tomo suli pi kama sona

NIMI PI TOMO NI

tsbohc

TOKI PONA: DISTRIBUTIONAL APPROACH TO SEMANTIC ANALYSIS OF A CONSTRUCTED LANGUAGE

tenpo ni la, lipu ni li pona ala

(spelling and general unenglishness will be corrected at a later date)

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 used to convey a wide range of ideas of varying complexity [Meulen 2021].

Problem

The publicly available dictionaries of Toki Pona are primarily based on the original official dictionary [Lang 2014] and do not fully reflect how the language is spoken today.

Goals

The primary goal of this paper is to comment on the semantics of the core Toki Pona vocabulary, as well as to organise individual words into groups based on their semantic relatedness and the context they are most prevalently used in.

The secondary goal is to discuss the semantics of the core vocabulary of Toki Pona — as it is spoken by the majority of the communety — in relation to the first official dictionary of the language.

- **Subject.** Semantic analysis and classification of vocabulary.
- **Object.** Toki Pona, a constructed language.
- **Methodology.** Distributional semantics and natural language processing, namely language modelling (word embedding).

Objectives

1. Define natural language processing and distributional semantics, discuss modern implementations as well as other concepts.
2. Define and classify constructed languages.
3. Describe Toki Pona, its philosophy, history, and unique features.
4. Obtain the necessary corpora.
5. Construct a vector space model of the language.
6. Make observations on the model.
7. Classify the words of the vocabulary based on the observed semantic relationships between them.

Relevance

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 or other resources concerned with teaching Toki Pona to new speakers could benefit from the findings of this research. New tools can be developed which will aid newcomers to the language.

The vector space model of Toki Pona developed in the course of this research can find further use in machine translation, 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

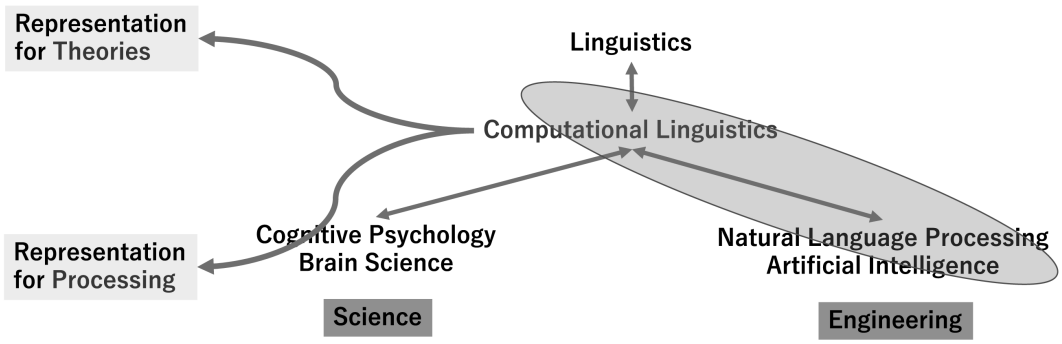

Figure .1: Language-related disciplines [Tsujii 2021]

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. <...>

⟶ algorithm

	dim1	dim2
sun	0.11023	0.53848
moon	0.21575	0.44034
typewriter	0.52834	0.05389

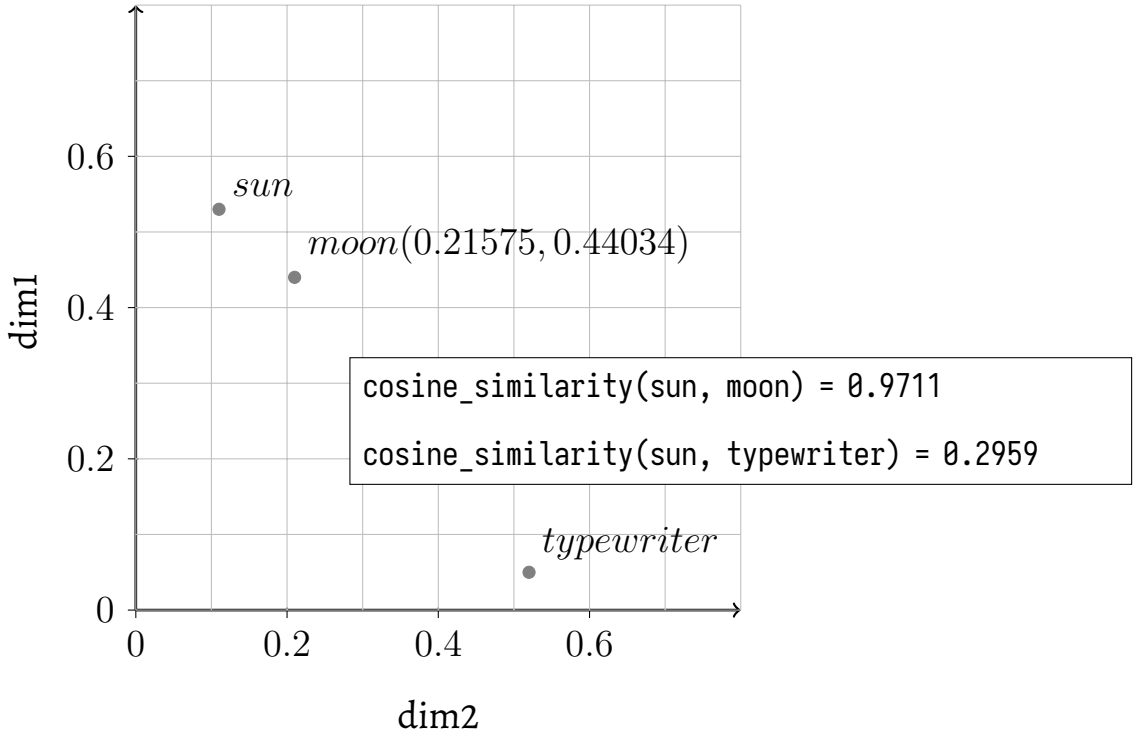

Figure .2: Distributional semantics, an illustrated overview

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, without the involvement of humans.

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 Notable implementations

2.2.1 Count vector model

The simplest implementatinos of distributional modeles feature 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 of Word2vec 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 continuous bag-of-words model is not ifluenced by history like the previous one. 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

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

Because this model architecture produces the prediction as output, the learned weights of the hidden layer is what represents the word vectors.

2.2.5 Continuous skip-gram model

The second architecture of Word2vec proposed by Mikolov has the opposite objective of the continuous bag-of-words model. The continuous skip-gram model predicts the surrounding context from the current word. Similar to the continuous bag-of-words model, when the continuous skip-gram model succeedes in predicting the context words, the semantic representation of the current word is reinforced. When it fails, it is corrected [Mikolov, Chen, Corrado 2013].

The training complexity of this model is

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

While the complexity of this model is greater, the accuracy is also much greater [Mikolov, Chen, Corrado 2013]. Similar to a continuous bag-of-words model, the weights of the hidden layer are the distributional representations.

2.3 Applications

The data provided by the distributional model can be used directly to analyse the semantics of a language:

1. **Semantic similarity.** By definition, distributional models provide data that quantifies semantic relatedness between individual words or expressions. This data can be interpreted by humans to draw conclusions about the meanings of words or used in other areas of natural language processing.

2. **Word clustering.** Semantic representations tend to form groups in the multi-dimensional space. Word clustering refers to the ways and means by which these groups can be extracted as formal clusters [Bekkerman, El-Yaniv, Tishby 2003].

3. **Automatic creation of thesauri.** The semantic similarity data can be further processed to produce lists of homonyms, synonyms, or even antonyms [Henestroza Anguiano, Denis 2011].

4. **Word sense disambiguation.** This refers to a problem in computational linguitics that is conserned with identifying which sense of a word is used in a particular sentence [Musto, Narducci, Basile 2011].

5. **Information retrieval.** Distributional models can be used to access semantically similar words to those of a query, expanding the retrieved results from exact word matching to semantically fuzzy matching [Silva, Maia 2019].

6. **Data mining.** In data mining, namely text mining, distributional models can provide means of identifying similar documents, thus narrowing the scope of a search [Dalianis 2018, p. 89].

