Toki Pona and Orders of Semantic Completeness

jan Kijupo
tomo sona sewi pi jan Kaneki en jan Melon
kijupo@cmu.edu

0 April 2024

Abstract

Natural language semantics is a field of linguistics that focuses on determining the meaning of linguistic expressions. Current research is ongoing to construct a natural semantic metalanguage (NSM) that could represent the meaning of any expression in any language. We present a notion of semantic completeness based on translatability and loss of information which may be useful in the pursuit of an NSM. This notion of completeness is built from the tools of formal semantics, a method for building abstract representations of meaning for complex expressions based on the meanings of their constituent parts. We propose the hypothesis that a majority of extant human languages are Toki Pona-complete, and we present the case for English specifically using the tools discussed in this article, building up a rigorous semantic model using $(\text{SM})^2\text{L}$ and showing that the symbolic formula for a lossless translation holds in general. Our research suggests the possibility of semantic complexity classes to create a well-defined notion of completeness as a relation between two languages, particularly a natural language and a constructed minimal language.

1 Background and Motivation

Computer scientists are obsessed with notions of completeness. Turing completeness can be used to evaluate whether a programming language or model of computation is robust enough for all general computational tasks. NP-completeness indicates that an NP problem is maximally hard, and algorithms to solve that particular problem can shed light more generally on all problems in NP. Idiot-completeness can shed light on how effectively you can communicate a seemingly simple concept to your coworkers, since if an idiot-complete explanation exists, you should be able to explain it to anyone (and if not, good luck). Many related notions of completeness have been defined for various types of complexity classes, but one notable field for which notions of completeness have not been defined, despite potential utility, is natural language semantics.

Semantics is, fundamentally, the study of meaning in natural language. Meaning can be thought of in this context as the correspondence between linguistic symbols[1] and the concepts these symbols represent. Formal semantics is a method to leverage the compositional nature of human languages to build representations of meaning for complex expressions based on the meanings of their constituent parts. It focuses on the use of formal constructs from logic, mathematics, and even theoretical computer science to build rigorous denotations – or representations of literal meaning – of linguistic expressions from smaller constituent symbols.

One actively researched field of semantics is the notion of semantic primes, which would be the smallest, most irreducible units of semantic meaning. All other concepts expressible in any human language would be built up from some composition of these primes, even if this composition is not transparent from the surface forms (Indeed, in some languages, the exponent for a semantic prime may contain multiple morphemes or even multiple words). Universality is key here, as the utility of semantic primes comes from the ability to represent the meaning of any expression in any language in terms of these primes. Semantic primes are the foundation for the theory of Natu-

[1] We use "symbol" as a generic term to encompass any mode of transmitting lingusitic information, including spoken, signed, written, telepathic, etc. Questions such as whether written text counts as "utterances", or whether "" is a word, are beyond the scope of this paper.

ral Semantic Metalanguage, which suggests that all human languages have a fundamental semantic core that can be represented entirely from prime concepts, and therefore a minimal language consisting of such primes would be no less expressive than any natural language[1]. The theory is still debated, but more than 60 hypothesized semantic primes have been discovered since the first research into NSM in the 1970s. This field research suggests that there is some shared semantic core universal to all human language, but this alone does not prove that *all* meaning can be reduced to semantic primes.

In order to have the desired expressive power, any proposed NSM should be **semantically complete**, in other words, at least as expressive as any other language. Every concept in the NSM should be easily, if not trivially, translatable into any human language. Any complex or culturally unique concept should have some representation in the NSM, albeit a potentially very long representation for concepts native speakers may find very simple. In fact, it should be a maximal covering the minimum common semantic space in order to cover all common areas of the human experience. A diagram of how NSM might fit into the semantic complexity classes of other languages is presented in Figure 1.1.

