaning that "the statistical distribution of linguistic items in context plays a key role in characterizing their semantic behavior" [Lenci 2018]. The aim of distributional semantics is exactly that, to learn the meanings of linguistics units from a corpus of text.

Distributional semantics was popularised by Firth in the 1950s. In a 1957 publication he wrote, "the placing of a text as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognise use. <...> You shall know a word by the company it keeps!" [Firth 1957, p. 11].

The ideas introduced by the distributional hypothesis have received attention in cognitive science [Mcdonald 2008] and language learning [Yarlett, Ramscar, Dye 2008].

Overview

Distributional semantics has become widespread with the adoption of information technology in the field of linguistic research.

Distributional semantics are most frequently applied by taking large amounts of text as input and pushing it through an abstraction algorithm to produce a distributional model as output [Emerson 2020].

Distributional models rely on context to produce semantic representations. That is, distributional models characterise the meanings of words through the context in which they have been observed [Erk 2016].

Planets of the solar system are orbiting the *sun*. The *moon* is orbiting the earth. It's his antique *typewriter* clacking. <...>

$\longrightarrow$ algorithm

	dim1	dim2
sun	0.11023	0.53848
moon	0.21575	0.44034
typewriter	0.52834	0.05389

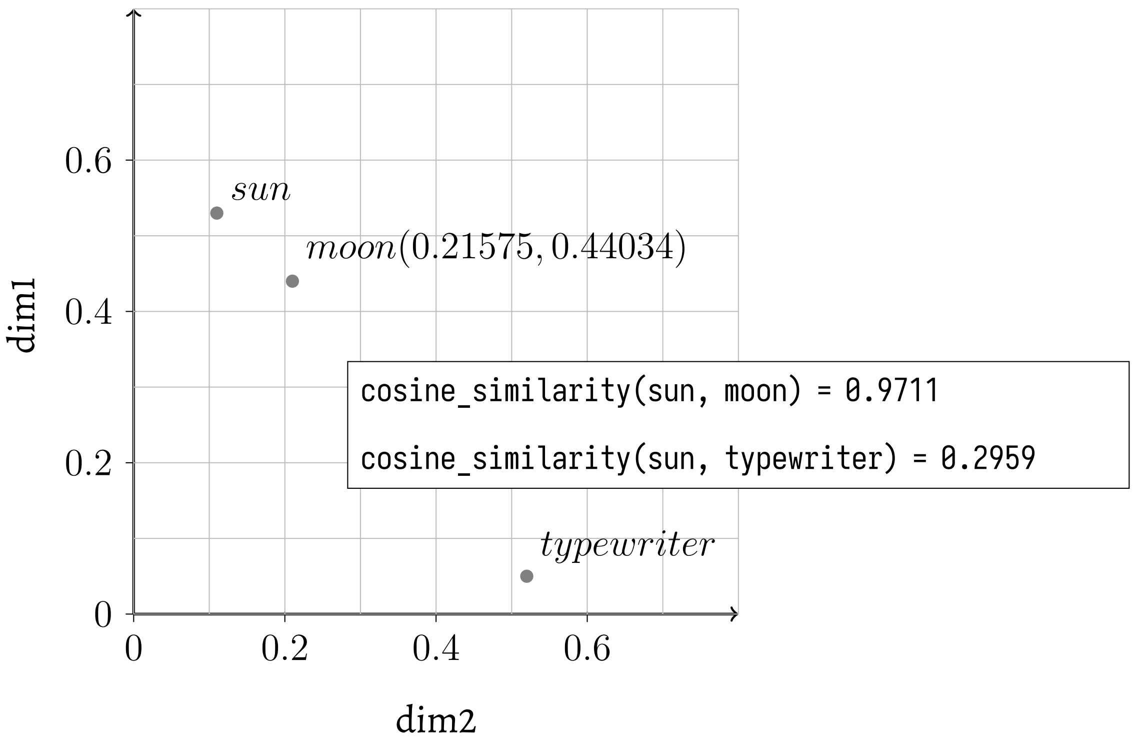

Figure .2: Distributional semantics, an illustrated overview

In a model, the semantic representations are stored in the form of vectors. Vectors are essentially lists of numbers that refer to points in a multi-dimensional space. These vectors are referred to as word vectors.