7. **Paraphrasing.** The data provided by distributional models can supply paraphrasing algorithms with vocabulary, or aid in judging the relative semantic similarity between two paraphrases on a sentence level basis [El Desouki, Gomaa 2019].

8. **Sentiment analysis.** Given a small list of words manually tagged with emotive potentials, distributional models can propagate these potentials through a corpus based on the semantic similarity between the tagged words [Alshari, Azman, Doraisamy 2017].

3 Artificial languages

The term 'artificial' is used to broadly refer to all languages that have been created through deliberate and conscious planning [Stria 2016, p. 41]. Under the broad umbreall of the term 'artificial' reside predicate calculus and programming languages such as Lisp and C++.

"In the interlinguistic literature the term 'artificial' as opposed to 'natural' is regarded as 'crudely misleading' (Schubert 1989) because it suggests that languages created to facilitate international communication are in fact identical to machine or formulaic languages [Stria 2016, p. 45]".

In an effort to avoid confusion with artifical languages of technical nature, the currently popular term 'constructed' will be used throughout the rest of this paper.

It should be noted that the abbreviation 'conlang' is also widely spread among the members of the community surrounding constructed languages.

3.1 The notion of a constructed language

Constructed languages are languages that have been purposely created to be similar or comparable in function to natural languages. The purpose behind the creation of constructed languages tends to vary greatly, ranging from the aim to create an auxiliary means of international communication, to artistic and philosophical expression.

3.2 Notable constructed languages throughout history

"The dream of a perfect language did not only obsess European culture. The story of the confusion of tongues, and of the attempt to redeem its loss through the rediscovery or invention of a language common to all humanity, can be found in every culture [Eco 1995, p. 1]".

Antiquity

The concept of a constructed language was first mentioned in writing by Athenaeus of Naucratis in his *Deipnosophistae* (*circa* AD 230). The languages he presented were not full languages, but languages of rudimentary type known as a naming language. These languages were collections of neologisms that could be used to replace existing vocabulary or to refer to things that otherwise had no name. Athenaeus further writes of other people who invented their own words [Sanders 2020].

Linguistic mysticism

Irish myths of the seventh century described the origin of Gaelic. According to *Auraicept na n-Éces*, King Fénius Farsaid of Scythia traveled to the Tower of Babel after God fragmented human language. King Fénius and his many scholars studied the remains of the human language and combined its best fragments into a new and more perfect language, Gaelic [Williams 2016].

The earliest attempt at language creation was Lingua Ignota by Hildegard of Bingen, a German Benedictine abbess and polymath, in the eleventh century. The underlying structure of the language was Latin, but the spelling was significantly altered. "She did compose one macaronic antiphon, "O orzchis Ecclesia," in which Latin and Lingua Ignota vocabulary alternate <...> Alternatively, if the Lingua had a second use as a secret language (possibly in the presence of outsiders) for Hildegard and her nuns, as some have suggested,

this reviewer submits that verbs are not always needed for the achievement of communication: “Enpholianz warinz nascutil” (bishop / wart / nose) provides, if not exactly a sentence, an entirely understandable lexical string [Straubhaar 2008]”.

“Hildegard also created Litteraæ Ignotæ ’unknown letters’, a constructed writing system or *neography*, which she used to represent Lingua Ignota [Sanders 2020]”. Despite the lack in grammatical degth by comparison to modern constructed languages, Lingua Ignota is widely praised as the first constructed language ever created.

From the belief in the connection between language creation and the divine arose the notion that languages inherited by humans from gods had become corrupted and now were causing confusion among the scholars and philosophers, impeding scientific progress. At the same time, by the 1600s, “Latin, which was the sole language of education and the only language taught, experienced a decline, and there was no other language to replace it anytime soon [Stria 2016, p. 51]”. This caused a search for a new language to replace Latin, and the subsequent creation of many philosopical constructed languages. These languages were centered around the idea of providing a top-down categorical view of the universe.

In his essay, Wilkins proposed one such universal philosopical language to replace Latin. This language was to be unambiguous and to encompass every concept in the universe, to overcome the curse of Babel [Wilkins 1668, ch. 2, p. 1].

Wilkins constructed a table of 40 major genera, which he then divided further into 251 characteristic differences. From them he derived 2030 species, which appear in pairs. For example, “starting from the major genus of Beasts, after having divided them into viviparous and oviparous, and after having subdivided the viviparous ones into whole footed, cloven footed and clawed, Wilkins arrives at the species Dog/Wolf [Eco 1995, p. 239]”.

Despite the initial acclaim and the interest it received from the king, the language soon fell into obscurity [Okrent 2010, p. 25]. This style of language creation continued into the seventeenth century and then was abandoned.

Early artistic languages

The sixteenth century also saw the beginning of artistic constructed languages. These languages "are designed to suit a creative goal, usually as flavorful adornment in a work of fiction [Sanders 2020]".

The prominent example of one of the earliest artistic constructed languages is the language of the fictional country Utopia (itself an invented word) from the 1516 novel by Sir Thomas More. The language was more than a naming language, though still mostly a relexification of Latin. It appears in the book only as a few isolated words in the text, as well as a four-line poem in the addendum written by More's friend Peter Giles [Sanders 2020].

The main focus constructed languages of this period was vocabulary, giving rise to many naming languages.

Early modern constructed languages

After a considerable decline of interest in philosophical languages, the efforts shifted towards a search for an idea auxiliarly language, which could be used as a lingua franca for people of different backgrounds.

The primary aim of the auxiliary constructed languages of this period was to become an international means of communication. Such languages are often referred to as international auxiliary languages.

One of the first fully fledged auxiliary languages of this time was Jean Pirro's Univeralglot (1868), it is notable for incorporating linguistic features from multiple other languages. The other notable language was Johann Martin Schleyer's Volapük, the most successful auxiliary language until it was

surpassed by Esperanto [Sanders 2020].

Esperanto was created by Ludwik Lejzer Zamenhof in 1887 and remains the most widely spoken auxiliary constructed language. It still hovewer, fell short of Zamenhof's expectations.

"Volapük and Esperanto spurred the creation of many auxiliary languages, especially by those who sought to improve upon previous auxlangs. The first offshoot of Esperanto was Jacob Braakman's Mundolinco (1888), but the most successful was Ido, the result of a battle amongo Esperanto enthusiasts over whether Esperanto should be, or could even be allowed to be, improved [Sanders 2020]". The modern critiques of Esperanto as an international auxiliary language note the irregularities of its grammar and prominent influences of Slavic languages on its design.

The artistic constructed languages of this time saw an increase in sophistication. The focus for still mainly on vocabulary, but the quantity and quality of the vocabulary have noticeably increased.

J. R. R. Tolkien

Tolkien saw language invention and myth-making as two interconnected forms of art. In a 1955 letter, Tolkien regarded his work as "*fundamentally* linguistic in inspiration [J. Tolkien, Carpenter, C. Tolkien 2000, p. 233]".

In a 1958 letter to his son, Tolkien wrote about *The Lord of the Rings*: "Nobody believes me when I say that my long book is an attempt to create a world in which a form of language agreeable to my personal aesthetic might seem real. But it is true [J. Tolkien, Carpenter, C. Tolkien 2000, p. 285]".

In a draft of a letter from 1967, Tolkien summed up his language invention: "It must be emphasized that this process of invention was/is a private enterprise undertaken to give pleasure to myself by giving expression to my personal linguistic 'aesthetic' or taste and its fluctuations [J. Tolkien,

Carpenter, C. Tolkien 2000, p. 411]".

The constructed languages of the time were reflected their intended practical use. Artistic languages were akin to flourishes added to a larger work of fiction, while auxiliary languages were conceived to support the full range of human communication. "Tolkien bridged the gap between these two extremes by creating fully formed languages, but without any larger functionality or purpose beyond the sheer intellectual you of doing so. However, he believed that his 'secret vice' would not be taken seriously on its own, so he wrote his Middle-earth novels as a way to showcase them. Thus, while other writers created conlangs for their fiction, Tolkien created fiction for his conlangs [Sanders 2020]".

Tolkien first revealed his love for language invention to the public in a 1931 essay entitled 'A secret vice'. The essay concludes with poems in Qenya and a fragment in Noldorin (which later became known as Quenya and Sildarin), [J. R. R. Tolkien, C. Tolkien 1983].