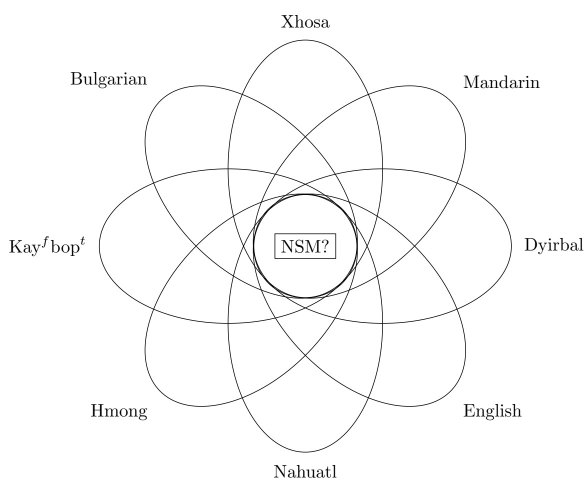

Figure 1.1: Rough Venn diagram of a handful of Semantic Complexity Classes. A natural semantic metalanguage should sit at the intersection of all of these.

Keep in mind that while we are using the term "complexity class" to describe semantic spaces of different languages, we do not suggest the existence of a hierarchy of complexity classes for different languages. Rather, we posit that different languages have different complexity classes, each of which is rich with culturally significant vocabulary and meanings that cannot necessarily be expressed concisely in other languages. Aside from artificial languages and simplifications, any two languages should have nonzero intersection in terms of semantic complexity, but neither should be a subset of any other. Part of the goal of NSM is to find this common intersection among the semantic spaces of all human languages.

1.1 Toki Pona

One language that has some of the desired properties of an NSM, although not itself designed to be an NSM, is Toki Pona. Toki Pona is a constructed language (conlang) created by linguist Sonja Lang in 2001, designed to be minimal in as many dimensions as possible while maintaining the expressive capabilities of other natural languages[2]. It is designed to be fairly easy to communicate common day-to-day experiences without introducing unnecessary complexity, breaking down advanced ideas into simpler concepts: "telo" can mean "water", but also "liquid", "wet", etc. Similarly, "moku" can mean "food", "to eat", "edible", "nutrition" or any number of related concepts. The semantics of Toki Pona have been characterized as being intentionally vague and subjective rather than ambiguous: "moku" doesn't itself mean "food" in some contexts and "edibility" in others, but rather encompasses the meanings of all its translated exponents simultaneously[3], making it a hypernym for a variety of concepts related to eating.

(1) jan li lukin e lipu
person COP look DOBJ book
'The person reads a book'

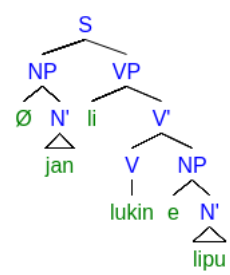

Figure 1.2: Gloss and Syntactic breakdown of a simple Toki Pona sentence

Due to the structure of Toki Pona and the naturally varied usage of most words that encompasses both noun-like and verb-like functions, some analyses of the language suggest that the distinction between content words and particles is more useful[4]. Content words in Toki Pona carry most of the semantic meaning of a sentence, while particles[2] mostly carry information about the syntactic structure. Consider the simple sentence in gloss (1) and the corresponding syntax tree in Figure 1.2, which is a good example of the basic structure of a transitive sentence in Toki Pona. The particles 'li' and 'e' denote the start of the main verb and the direct object, respectively, each serving as the specifier for that syntactic position. In some cases, as with the main subject, the specifier is empty, leaving only the head of the phrase and the word order to denote the syntax.

2 Definitions

2.1 Translatability

We will use translatability as a metric to construct relations to indicate semantic compatibility and complexity. Translations are notorious for their difficulty to accurately construct: loss of shades of meaning, introduction of ambiguity not present in the source, and differences in cultural context that create meaningfully different interpretations of deceptively similar forms. We can measure and quantify this loss of information to compare the quality of different translation schemas. Whether truly accurate translations that don't lose information are even possible is still a hotly debated topic: opponents argue that any translation necessarily loses some of the nuances present in the source language. However, for the sake of this paper, I will adopt a framework that implies that a translation that loses no information is possible given sufficient length or a sufficiently large context. For example, the Finnish word "sisu" is generally considered to have no English translation, but given a sentence such as "Hänellä on paljon sisua" (see Gloss 2), we can translate the sentence trivially by translating "sisu"⇝"*sisu*" and appending the text of the English Wikipedia article on the Finnish concept of Sisu to the end of the sentence[5]. Arguably, very little (if any) information is lost here, but translating a four-word sentence in one language into a similarly sized sentence that requires a 2500-word footnote to understand in the target language is unwieldy in practice. Information may be preserved, but understanding is

(2) Hän-ellä on paljon sisu-a
3sg-ADE be much sisu-PART
'They have a lot of *sisu**'
*See [5] for explanation of '*sisu*'

Figure 2.1: Gloss for translation-by-footnote example

likely to be hampered.