In the illustrated example, the model only has the dimensionality of two and thus can be mapped onto a two dimensionsional plane without any further processing.

If this is not the case, the multi-dimensionality of the word vector encodings can be reduced to only two or three dimensions. The resulting dimensions can then be used to create a projection of the model which can be observed by the human eye.

All of the approaches to distributional semantics share the quality of learning semantic representations from a corpus in an unsupervised manner. Meaning that it is not required for the corpus to be preproccessed by hand.

2.1 Distributional representations

Distributional representations are mathematic encodings of the distributional properties of words. Typically, in the form of a sequence of numbers. This sequence of numbers can be viewed as a multi-dimensional vector for the purposes of applying to them principles derived from liner algebra.

"Word vectors represent words as multidimensional continuous floating point numbers where semantically similar words are mapped to proximate points in geometric space" [Ahire 2018].

In simpler terms, a word vector is a numerical representation of a word in a corpus relative to every other word in that corpus.

"Vectors have geometrical interpretations: Vectors with n components define points (or arrows) in n-dimensional spaces. Therefore, distributional representations are geometrical representations of the lexicon in the form of a distributional vector space. The positions of lexemes in a distributional semantic space depend on their co-occurrences with linguistic contexts" [Lenci 2018].

2.1.1 Context types

Distributional representations output by a distributional model differ with respect to how the linguistic context is defined.

The contexts can be of the following types [Lenci 2018]:

- **Undirected window-based collocate.** This context type includes words around the current word. No information as to whether the context words precede or follow after the current word is provided to the model. The window size typically ranges from 2 to 10.
- **Directed window-based collocate.** Unlike the previous context type, directed window-based contexts provide the direction in which the context word was seen relative to the current word.
- **Dependency-filtered syntactic collocate.** This context restricted the words which are analysed by the algorithm based on their syntactic roles. This information is hovewer not provided to the model.
- **Dependency-typed syntactic collocate.** This context type provides the previously omitted syntactic type to the model.
- **Text region.** A text region context can represent any text sample that is uniquely identifiable: book chapters, web pages, or simply text portions of any fixed size.

The term window provides a physical analogy to a linguistic context. As the algorithm processes the corpus, the window of the context slides across the text, accounting for the words that can be seen through it.

2.1.2 Semantic similarity metric

The semantic similary between two vectors is primary measured in two ways: using cosine similarity or the Euclidean distance.

The primary advantage of using one of these two methods is that they can be calculated for vectors of any dimensionality.

Euclidean distance

The Euclidean distance between two points is the length of a line segment between the two points. It can also be defined as the shortest distance between two points in an n-dimensional space. For the purposes of calculating the Euclidean distance, the vectors are viewed as point coordinates [Oduntan, Adeyanju 2018].

$$d_{Euc}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

Cosine similarity

Cosine similarity is a measurement of similarity between two sequences of numbers. When calculating cosine similarity, the two sequences of numbers are viewed as vectors. Cosine similarity is equal the cosine of the angle between two vectors, that is, the dot product of the vectors devided by the product of their lengths [Oduntan, Adeyanju 2018].

Cosine similarity always falls into the interval $[-1, 1]$. Two parallel vectors have a cosine similarity of 1, two orthogonal (perpendicular to each other) vectors have a cosine similarity of 0, while two opposite vectors have a cosine similarity of -1.

$$s_{cos}(A, B) := cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

This method was chosen to simplify the process of comparing similarities between vector pairs. Where the Euclidean distance provides an absolute value, the cosine similarity provids a fraction.

2.1.3 Curse of dimensionality

The curse of dimensionality refers to the phenomena that arise when organising data in high-dimensional spaces.