Tolkien was a prolific language inventor. In his time, he created at least fifteen languages and dialects. His work did not stop merely at language creation but involved establishing both the detailed histories and the intricate interconnections between the languages of the Middle-earth. To an extent that Tolkien's 'Tree of Tongues' was meant to reproduce a Indo-European genealogical tree model [Fimi 2009, p. 101].

It is worth mentioning that within the lore of Middle-earth, the Black Speech was created by Sauron in mockery of the Elvish languages [J. Tolkien, Fimi, Higgins 2019, p. 20], "to unite the forces of Morder, making it one of the most notable examples of a conlang designed to be understood within its associated fictional setting as an actual conlang [Sanders 2020]," rather than a natural language.

Modern constructed languages

Tolkien's work shaped the further history of constructed languages. Long viewed as a mere pasttime, language invention has become a subject of academic study within linguistics.

From the 70s and well into the 2000s, the world saw the airing of many television series that featured constructed languages. Some of the most notable of them are *Star Trek III: The Search for Spock* (1984, featuring Klingon) and *Game of Thrones* (2011, featuring Dothraki).

"Aided by the expansion of the Internet in the 1990s, and especially the cration of the Conlang email list in 1991, modern conlangers have developed robust community for exchanging ideas, critiques, and tools, allowing them to develop increasingly sophisticated and experimental conlangs [Sanders 2020]".

3.3 Classification

3.3.1 Traditional: structure and source material

The classification proposed by Couturat and Leau groups constructed languages based on their relationship with source material or a lack there of [Couturat, Leau 2014]:

- **A priori.** Constructed languages that are not based on the elements of natural languages. "A priori languages start from scratch with new symbols, signs or other elements devised to represent essential concepts [Lo Bianco 2004]".
- **A posteriori.** Constructed languages are based on the elements of natural languages. "A posteriori languages draw their building blocks from existing languages [Lo Bianco 2004]".
- **Mixed.** A combination of the two.

Several linguists have adopted this classification, most notably Janton, who provided several additions [Janton, Tonkin, Edwards 1993, p. 5]:

1. **A priori.** Metalanguages, schematic languages. These constructed languages are "characterized by largely artificial, nonethnic word roots, schematic derivation, and fixed word categories (i.e., philosophical languages) [Janton, Tonkin, Edwards 1993, p. 6]".

2. **A posteriori.** Naturalistic languages, pseudolanguages. "These languages consciously imitate, in varying degrees, natural languages [Janton, Tonkin, Edwards 1993, p. 5]".

    - **Minimal languages.** Simplified natural languages, living or dead.
    - **Mixed languages.** Languages that use natural and non-natural roots.
        - Languages that were schematically derived with natural word roots in distorted form (Volapük, 1880) or with both artistic and natural word roots (Perio, 1904).
        - Languages that are partly schematic and partly naturalistic. Natural word roots in this group are rarely distorted (Esperanto, 1887).
    - **Naturalistic languages.**
        - Languages with some schematic traits (Unial, 1903; Novial, 1928)
        - Languages with natural derivation (Interlingua, 1940s).

This classification presents a scale of artificiality, with languages derived from natural languages on one end and deliberately designed languages on the other. It is also primarily concerned with morphological and syntactic natures of derivation, not the intention with which the language was created.

3.3.2 Traditional: purpose

"Another type of classification categorises artificial languages according to the purpose of creation [Stria 2016, p. 93]". Kennaway provides the following division: universal languages (from a strive for perfection), international languages, languages of fiction, languages as recreation [Kennaway 2013].

The primary concern with this way of classifying constructed languages this way lies in the fact that purpose is rarely binary. The prominent example are the languages which were constructed as recreation and later served as a tool in fictional worldbuilding.

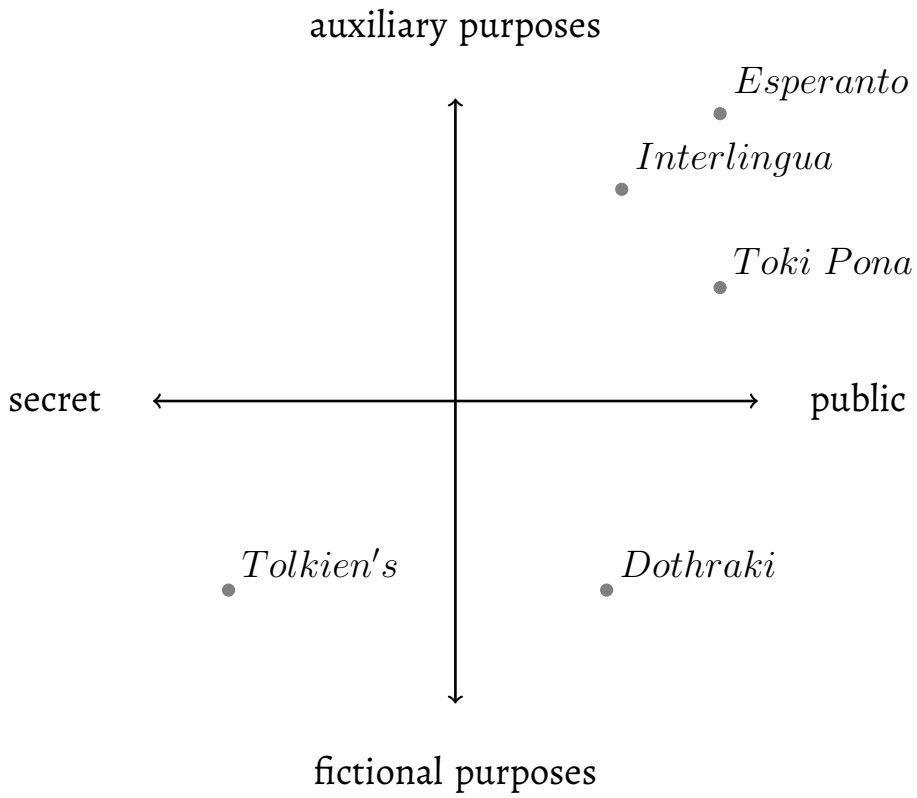

Figure .3: Federico Gobbo's coordinate system

What followed were classifications that took a far more granular approach. "A detailed typology is proposed by Albani and Buonarotti (1994), where a division is made into sacred and non-sacred languages. Sacred languages are further divided into structured (Bālaibalan) and non-structured

with six subdivisions. Nonsacred languages split into languages with communicative and expressive goals both with further detailed subdivisions [Stria 2016, p. 93]".

The classifications that followed saw a decrease in complexity and a shift to a more visual approach. Gobbo proposed a coordinate system which placed constructed languages between the "secret" and "public" extremes on the *x axis* and "auxiliary purposes" and "fictional purposes" on the *y axis* [Gobbo 2012]. This however, is also not ideal, as the publicity of languages tends to shift greatly over time.

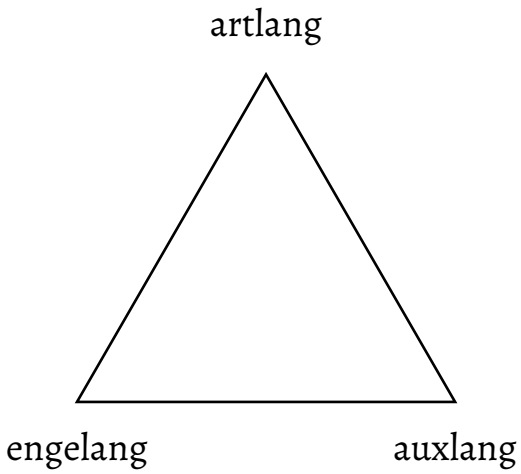

Figure .4: The modified Gnoli triangle

"One of the newest propositions widely spread on the Internet is the so-called Gnoli triangle. Claudio Gnoli, dissatisfied with the fact that his constructed language Liva was not easily classified, came up with an idea of a triangle whose vertices were labelled 'artlang' (artistic language), 'auxlang' (auxiliary language) and 'loglang' (logical language; the term 'engelang' was proposed later by And Rosta, apparently in 2001) [Stria 2016, p. 97]".

Constructed languages placed on the Gnoli triangle can be characterised as a gradient between three distinct purposes, depending on where on the triangle they are located.