Because of this unwieldiness, we will only consider the ability to interpret translations that are *efficient* by some metric. Intuitively, we would want the number of clauses in the source to be about the same as the number of clauses in the translation, and the number of morphemes in the translation should be within a constant factor of the number of morphemes in the source. In other words, if a sentence in the source language contains n symbols, then the translation of that sentence should have length $O(n)$. We can include context in the translation length by way of "footnotes", where each symbol token in each footnote counts towards the overall length. Thus, a translation of $O(n)$ length permits $O(n)$ tokens in the translation text itself, $O(n)$ tokens in a global footnote, and $O(1)$ tokens in individual symbol footnotes. Any footnotes for a specific symbol cannot reference other specific symbols, eg. to capture a relationship between two or more symbols, so judgments like this would necessarily be in the global context. The length of a footnote containing judgments relating k different symbols is at worst $O(n^k)$, since there are at most $O(n^k)$ ways of selecting k symbols.

In practice, for most[3] language pairs, the differences in inflection amount to a bounded number of checks to perform for each word: things like gender, plurality, case, tense, status as a penguin, etc. which we can encode as a constant number of simple predicates about each word in the sentence. Extending this, we can consider other types of predicates that carry non-trivial information about a word. The English word "bank", for example, may refer to a financial institution or to the side of a river, whereas other languages have distinct words for each of these senses. However, if you're translating English into, say, Chinese, you can easily decide whether to translate "bank" as

[2] Including prepositions, preverbs, and semiparticles when not used as content words

[3] We ran into a snag with Kay^{f}bop^t to English, which theoretically requires $\Omega(n(\log\log v + \log\sigma))$ footnote space for any sentence, where v is the market value of the most expensive noun in the sentence, and σ is the number of standard deviations from the mean of the most extreme adjective in the sentence[8]. The impact of this on lossless translatability is unclear.

(3) Jeohui abeoji-kkeseo o-syeoss-seumnida (*Hangul*: 저희 아버지-께서 오-셨-습니다)
1PL.POL father-NOM.HON come-PST-FORMAL.POL
'Our father has arrived'

(4) Ci vediamo alle 19.00
1PL see.1PL at 7:00.pm
'I'll see you at 7 PM'

(5) sel^{f}bush^{b}rub^{t}sip^{p}mif^{f}guv^p-san^{t}ap^t-vlir^t-sang^b-es^{p}u^t-vom^{b}ngo^p-gag^b-yam^{t}ack^{t}vlim^p-kay^{f}kay^f kay^{f}dan^{f}tuw^{t}fob^{p}san^t-ap^t
humanity-NOM-FIN-EXP:Y-NFOOD-HUM-DEATH:WOMBAT-COOL-$77 language-ACC-FIN-EXP:Y-N
'All of humanity had one language'

(6) sel^{f}bush^{b}rub^{t}sip^{p}mif^{f}guv^{p}san^t-ap^t-vlir^t-sang^b-es^{p}u^t-vom^{b}ngo^p-gag^b-yam^{t}ack^{t}vlim^p-kay^{f}kay^f kay^{f}dan^{f}tuw^{t}fob^{p}san^{t}ap^{t}v
stranger-ACC-FIN-EXP:Y-NFOOD-HUM-DEATH:WOMBAT-COOL-$77 to.PREP
'We are not strangers to love'

Figure 2.2: Glosses for a selection of sentences that cannot be losslessly translated. See [6], [7] for more on these examples.