In the context of distributional models, dimensionality is determined by how many word relationships are accounted for by the model.

As the dimensionality of representations increases, the volume of the space they take up increases so fast that the available data becomes sparse. In other words, it becomes hard to make sense of the data as it becomes spread too thinly across the multi-dimensional space [Venkat 2018].

A common solution to this is dimensionality reduction.

2.2 Implementations

2.2.1 Count vector model

The earliest implementatinos of distributional modeles featured simple counting algorithms. “<…> these models just record other words that have been observed in the vicinity of a target word in large text corpora, and form some sort of aggregate over the recorded context items. They then estimate the semantic similarity between words based on contextual similarity" [Erk 2016]. These models are referred to as count models.

"Context items are counted only if they appear close to the target word, that is, if they are within the relevant context" [Erk 2016].

The count models operate on window-based context. The window size is typically narrow (2-4 words). The window can be allowed to cross the boundaries of sentences or not [Baroni, Dinu, Kruszewski 2014].

2.2.2 Neural probabilistic language model

In the recent years, the distributional model architecture has seen as notable shift to machine learning algorithms. With the improvements of hardware performance, the training of complex neural networks on corpora of larger sizes has become possible.

The earlier machine learning based models were plagued by the curse of dimensionality. This problem was solved in the model proposed by [Bengio, Ducharme, Vincent 2000]. The proposed neural network model learns distributional representations and the generalisation function at the same time. The generalisation function is based on the estimates of probablity of a word appearing in the given context.

The architecture of this model "consists of input, projection, hidden and output layers. At the input layer, N previous words are encoded using 1-of-V coding, where V is size of the vocabulary. The input layer is then projected to a projection layer P that has dimensionality $N \times D$, using a shared projection matrix. As only N inputs are active at any given time, composition of the projection layer is a relatively cheap operation" [Mikolov, Chen, Corrado 2013].

The training complexity of this model is

$$Q = N \times D + N \times D \times H + H \times V$$

where H is the size of the hidden layer.

2.2.3 Recurrent neural net language model

This model artcitecture contains recurrent neural networks, meaning that as the model learns from the input, it produces output that is fed back into the model as input. The recurrent matrix connects hidden layers to itself using time-delayed connections. "This allows the recurrent model to form some kind of short term memory, as information from the past can be represented by the hidden layer state that gets updated based on the current input and the state of the hidden layer in the previous time step" [Mikolov, Chen, Corrado 2013].

This model architecture consists of only input, hidden, and output layers, thus allowing for a reduction of complexity when compared to the neural probabilistic language model [Mikolov, Chen, Corrado 2013].

The training complexity of this model is

$$Q = H \times H + H \times V$$

2.2.4 Continuous bag-of-words model

The first architecture proposed by Mikolov removes the non-linear hidden layer, further reducing complexity. The projection layer is shared for all words [Mikolov, Chen, Corrado 2013].

The output of this model is not ifluenced by history, like in the previous model. In a continuous bag-of-words model not only the words preceding the current word are used for context, but also the words that follow it.

This model attempts to predict the current word from the sum of the context vectors. This sum of vectors is referred to as a "bag of words", giving the name to the model. If the prediction of the word is correct after comparing it with the current word, its distributional representation is reinforced. If the prediction is wrong, the distributional representation is corrected.