- **Auxlang.** An international auxiliary language, that is, a language which

was devised with the intention of beaing a means of international communication. The majority of international auxiliary languages are meant to be second languages and not to replace native languages [Libert 2018].

- **Artlang.** A language created as a form of artistic expression or to fill an artistic role. Artistic languages often have irregular grammar systems, much like natural languages.
- **Engelang.** A language whose grammar or other feature is based on logic (a loglang) or a language which was created to as an experiment or to prove a hypothesis of how languages function.

3.3.3 Blanke's functional classification

The classification proposed by Blanke divides artificial languages into into invented projects and fully-fledged languages [Stria 2016]:

- **International auxiliary languages.** Languages intended to be used as a means of international communication.
- **Artistic languages.** Languages created for aesthetic reasons.
- **Constructed languages.** Languages invented to exercise the limits of language.
- **Exerimental languages.** Languages created to exercise a philosophical idea.

Final note on classifications

Within the conlanging community, the terms from the above classifications are used by the principles of compositionality. Individual

constructed languages are often described by having particular qualities borrowed from multiple classification systems.

4 Summary

LANGUAGE MODELLING AND TOKI PONA

1 Toki Pona

Toki Pona is a philosophical artistic language created by Sonja Lang, a Canadian linguist and translator. The core vocabulary of Toki Pona consists of around 120 words and focuses on simple, near-universal concepts. Although Lang herself never planned the language as an international auxiliary language, it might indeed be considered as such [Stria 2016, p. 100].

Lang has published two official books on Toki Pona. The first one, *Toki Pona: The Language of Good* was published in May 2014 and is known as *pu*. The second one, *Toki Pona Dictionary* was published in July 2021 and is known as *ku*.

1.1 History

1.1.1 Pre-pu

In the early days of Toki Pona, Lang experimented with both grammr and vocabulary. Numerous words from this period would be discarded before the online publication of the first draft of Toki Pona [Pake 2019].

Lang first revealed the language to the public in a 2001 draft [Lang 2001]. Multiple words are removed, replaced, and added. Toki Pona gains a small online following. The "community primarily meets to discuss the language on Yahoo Groups, experimenting with and fleshing out the language [Pake 2019]".

Around this time, Lang performs the first grammatical reform of Toki Pona. "The words 'en', 'kin', 'kan' and the concept of 'and' have been relatively unstable and confusing in Toki Pona, as I have been experimenting to find the best system <...> There is no longer a way to divide between modifiers. This is

no longer necessary. A 'tall and good' person is simply a tall 'good person' or a good 'tall person', as you will. [Lang 2002]".

The community surrounding Toki Pona grows, as well its corpus. This times sees the appearance of many early translations.

After more words are removed, for a while "the total word count stabilises at 118 [Pake 2019]". "The word *pu* makes its first appearance but remains undefined until the publishing of *Toki Pona: The Language of Good* [Pake 2019]".

The preface of the book reads: "Toki Pona is my philosopical attempt to understand the meaning of life in 120 words [Lang 2014]". Lang continues to further outline the primary characteristics of the language:

- "Toki Pona is semantically, lexically, and phonetically minimalist".
- "In many ways, Toki Pona resembles a pidgin. When people from different cultures need to communicated, they must focus on the elements that are most universal to our human experience".
- "Toki Pona offers a path for semantic reduction. <...> we can distill our thoughts to their most fundamental units to discover what things really mean. We can understand complex ideas in terms of their smaller parts".
- "An inherent idea of goodness is transparent throughout the language."

1.1.2 Post-pu

After the publication of the first book, a few new words become official. Some portions of the grammar undergo significant changes, and many words receive new or altered definitons [Pake 2019].

As the Toki Pona community rapidly grows, many idiolects permeate the way the language is spoken. Over a hundred of new words is invented by the community. Most of them are relatively well adopted.

The first issue of the *lipu tenpo* magazine is published online [Sonatan 2021]. The second official Toki Pona book *Toki Pona Dictionary* is published [Lang, Sartirani 2021]. Toki Pona is officially recognised by the ISO 639-3 registry under the code *tok* [SIL International 2022].

The official Toki Pona subreddit sees an influx of members, reaching almost 11.000 subscribers at the time of writing. While longer directly influencing the language, Lang remains an active member of the community.

1.2 Etymology

2 Distributional model

2.1 Pre-processing

In the context of machine learning, noise contained within the input data is detrimental to the output. Before the data can be used as input, it has to be thoroughly cleaned from noise. This process is referred to as text pre-processing [Vajjala, Majumder, Gupta 2020, p. 49].

The noise includes leftover formatting and HTML (HyperText Markup Language) tags, raw Unicode sequences, emojis, emoticons, punctuation, text in other languages, etc.

The Natural Language Toolkit Python library was used for the majority of text pre-processing tasks. "NLTK is a leading platform for building Python programs to work with human language data [NLTK Team 2022]".

Below is an example of a corpus entry at this stage:

```
Pilin mi la, ni li pona: i<em>think</em> lon li lon is a goood way of saying c'
   est la vie :)  \n\n*also* TOKI!!! pan suwi &#xe339 li pona mute tawa mi a a a
   ! :D &amp mi pali pona e pan suwi !! \n
```

2.1.1 Tokenisation

Tokenisation is the process of splitting text into individual tokens, usually representing words and punctuation marks [Vajjala, Majumder, Gupta 2020, p. 49].

Usually, this step is performed after the text has been split into sentences. The choice to perform tokenisation at this stage is supported by the fact that the source data contains an unusually large amount of text in other languages. Often, there is no separation between Toki Pona and other languages. If text tokenisation is performed first, it will prove easier to create custom rules for sentence segmentation at a later stage.

After being tokenised, the example entry takes the shape of an array:

```
['Pilin', 'mi', 'la', ',', 'ni', 'li', 'pona', ':', 'i', '<', 'em', '>', 'think
   ', '<', '/em', '>', 'lon', 'li', 'lon', 'is', 'a', 'goood', 'way', 'of', '
   saying', "c'est", 'la', 'vie', ':', ')', '*', 'also', '*', 'TOKI', '!', '!',
   '!', 'pan', 'suwi', '&', '#', 'xe339', 'li', 'pona', 'mute', 'tawa', 'mi', 'a
   ', 'a', 'a', '!', ':', 'D', '&', 'amp', 'mi', 'pali', 'pona', 'e', 'pan', '
   suwi', '!', '!']
```

2.1.2 Noise removal

Every sequence that only contains symbols or a mixture of letters and symbols that is longer than one characters is removed. The tokens which consist of only one punctuation mark are treated differently. Exclamation points and question marks are replaced with periods. The rest of the punctuation marks are removed

The tokens of the entry are then compared with a whitelist of allowed tokens. The whitelist includes all of the Toki Pona words, including the vocabulary that is not considered core vocabulary. When the matching is

case-insensitive, meaning that it converts each token to lowercase before matching. The case is preserved.

The words were not found to be in the Toki Pona vocabulary, are replaced with a period. The consecutive duplicate tokens which are not found in the whitelist are then removed from the sequence.

After this is done, the entry looks like this:

```
['Pilin', 'mi', 'la', 'ni', 'li', 'pona', '.', 'lon', 'li', 'lon', '.', 'a',
   '.', 'la', '.', 'TOKI', '.', 'pan', 'suwi', 'li', 'pona', 'mute', 'tawa', 'mi
   ', 'a', 'a', 'a', '.', 'mi', 'pali', 'pona', 'e', 'pan', 'suwi', '.']
```

2.1.3 Sentence segmentation

Sentence tokenisation is the process of splitting text based on the presence of punctuation marks [Vajjala, Majumder, Gupta 2020, p. 51].

Because there are no punctuation marks left in the entry except for the periods, the entry can be split naively on them:

```
[['Pilin', 'mi', 'la', 'ni', 'li', 'pona'], ['lon', 'li', 'lon'], ['a'], ['la'],
    ['TOKI'], ['pan', 'suwi', 'li', 'pona', 'mute', 'tawa', 'mi', 'a', 'a', 'a
   '], ['mi', 'pali', 'pona', 'e', 'pan', 'suwi']]
```

2.1.4 Normalisation

The first word of each sentence is brought to lowercase and compared with the whitelist. If there is a match, the word is replaced with its lowercase. Same is done for fully uppercase words in any part of the sentence.