"银行" or "岸": even in cases where context doesn't make the choice clear, you can find out with a single question. This brings us to what we call the "20-questions" test: if, given any word in language A, you can determine the contextually correct translation in language B (provided that it exists) by asking 20 or fewer yes-or-no questions about the context, then you should be able to translate any sentence from A to B with a context of size $O(n)$. The important point isn't 20 specifically, but more broadly the idea that there is a fixed (and ideally small) upper bound to the number of classifications we need to make for any word.

This intuition on lossless vs. lossy translations should suffice to provide clear definitions for our reduction, but the formal definitions require some tools that will be introduced in Section 3. In the meantime, we provide some further examples to strengthen these intuitions in Figure 2.2, which are discussed below.

2.1.1 Examples of untranslatable pairs

There are several cases in which a sentence may be considered untranslatable, most of which stem from fundamental cultural differences. One, as in the Finnish example with "sisu", is where a specific word or phrase carries a large amount of meaning unique to the culture of the speakers. Another related issue is honorific systems, and context-sensitive variations for pronouns or inflection, as in the Korean example in gloss (3). Notably, each word in the sentence contains either an honorific or politeness marker, and changing the sentence to remove these would be a valid construction, but would affect the meaning in a way not easily captured by an English translation. While an argument can be made that you can determine the correct honorific with a finite-depth decision tree, such a decision tree can be very large and have poor worst-case performance with respect to number of judgements necessary, so we will not count these as passing the "20-questions" test[4].

Other cases include issues of subtext, which is often heavily reliant on culture. These differences can introduce different interpretations of sentences with denotations that seem nearly identical on the surface. An example of this is chronicity: most English-speaking cultures view time monochronically, meaning there is an expectation of closely adhering to scheduled commitments and doing only one thing at a time, whereas many other cultures are polychronic, which means multiple commitments can be scheduled at the same time, and people arriving late to the function[5] is to be expected[6]. We see in gloss (4) that nothing seems to explicitly indicate the polychronicity of the Italian sentence or the monochronicity of the English translation, so we can hypothesize that the meaning of the part glossed "at 7:00 PM" means different things in each language, and that the two cannot be easily translated without an explanation about chronicity or the speaker's expectations.

Another major consideration is ambiguity, which

[4] In fact, since honorific and polite forms usually capture a relationship between two or more people, the judgements necessary would need a pair of inputs, yielding $O(n^2)$ when this type of honorific is used.

[5] In the proverbial sense of an event to attend, rather than a total, unique mapping between two sets or types with no meaningful changes to external state.

while common enough in monolingual settings, can become further exacerbated when trying to construct a translation. A good example is Kay^{f}bop^t, which is notorious for its unique ways of dividing up semantic space (among other things)[7]. For example, the sentence in glosses (5) and (6) can be alternatively translated as "All of humanity had one language" or "We are not strangers to love" depending on the context, with neither really being a more accurate translation in general. Ambiguity is generally considered by semanticists to be a feature of natural language and not a bug[9], but ambiguity may be expressed differently in different languages. This is especially true for native speakers of Kay^{f}bop^t (and other languages in the Vlim^{p}kay^{f}sna^f family) who wholeheartedly embrace the ambiguity present in their language, with some roots having up to 18 different meanings[10].

2.2 Reductions

From the notion of lossy and lossless translations, we can define our reduction relation:

Definition 2.1 (B-Completeness)**.** For languages A and B, we have that A is B**-complete** ($A \geq_S B$) if and only if for any expression $s \in B$, there exists a lossless translation $s' \in A$. Translations from A to B are possible, but may be lossy.

In short, if we can find a lossless translation for every sentence in B, then a language is B-complete. If any sentence cannot be losslessly translation, completeness fails. Based on this definition, the desired completeness property for an NSM is that all other human languages are NSM-complete. This would guarantee that translations (by way of exponents) of such an NSM into any language would lose no information[6], and therefore concepts expressed in NSM could be understood by speakers of any language.

There are two other interesting relations that can also be constructed. Notably, these both form equivalence classes.

Definition 2.2 (Compatibility)**.** Languages A and B are **semantically compatible** ($A \simeq_S B$) iff any expression $s \in A$ has a potentially lossy translation $s' \in B$, and vice versa.