The training complexity of this model is

$$Q = N \times D + D \times \log_2 V$$

2.2.5 Continuous skip-gram model

3 Constructed languages

3.1 The notion of a constructed language

3.2 Classification

LANGUAGE MODELLING AND TOKI PONA

1 Toki Pona

1.1 History

1.2 Phonology

1.3 Grammar

1.4 Vocabulary

2 Vector space model

2.1 Corpus acquisition

2.2 Text normalisation

2.3 Model construction

2.4 Projection and visualisation

2.5 Observations

REFERENCES

1. *Ahire J. B.* Introduction to Word Vectors // Retrieved March. — 2018. — Vol. 12. — P. 2018.
2. *Baroni M.*, *Dinu G.*, *Kruszewski G.* Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors //. Vol. 1. — 06/2014. — P. 238–247. — DOI: 10.3115/v1/P14-1023.
3. *Bengio Y.*, *Ducharme R.*, *Vincent P.* A Neural Probabilistic Language Model //. Vol. 3. — 01/2000. — P. 932–938. — DOI: 10.1162/153244303322533223.
4. *Emerson G.* What are the Goals of Distributional Semantics? — 2020. — DOI: 10.48550/ARXIV.2005.02982. — URL: https://arxiv.org/abs/2005.02982.
5. *Erk K.* What do you know about an alligator when you know the company it keeps // Semantics and Pragmatics. — 2016. — Vol. 9.
6. *Firth J.* A Synopsis of Linguistic Theory, 1930-1955. — 1957. — URL: https://books.google.ru/books?id=T8LDtgAACAAJ.
7. *Hagiwara M.* Real-World Natural Language Processing: Practical Applications with Deep Learning. — Manning, 2021. — ISBN 9781617296420. — URL: https://books.google.ru/books?id=Ok5NEAAAQBAJ.
8. *Harris Z. S.* Distributional Structure // WORD. — 1954. — Vol. 10, no. 2/3. — P. 146–162. — DOI: 10.1080/00437956.1954.11659520. — eprint: https://doi.org/10.1080/00437956.1954.11659520. — URL: https://doi.org/10.1080/00437956.1954.11659520.
9. *Lang S.* Toki Pona: The Language of Good. — Sonja Lang, 2014. — ISBN 9780978292300. — URL: https://books.google.ru/books?id=5P0ZjwEACAAJ.

10. *Lenci A.* Distributional Models of Word Meaning // Annual Review of Linguistics. — 2018. — Feb. — Vol. 4. — DOI: 10.1146/annurev-linguistics-030514-125254.

11. *Mcdonald S.* Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. — 2008. — July.

12. *Mikolov T.*, *Chen K.*, *Corrado G.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. — 2013. — Jan. — Vol. 2013.

13. *Oduntan O.*, *Adeyanju O.* A Comparative Analysis of Euclidean Distance and Cosine Similarity Measure for Automated Essay-Type Grading // Journal of Engineering and Applied Sciences. — 2018. — July. — Vol. 13. — P. 4198–4204. — DOI: 10.3923/jeasci.2018.4198.4204.

14. *Tsujii J.* Natural Language Processing and Computational Linguistics // Computational Linguistics. — 2021. — Dec. — Vol. 47, no. 4. — P. 707–727. — ISSN 0891-2017. — DOI: 10.1162/coli_a_00420. — eprint: https://direct.mit.edu/coli/article-pdf/47/4/707/1979478/coli_a_00420.pdf. — URL: https://doi.org/10.1162/coli%5C_a%5C_00420.

15. *Vajjala S.*, *Majumder B.* Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. — O'Reilly Media, 2020. — ISBN 9781492054009. — URL: https://books.google.ru/books?id=hPrrDwAAQBAJ.

16. *Venkat N.* The Curse of Dimensionality: Inside Out. — 09/2018. — DOI: 10.13140/RG.2.2.29631.36006.

17. *Yarlett D.*, *Ramscar M.*, *Dye M.* Language Learning Through Similarity-Based Generalization. — 2008. — Jan.

SUPPLEMENTARY MARTERIAL

1 Vector space model

The two-dimensional projection of the semantic model constructed as a result of this reseach.

2 Dictionaries

2.1 nimi pu

The dictionary of Toki Pona as it appears in Toki Pona: The Language of Good [Lang 2014, p. 125–134]. This dictionary is licensed under public domain.