The vocabulary of Toki Pona includes the word ‘ali’, which is an alternative spelling of the word ‘ale’. Within the corpus, in all cases where ‘ali’ appears, it is

replaced with ‘ale’.

All of the proper names are removed from the corpus. All of the sentences shorter than three tokens are removed from the corpus as well.

The resulting entry takes the following shape:

```
pilin mi la ni li pona
lon li lon
pan suwi li pona mute tawa mi a a a
mi pali pona e pan suwi
```

2.2 Vector space model

2.2.1 Training

The model is trained via the Gensim library. “Gensim is an open-source library for unsupervised topic modeling and natural language processing, using modern statistical machine learning. [Řehůřek 2022]”. Gensim is chosen as it comes prepackaged with multiple distributional model architecture and an simple interface for training them.

Word2vec trains individual vectors for each word form. While this would be detrimental to training a model of a highly inflected language, no words in Toki Pona are ever inflected.

Before the model is trained, all low frequency vocabulary is removed from the corpus. Consequently, sentences which are reduced to less than four words are removed as well. The vocabulary is kept to the set of words present in the original dictionary of the language.

The model is trained with the following parameters:

- **Input.** 270903 (271K) sentences, 2239370 (2.2M) words total.
- **Window size.** 5 words.

- **Vector dimensionality.** 114 dimensions, to match the vocabulary size of the corpus.

2.2.2 Dimensionality reduction

The Scikit-learn library is used to reduce the dimensionality of the model to two dimensions. Scikit-learn "features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy [Wikipedia 2022]".

Scikit-learn provides several methods for dimensionality reduction [scikit-learn developers 2022]. Among truncated singular value decomposition (Truncated SVD), principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE) the best separation was achieved using t-SNE. Moreover, t-SNE "takes a high dimensional data set and reduces it to a low dimensional graph that retains a lot of the original information [Sivarajah 2020]".

The following parameters were used:

- **Number of components.** 2, the dimension of the embedded space.
- **Perplexity or number of nearest neighbours.** 10, as no more than 10 words were expected to belong in the same semantic group.
- **Number of iterations.** 2500, for a higher degree of optimisation.

2.2.3 K-means clustering

The primary objective of K-means clustering is to group similar data points in multi-dimensional space. These groups can then be assigned colours.

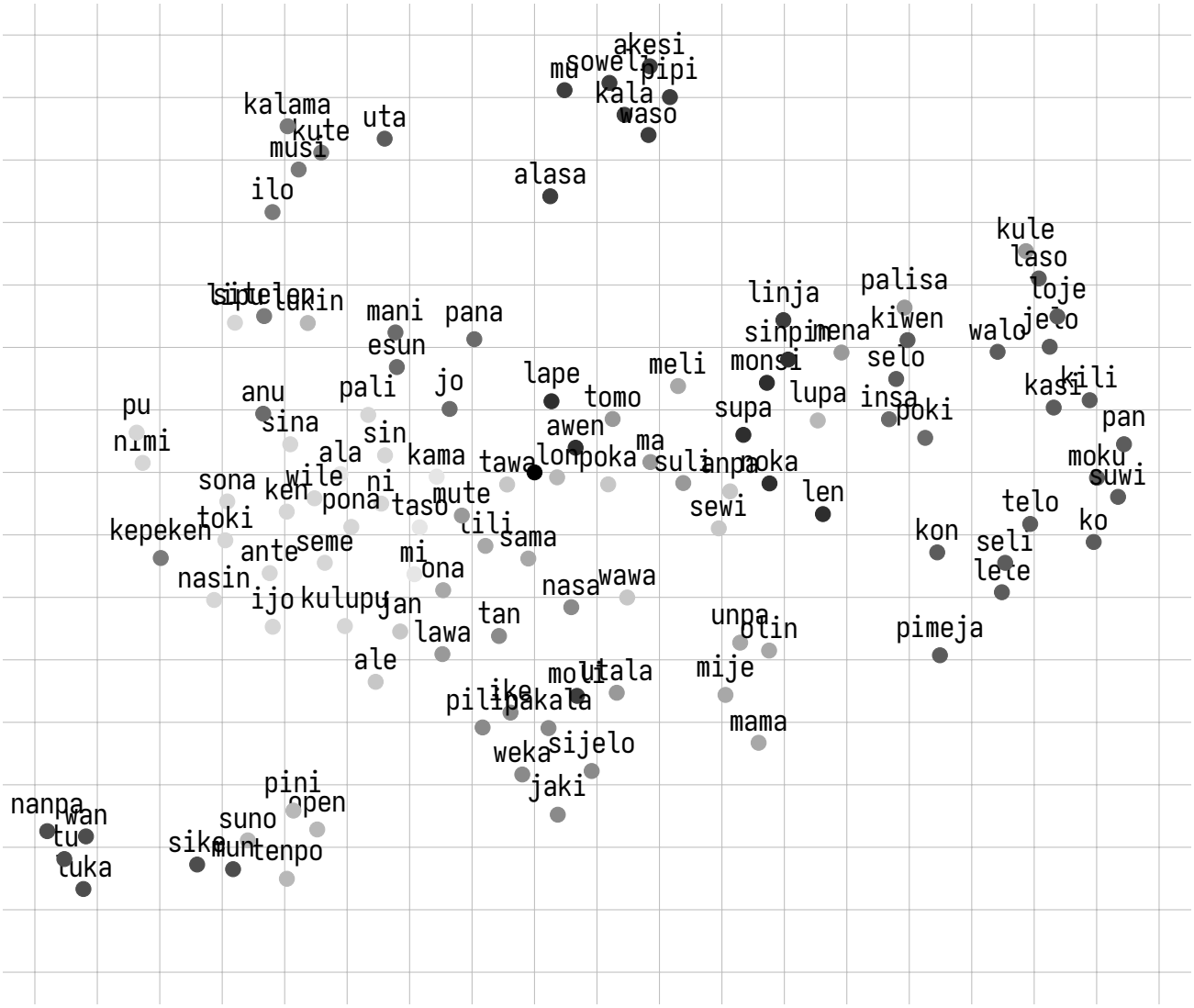

Figure .1: Toki Pona vector space model in two-dimensional space

The colours can then be used to draw the data points in any dimensionality, preserving the patters.

The KMeansClusterer module from NLTK is used to achieve this. "The K-means clusterer starts with k arbitrary chosen means then allocates each vector to the cluster with the closest mean. It then recalculates the means of each cluster as the centroid of the vectors in the cluster. This process repeats until the cluster memberships stabilise [NLP APIs 2016]".

The following parameters were used:

- **Number of clusters.** 13, roughly 10 words in each cluster.

- **Distance.** Cosine distance.

2.3 Observations

1. The projection of the model is observed. The relative positions of words and their assigned clusters are noted down.
2. For each analysed group, words are exctracted from the projection as a semantic field based on their expected distribution.
3. The semantic similarities between the words in the semantic field are calculated via cosine similarity from the perspective of the model, not the projection.
4. The deviations of cosine similarity values returned by the model are correlated with the patterns of the vocabulary use and the dictionary definitions.

The cosine similarity between the vector of the $\texttt{word}_a$ and the vector of the $\texttt{word}_b$ is shown as $\texttt{cs(word}_a\texttt{, word}_b\texttt{) = cosine similarity}$. Cosine similarity ranges from $[-1]$ to $[1]$.

2.3.1 kule

There are only five words in Toki Pona that can denote **kule** (colors): **pimeja** (*adj.* black, dark, unlit), **walo** (*adj.* white, whitish; light-coloured, pale), **loje** (*adj.* red, reddish), **jelo** (*adj.* yellow, yellowish), **laso** (*adj.* blue, green). For the rest of the colours, the rules of colour mixing apply: **jelo loje** can be used to mean *orange*. Some of these words have broader meanings that describe not only the colour, but its vibrance, e.g **loje pimeja** can be used to mean *dark red*.

The presence of **kasi** (*n.* plant, vegetation; herb, leaf) among the colours can be attributed to Toki Pona having only one word for blue and green – **laso**.

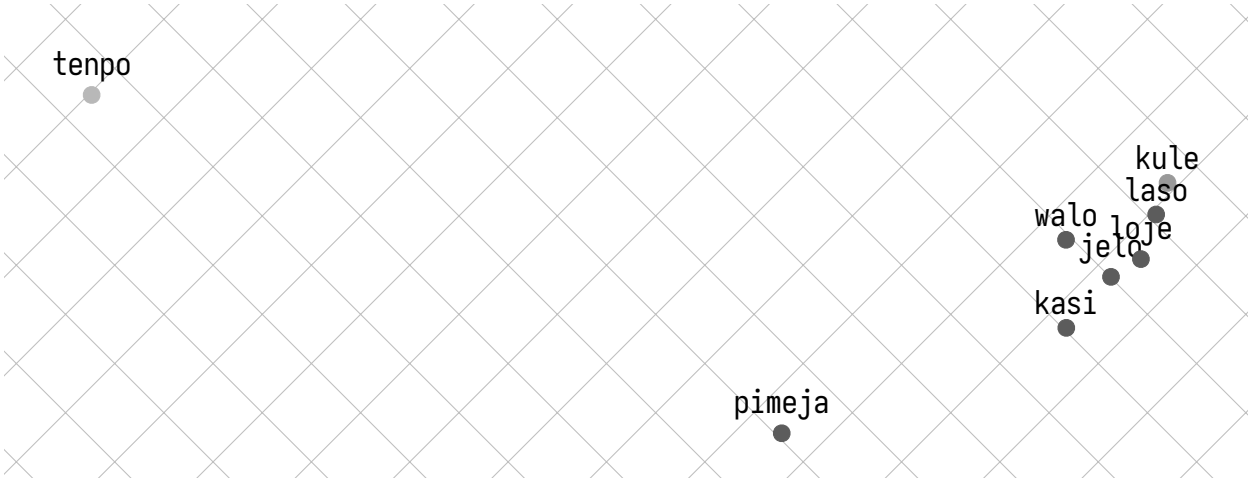

Figure .2: kule, $-45°$

Despite the existence of many languages that also have one word for blue and green, the distribution is evident of the fact that many Toki Pona speakers opt to refer to the color green unambiguously `cs(kasi, kule) = 0.51`.

The word **pimeja** is considerably separated from the rest of the colours. This is explained by it often appearing near **tenpo** (*n.* time, duration, moment, occasion, period, situation). In turn, **tenpo pimeja** is a common way of referring to *nighttime*:

(1) tenpo pimeja la, mi lape.
time black *cm* I sleep
I sleep at night.

Among words for colours, this property is unique to **pimeja**: `cs(walo, tenpo) = -0.03`, `cs(loje, tenpo) = -0.07`, `cs(jelo, tenpo) = -0.11`, `cs(laso, tenpo) = -0.06`. Interestingly, **pimeja** experiences an even semantic split between **kule** and **tenpo**: `cs(pimeja, kule) = 0.24`, `cs(pimeja, tenpo) = 0.24`.

2.3.2 nanpa

There are only four words which are related to **nanpa** (numerals) in the vocabulary of Toki Pona: **wan** (*adj.* unique, united. *num.* one) used to mean *one*,

tu (*num.* two) used to mean *two*, **luka** (*n.* arm, hand, tactile organ. *num.* five) used to mean *five*, and **mute** (*adj.* many, a lot, more, much, several, very. *n.* quantity). With the exception of **mute**, which can be used to denote any large quantity, these words form an independent cluster in the lower left corner of the projection.

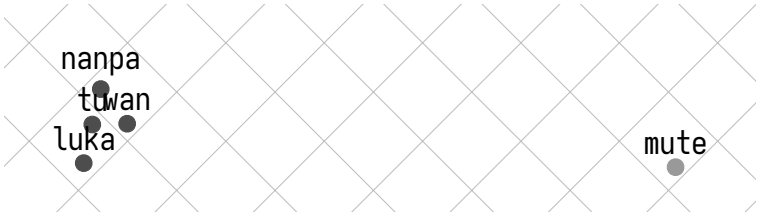

Figure .3: nanpa, $-45°$

The word **luka** is semantically detached from its lower body counterpart, **noka** (*n.* foot, leg, organ of locomotion; bottom, lower part): `cs(luka, noka) = 0.35`, `cs(luka, nanpa) = 0.76`. This highlights the unique ability of **luka** to denote numerals among other words that indicate parts of the body.

Despite one of the word sense of **mute** being *quantity*, it experiences little semantic similarity with **nanpa**: `cs(mute, nanpa) = 0.25`, as **mute** is most commonly used for specifying *intensity*:

(2) mi pilin pona mute a.
I feel good very em
I am feeling very good!

One of the secondary uses of **nanpa** and **wan** is involved in expressing high degree of positivity towards something:

(3) kili ma li pona nanpa wan.
vegetable land pm good number one
Potatoes are the best.

The distribution shows that this use of **nanpa wan** is farily uncommon: `cs([nanpa, wan], [pona, mute]) = 0.14`.

2.3.3 soweli

The cluster **soweli** (*n.* animal, beast, land mammal) contains words that refer to animals: **akesi** (*n.* non-cute animal; reptile, amphibian), **kala** (*n.* fish, marine animal, sea creature), **waso** (*n.* bird, flying creature, winged animal), **pipi** (*n.* bug, insect, ant, spider), **akesi** (*n.* non-cute animal; reptile, amphibian).

Figure .4: soweli, $-45°$

The onomatopoeia **mu** (*p.* (animal noise or communication)) is also present in this cluster: `cs(soweli, mu) = 0.54`. Among other words in the cluster, **pipi** is the least semantically related to **mu**: `cs(pipi, mu) = 0.29`.

The dictionary entry for **akesi** contains the now considered outdated *non-cute animal* word sense. The fact that this definition was in use at a time is evident from the semantic similarity between **suwi** (*adj.* sweet, fragrant; cute, innocent, adorable) and **akesi**: `cs(akesi, suwi) = 0.27`; compared to the other members of the cluster: `cs(soweli, suwi) = 0.53`, `cs(kala, suwi) = 0.51`, `cs(waso, suwi) = 0.49`, `cs(pipi, suwi) = 0.45`.

The word **mani** (*n.* money, cash, savings, wealth; large domesticated animal) has the secondary definition of a *large domenticated animal*. This could be attributed to the themes of hunter-gatherer culture that permeate the language. In the model, only a very weak semantic connection between **mani** and **soweli** is present: `cs(mani, soweli) = 0.18`.

2.3.4 kalama

A small cluster of **kalama** (*v.* to produce a sound; recite, utter aloud) includes words related to sound production and perception: **uta** (*n.* mouth, lips, oral cavity, jaw), **kute** (*n.* ear. *v.* to hear, listen; pay attention to, obey). The earlier discussed **mu** is also in the close proximity of this cluster.

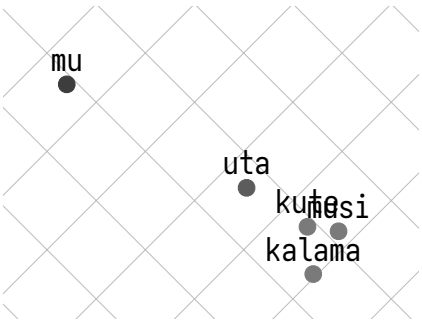

Figure .5: kalama, 135°

The presense of **musi** (*adj.* artistic, entertaining, frivolous, playful, recreational) in the cluster is expected due to *music* being often described as **kalama musi**:

(4) mi kute e kalama musi la, mi wile tawa musi.
I hear *do* sound entertaining *cm* I want move entertainingly.
When I'm listening to music, I want to dance.

Music is among the most universal forms of entertainment that is often discussed online, and the semantic distribution is the evidence of that. The semantic similarity between the two words is `cs(kalama, musi) = 0.66`.