Word		Definition
a or kin	PARTICLE	(emphasis, emotion or confirmation)
akesi	NOUN	non-cute animal; reptile, amphibian
ala	ADJECTIVE	no, not, zero
alasa	VERB	to hunt, forage
ale or ali	ADJECTIVE	all; abundant, countless, bountiful, every, plentiful
	NOUN	abundance, everything, life, universe
	NUMBER	100
anpa	ADJECTIVE	bowing down, downward, humble, lowly, dependent
ante	ADJECTIVE	different, altered, changed, other
anu	PARTICLE	or
awen	ADJECTIVE	enduring, kept, protected, safe, waiting, staying

	PRE-VERB	to continue to
e	PARTICLE	(before the direct object)
en	PARTICLE	(between multiple subjects)
esun	NOUN	market, shop, fair, bazaar, business transaction
ijo	NOUN	thing, phenomenon, object, matter
ike	ADJECTIVE	bad, negative; non-essential, irrelevant
ilo	NOUN	tool, implement, machine, device
insa	NOUN	centre, content, inside, between; internal organ, stomach
jaki	ADJECTIVE	disgusting, obscene, sickly, toxic, unclean, unsanitary
jan	NOUN	human being, person, somebody
jelo	ADJECTIVE	yellow, yellowish
jo	VERB	to have, carry, contain, hold
kala	NOUN	fish, marine animal, sea creature
kalama	VERB	to produce a sound; recite, utter aloud
kama	ADJECTIVE	arriving, coming, future, summoned
	PRE-VERB	to become, manage to, succeed in
kasi	NOUN	plant, vegetation; herb, leaf
ken	PRE-VERB	to be able to, be allowed to, can, may
	ADJECTIVE	possible
kepeken	PREPOSITION	to use, with, by means of
kili	NOUN	fruit, vegetable, mushroom
kiwen	NOUN	hard object, metal, rock, stone
ko	NOUN	clay, clinging form, dough, semi-solid, paste, powder

kon	NOUN	air, breath; essence, spirit; hidden reality, unseen agent
kule	ADJECTIVE	colourful, pigmented, painted
kulupu	NOUN	community, company, group, nation, society, tribe
kute	NOUN	ear
	VERB	to hear, listen; pay attention to, obey
la	PARTICLE	(between the context phrase and the main sentence)
lape	ADJECTIVE	sleeping, resting
laso	ADJECTIVE	blue, green
lawa	NOUN	head, mind
	VERB	to control, direct, guide, lead, own, plan, regulate, rule
len	NOUN	cloth, clothing, fabric, textile; cover, layer of privacy
lete	ADJECTIVE	cold, cool; uncooked, raw
li	PARTICLE	(between any subject except mi alone or sina alone and its verb; also to introduce a new verb for the same subject)
lili	ADJECTIVE	little, small, short; few; a bit; young
linja	NOUN	long and flexible thing; cord, hair, rope, thread, yarn
lipu	NOUN	flat object; book, document, card, paper, record, website
loje	ADJECTIVE	red, reddish
lon	PREPOSITION	located at, present at, real, true, existing
luka	NOUN	arm, hand, tactile organ

	NUMBER	five
lukin or oko	NOUN	eye
	VERB	to look at, see, examine, observe, read, watch
	PRE-VERB	to seek, look for, try to
lupa	NOUN	door, hole, orifice, window
ma	NOUN	earth, land; outdoors, world; country, territory; soil
mama	NOUN	parent, ancestor; creator, originator; caretaker, sustainer
mani	NOUN	money, cash, savings, wealth; large domesticated animal
meli	NOUN	woman, female, feminine person; wife
mi	NOUN	I, me, we, us
mije	NOUN	man, male, masculine person; husband
moku	VERB	to eat, drink, consume, swallow, ingest
moli	ADJECTIVE	dead, dying
monsi	NOUN	back, behind, rear
mu	PARTICLE	(animal noise or communication)
mun	NOUN	moon, night sky object, star
musi	ADJECTIVE	artistic, entertaining, frivolous, playful, recreational
mute	ADJECTIVE	many, a lot, more, much, several, very
	NOUN	quantity
nanpa	PARTICLE	-th (ordinal number)
	NOUN	numbers
nasa	ADJECTIVE	unusual, strange; foolish, crazy; drunk, intoxicated
nasin	NOUN	way, custom, doctrine, method, path, road