2.3.5 lawa

2.3.6 sijelo

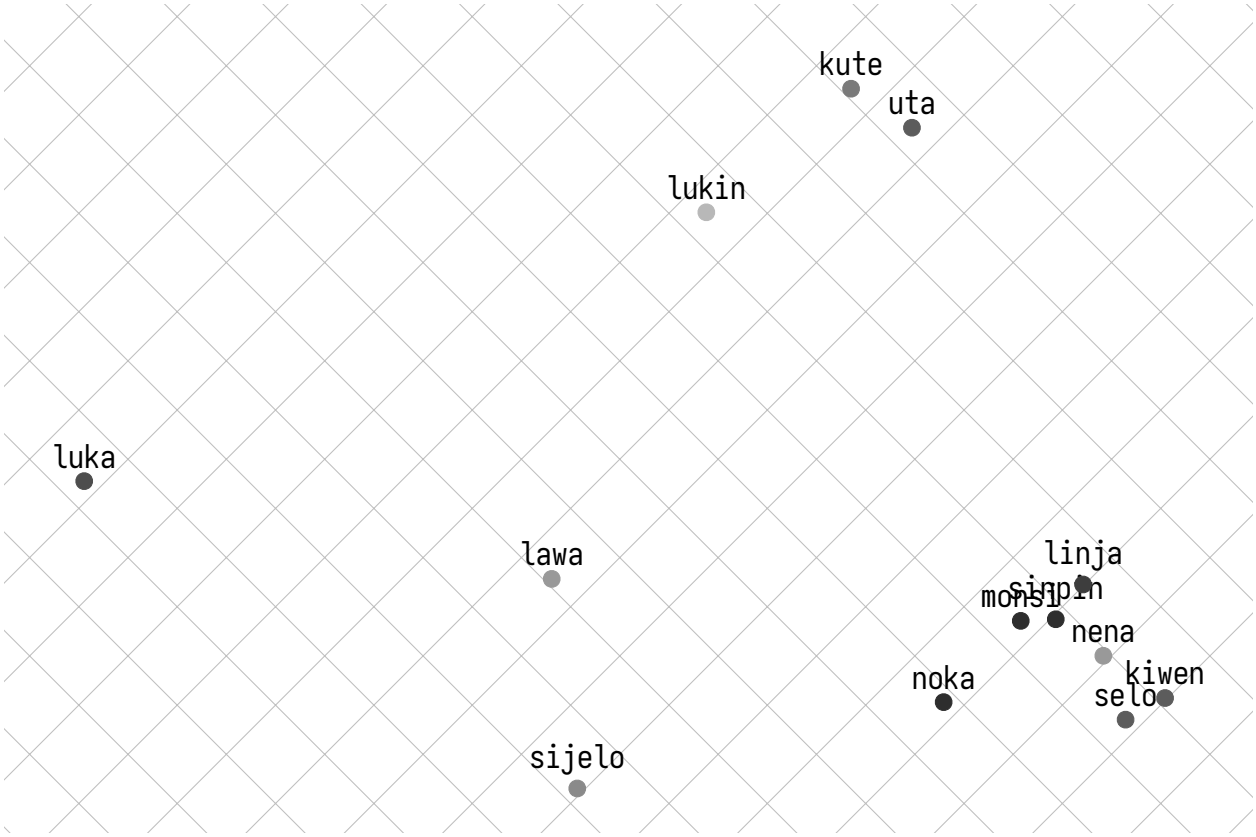

Figure .6: sijelo, $-45°$

2.3.7 sijelo

2.3.8 sinpin

2.3.9 ike

3 Summary

CONCLUSION

REFERENCES

1. *Ahire J. B.* Introduction to Word Vectors // Retrieved March. — 2018. — Vol. 12. — P. 2018.
2. *Alshari E. M., Azman A., Doraisamy S.* Improvement of sentiment analysis based on clustering of Word2Vec features // 2017 28th international workshop on database and expert systems applications (DEXA). — IEEE. 2017. — P. 123–126.
3. *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.
4. *Bekkerman R., El-Yaniv R., Tishby N.* Distributional Word Clusters vs. Words for Text Categorization // Journal of Machine Learning Research. — 2003. — Mar. — Vol. 3. — P. 1183–1208. — DOI: 10.1162/153244303322753625.
5. *Bengio Y., Ducharme R., Vincent P.* A Neural Probabilistic Language Model //. Vol. 3. — 01/2000. — P. 932–938. — DOI: 10.1162/153244303322533223.
6. *Couturat L., Leau L.* Histoire de La Langue Universelle (1903). — Literary Licensing, LLC, 2014. — ISBN 9781498147248. — URL: https://books.google.ru/books?id=s2%5C_poQEACAAJ.
7. *Dalianis H.* Clinical text mining: Secondary use of electronic patient records. — Springer Nature, 2018.
8. *Eco U.* The search for the perfect language. — Search, 1995.

9. *El Desouki M. I., Gomaa W. H.* Exploring the recent trends of paraphrase detection // International Journal of Computer Applications. — 2019. — Vol. 975, S 8887.

10. *Emerson G.* What are the Goals of Distributional Semantics? — 2020. — DOI: 10.48550/ARXIV.2005.02982. — URL: https://arxiv.org/abs/2005.02982.

11. *Erk K.* What do you know about an alligator when you know the company it keeps // Semantics and Pragmatics. — 2016. — Vol. 9.

12. *Fimi D.* Tolkien, race, and cultural history: from fairies to Hobbits. Vol. 146. — Springer, 2009.

13. *Firth J.* A Synopsis of Linguistic Theory, 1930-1955. — 1957. — URL: https://books.google.ru/books?id=T8LDtgAACAAJ.

14. *Gobbo F.* Alan Turing creator of Artificial Languages // InKoj. — 2012. — Sept. — Vol. 3. — DOI: 10.13130/2037-4550/2385.

15. *Hagiwara M.* Real-World Natural Language Processing: Practical Applications with Deep Learning. — Manning, 2021. — ISBN 9781617296420. — URL: https://books.google.ru/books?id=0k5NEAAAQBAJ.

16. *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.

17. *Henestroza Anguiano E., Denis P.* FreDist: Automatic construction of distributional thesauri for French. — 2011. — June.

18. *Janton P., Tonkin H., Edwards J.* Esperanto: Language, Literature, and Community. — State University of New York Press, 1993. — ISBN 9781438407807. — URL: https://books.google.sk/books?id=R%5C_ZGfG2a1tIC.

19. *Kennaway R.* Artificial Languages //. — 05/2013. — ISBN 978-0-415-42432-5.

20. *Lang S.* Toki Pona: The Language of Good. — Sonja Lang, 2014. — ISBN 9780978292300. — URL: https://books.google.ru/books?id=5P0ZjwEACAAJ.

21. *Lang S., Sartirani V.* Toki Pona Dictionary. — Amazon Digital Services LLC - KDP Print US, 2021. — (Official Toki Pona). — ISBN 9780978292362. — URL: https://books.google.ru/books?id=ybqPzgEACAAJ.

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

23. *Libert A. R.* Artificial Languages. — 06/2018. — DOI: 10.1093/acrefore/9780199384655.013.11. — URL: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-11.

24. *Lo Bianco J.* Invented languages and new worlds // English Today. — 2004. — Vol. 20, no. 2. — P. 8–18. — DOI: 10.1017/S0266078404002032.

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

26. *Meulen S. v. d.* Request for New Language Code Element in ISO 639-3 // ISO 639-3 Registration Authority. — 2021. — Aug. — Vol. 2021. — P. 15.

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

28. *Musto C., Narducci F., Basile P.* Cross-Language Information Filtering: Word Sense Disambiguation vs. Distributional Models //. Vol. 6934. — 09/2011. — P. 250–261. — ISBN 978-3-642-23953-3. — DOI: 10.1007/978-3-642-23954-0_24.

29. *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.

30. *Okrent A.* In the Land of Invented Languages: Adventures in Linguistic Creativity, Madness, and Genius. — Random House Publishing Group, 2010. — (Spiegel & Grau trade paperbacks). — ISBN 9780812980899. — URL: https://books.google.ru/books?id=3anWeY0G2moC.

31. *Sanders N.* A primer on constructed languages //. — 08/2020. — P. 6–26. — ISBN 9780198829874. — DOI: 10.1093/oso/9780198829874.003.0002.

32. *Silva F. T. da, Maia J. E.* Query Expansion in Text Information Retrieval with Local Context and Distributional Model. // J. Digit. Inf. Manag. — 2019. — Vol. 17, no. 6. — P. 313.