nena	NOUN	bump, button, hill, mountain, nose, protuberance
ni	ADJECTIVE	that, this
nimi	NOUN	name, word
noka	NOUN	foot, leg, organ of locomotion; bottom, lower part
o	PARTICLE	hey! O! (vocative or imperative)
olin	VERB	to love, have compassion for, respect, show affection to
ona	NOUN	he, she, it, they
open	VERB	to begin, start; open; turn on
pakala	ADJECTIVE	botched, broken, damaged, harmed, messed up
pali	VERB	to do, take action on, work on; build, make, prepare
palisa	NOUN	long hard thing; branch, rod, stick
pan	NOUN	cereal, grain; barley, corn, oat, rice, wheat; bread, pasta
pana	VERB	to give, send, emit, provide, put, release
pi	PARTICLE	of
pilin	NOUN	heart (physical or emotional)
	ADJECTIVE	feeling (an emotion, a direct experience)
pimeja	ADJECTIVE	black, dark, unlit
pini	ADJECTIVE	ago, completed, ended, finished, past
pipi	NOUN	bug, insect, ant, spider
poka	NOUN	hip, side; next to, nearby, vicinity
poki	NOUN	container, bag, bowl, box, cup, cupboard, drawer, vessel

pona	ADJECTIVE	good, positive, useful; friendly, peaceful; simple
pu	ADJECTIVE	interacting with the official Toki Pona book
sama	ADJECTIVE	same, similar; each other; sibling, peer, fellow
	PREPOSITION	as, like
seli	ADJECTIVE	fire; cooking element, chemical reaction, heat source
selo	NOUN	outer form, outer layer; bark, peel, shell, skin; boundary
seme	PARTICLE	what? which?
sewi	NOUN	area above, highest part, something elevated
	ADJECTIVE	awe-inspiring, divine, sacred, supernatural
sijelo	NOUN	body (of person or animal), physical state, torso
sike	NOUN	round or circular thing; ball, circle, cycle, sphere, wheel
	ADJECTIVE	of one year
sin or namako	ADJECTIVE	new, fresh; additional, another, extra
sina	NOUN	you
sinpin	NOUN	face, foremost, front, wall
sitelen	NOUN	image, picture, representation, symbol, mark, writing
sona	VERB	to know, be skilled in, be wise about, have information on
	PRE-VERB	to know how to
soweli	NOUN	animal, beast, land mammal
suli	ADJECTIVE	big, heavy, large, long, tall; important; adult

suno	NOUN	sun; light, brightness, glow, radiance, shine; light source
supa	NOUN	horizontal surface, thing to put or rest something on
suwi	ADJECTIVE	sweet, fragrant; cute, innocent, adorable
tan	PREPOSITION	by, from, because of
taso	PARTICLE	but, however
	ADJECTIVE	only
tawa	PREPOSITION	going to, toward; for; from the perspective of
	ADJECTIVE	moving
telo	NOUN	water, liquid, fluid, wet substance; beverage
tenpo	NOUN	time, duration, moment, occasion, period, situation
toki	VERB	to communicate, say, speak, say, talk, use language, think
tomo	NOUN	indoor space; building, home, house, room
tu	NUMBER	two
unpa	VERB	to have sexual or marital relations with
uta	NOUN	mouth, lips, oral cavity, jaw
utala	VERB	to battle, challenge, compete against, struggle against
walo	ADJECTIVE	white, whitish; light-coloured, pale
wan	ADJECTIVE	unique, united
	NUMBER	one
waso	NOUN	bird, flying creature, winged animal
wawa	ADJECTIVE	strong, powerful; confident, sure; energetic, intense
weka	ADJECTIVE	absent, away, ignored

wile	PRE-VERB	must, need, require, should, want, wish