33. *Straubhaar S. B.* Sarah Higley, Hildegard of Bingen's Unknown Language: An Edition, Translation and Discussion.(The New Middle Ages series.) Palgrave Macmillan, 2007 // Medieval Feminist Forum: A Journal of Gender and Sexuality. Vol. 44. — Society for Medieval Feminist Scholarship. 2008. — P. 158–161.

34. *Stria I.* Inventing languages, inventing worlds. Towards a linguistic worldview for artificial languages. — 01/2016. — ISBN 978-83-947609-1-5. — DOI: 10.14746/9788394760915.

35. *Tolkien J. R. R., Tolkien C.* The Monsters and the Critics, and Other Essays //. — 1983.

36. *Tolkien J., Carpenter H., Tolkien C.* The Letters of J.R.R. Tolkien: A Selection. — Houghton Mifflin Company, 2000. — ISBN 9780618056996. — URL: https://books.google.ru/books?id=pw4nlr1ieVEC.

37. *Tolkien J., Fimi D., Higgins A.* A Secret Vice: Tolkien on Invented Languages. — HarperCollins Publishers, 2019. — ISBN 9780008348090. — URL: https://books.google.ru/books?id=iRokwQEACAAJ.

38. *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.

39. *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.

40. *Vajjala S., Majumder B., Gupta A.* 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.

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

42. *Wilkins J.* An Essay Towards a Real Character, and a Philosophical Language. — Sa: Gellibrand, for John Martyn printer to the Royal Society, 1668. — URL: https://books.google.ru/books?id=BCCtZjBtiEYC.

43. *Williams M.* Ireland's Immortals: A History of the Gods of Irish Myth. — Princeton University Press, 2016. — ISBN 9781400883325. — URL: https://books.google.ru/books?id=o5cTDAAAQBAJ.

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

ONLINE RESOURCES

45. *Lang S.* Toki Pona Word List. — 2001. — URL: https://web.archive.org/web/20070818031155/http://bknight0.myweb.uga.edu/toki/about/olddict.html (visited on 10/05/2022).

46. *Lang S.* the words "en", "kin" and "kan". — 2002. — URL: http://forums.tokipona.org/viewtopic.php?f=33&t=81&sid=2a993bc8cd0daccec16ca43739ad5c89 (visited on 10/05/2022).

47. *NLP APIs*. nltk.cluster.KMeansClusterer. — 2016. — URL: https://tedboy.github.io/nlps/generated/generated/nltk.cluster.KMeansClusterer.html.

48. *NLTK Team*. Natural Language Toolkit. — 2022. — URL: https://iso639-3.sil.org/code_tables/639/data/t?title=tok&field_iso639_cd_st_mmbrshp_639_1_tid=All&name_3=&field_iso639_element_scope_tid=All&field_iso639_language_type_tid=All&items_per_page=200 (visited on 11/05/2022).

49. *Pake k.* The Evolution of Toki Pona. — 2019. — URL: https://www.reddit.com/r/tokipona/comments/iw2xp6/the_evolution_of_toki_pona/ (visited on 10/05/2022).

50. *Řehůřek R.* What is Gensim? — 2022. — URL: https://radimrehurek.com/gensim/intro.html#what-is-gensim.

51. *scikit-learn developers*. sklearn.manifold.TSNE. — 2022. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

52. *SIL International*. ISO 639 Code Tables. — 2022. — URL: https://iso639-3.sil.org/code_tables/639/data/t?title=tok&field_iso639_cd_st_mmbrshp_639_1_tid=All&name_3=&field_iso639_element_scope_tid=All&field_iso639_language_type_tid=All&items_per_page=200 (visited on 10/05/2022).

53. *Sivarajah S.* Dimensionality Reduction for Data Visualization: PCA vs TSNE vs UMAP vs LDA. — 2020. — URL: https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29.

54. *Sonatan j.* lipu tenpo li lon. — 2021. — URL: https://www.reddit.com/r/tokipona/comments/la0owt/lipu_tenpo_li_lon/ (visited on 10/05/2022).

55. *Wikipedia*. scikit-learn. — 2022. — URL: https://en.wikipedia.org/wiki/Scikit-learn.

SUPPLEMENTARY MARTERIAL

1 Model vectors

The two-dimensional projection of the semantic model constructed in the course of this reseach.

word	x	y	word	x	y
akesi	36.91	130.0	ala	-62.14	-0.54
alasa	5.03	88.39	ale	-50.87	-67.04
anpa	62.62	-6.07	ante	-84.8	-32.24
anu	-86.88	18.78	awen	13.16	7.89
esun	-44.08	33.71	ijo	-83.83	-49.42
ike	-7.69	-76.89	ilo	-83.9	83.32
insa	113.5	17.01	jaki	7.45	-109.57
jan	-43.02	-50.94	jelo	164.97	40.21
jo	-27.19	20.31	kala	28.83	114.46
kalama	-79.08	110.8	kama	-31.34	-1.38
kasi	166.25	20.75	ken	-79.34	-12.56
kepeken	-119.69	-27.41	kili	177.77	23.11
kiwen	119.44	42.33	ko	179.06	-22.28
kon	128.95	-25.61	kule	157.36	70.82
kulupu	-60.73	-49.2	kute	-68.3	102.37
lape	5.43	22.77	laso	161.47	62.08
lawa	-29.5	-58.17	len	92.35	-13.35
lete	149.66	-38.36	lili	-15.71	-23.53
linja	79.68	48.74	lipu	-95.93	47.91
loje	167.44	49.91	lon	7.24	-1.58

Table .1: Toki Pona model projection

word	x	y	word	x	y
luka	-144.45	-133.36	lukin	-72.61	47.81
lupa	90.7	16.62	ma	37.1	3.31
mama	71.76	-86.55	mani	-44.53	44.8
meli	45.97	27.63	mi	-38.46	-32.59
mije	61.15	-71.23	moku	180.11	-1.72
moli	13.62	-71.54	monsi	74.41	28.67
mu	9.62	122.35	mun	-96.53	-127.01
musi	-75.53	96.95	mute	-23.28	-13.83
nanpa	-156.04	-114.84	nasa	11.77	-43.15
nasin	-102.6	-40.84	nena	98.3	38.36
ni	-49.09	-10.0	nimi	-125.47	3.03
noka	75.25	-3.53	olin	75.11	-57.03
ona	-29.25	-37.71	open	-69.58	-114.3
pakala	4.48	-81.85	pali	-53.25	18.4
palisa	118.49	52.76	pan	188.76	9.02
pana	-19.28	42.64	pilin	-16.63	-81.64
pimeja	129.82	-58.54	pini	-77.21	-108.26
pipi	43.36	120.13	poka	23.59	-3.83
poki	125.1	11.05	pona	-58.67	-17.45
pu	-127.39	12.73	sama	-1.99	-27.58
seli	150.71	-28.95	selo	115.81	29.9
seme	-67.16	-28.94	sewi	59.0	-17.91
sijelo	18.3	-95.59	sike	-108.03	-125.53
sin	-47.87	5.46	sina	-78.22	9.0
sinpin	81.2	36.12	sitelen	-86.58	50.03
sona	-98.38	-9.31	soweli	23.99	124.68

Table .1: Toki Pona model projection

word	x	y	word	x	y
suli	47.63	-3.41	suno	-91.8	-117.79
supa	66.88	12.01	suwi	186.86	-7.83
tan	-11.37	-52.37	taso	-36.77	-17.58
tawa	-8.77	-3.88	telo	158.72	-16.5
tenpo	-79.3	-130.08	toki	-99.01	-21.74
tomo	25.01	17.09	tu	-150.56	-123.77
unpa	65.85	-54.5	uta	-47.99	106.82
utala	26.33	-70.54	walo	148.29	38.6
wan	-143.63	-116.52	waso	36.51	108.0
wawa	29.68	-40.05	weka	-3.92	-96.7
wile	-70.47	-8.29			

2 Toki Pona dictionary

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

Table .2: nimi pu

Word		Definition
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

Table .2: nimi pu

Word		Definition
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

Table .2: nimi pu

Word		Definition
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

Table .2: nimi pu

Word		Definition
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

Table .2: nimi pu

Word		Definition
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

Table .2: nimi pu

Word		Definition
	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

Table .2: nimi pu

Word		Definition
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