Definition 2.3 (Equivalence)**.** Languages A and B are **semantically equivalent** ($A \equiv_S B$) iff $A \geq B$ and $B \geq A$, ie. lossless translations from A to B and from B to A are possible.

The former relation is interesting because it presents a way of formalizing the compatibility problem: To what extent can humans communicate with other species? This would be particularly useful for communication with any extraterrestrial intelligence we may encounter in the future, as well as research into interspecies communication with dolphins, octopi, birds, apes, etc. is possible. Current research suggests that all human languages are semantically compatible with each other, so for the purposes of this paper, we will take it as given. Actually proving this is left as a task for future research.

Semantic compatibility is also an essential property for full-scale conlangs, including any NSM constructions, since not being semantically compatible with other human languages represents the absence (or addition) of a fundamental category of meaning that cannot be expressed otherwise. If, for a given proposed NSM, it is possible to find a natural language that is NSM-complete but the NSM is not semantically compatible with it, this means that the proposed NSM is incapable of representing all possible denotations present in human languages, so it can be rejected as an NSM. Strictly speaking, this does conflict with intuition about the meaning of completeness, which suggests providing another term, maybe "hardness", to represent the relation of $A \gtrsim_S B$, which indicates that A is fundamentally more expressive than B.

The latter relation may be used to compare how closely related different language varieties are to each other. Even if two varieties of a language are not mutually intelligible, if they are semantically equivalent, they are probably very closely connected. Conversely, it can be used to test whether a conlang is a relex of another existing language, natural or constructed. For instance, a conlang that is semantically equivalent to English is most likely just a relex of English, and will probably be received as unoriginal and lacking in creativity, except if being a relex is intentional. Pig Latin, for instance, is semantically equivalent to English, which we can show easily by simply noting that we can map every English word to its translation in Pig Latin by applying a simple, reversible transformation. The proof of this is left as an exercise for the reader.

For some intuition, all Turing-complete programming

[6] Modulo cases of polysemy, which would be resolved by the 20-questions test and corresponding context

languages are, by definition, semantically compatible, but not necessarily semantically equivalent. While it is theoretically possible to rewrite, for example, the entire PyTorch library in Brainfuck, it would probably be a bad idea.

3 Formalization

Many of you might look at this and think "great, we have a theoretical model, let's set up some language pairs with semantic vectors and see what similarities we can find!" While you certainly could do this, and I'm not stopping you[7], I'd prefer to use the tools of formal semantics to present a deeper theoretical formalization of the reduction framework discussed in Section 2. A formal semantic analysis will allow us to reason about languages more abstractly, which is particularly useful when trying to prove statements about translatability.

3.1 ILL and MSML

In order to formalize the notion of meaning, we will be using an Intensional Logical Language (ILL) as an intermediary between object language symbols and metalinguistic representations of meaning. ILL follows the syntactic conventions of the lambda calculus, but with some additional tools that represent quantifiers and context-sensitive determiners that don't easily or efficiently translate into pure computations. Nevertheless, we can reason about these constructs roughly like a computer language with modal separation: ILL to represent linguistic expressions, and some kind of semantic metalanguage to determine the meaning of those expressions by parameterizing encapsulated ILL expressions and evaluating their value. To avoid confusion with NSM, we will call this the Montaguean Semantic Metalanguage (MSML) after the creator, Richard Montague [11]. If we allow the language to be evaluated nondeterministically and hand-wave away any loops over potentially infinite[8] data, then it's perfectly fine (if unwieldy) as a programming language!

Translations from an object language (usually a natural language) into ILL are generally done syntactically. Declarative sentences are assumed to have type t representing truth values[9], named entities to have type e, and most other constituents are assigned a function type $\tau_1 \to \tau_2$[10]. Once all the constituents have a type, we can apply a series of rules at each branch in the syntax tree to determine how they compose together. In a simple ILL for English, we might have the following rules for a branch connecting expressions a and b:

- For $a \dot{\sim} \tau_1 \to \tau_2,\ b \dot{\sim} \tau_1$, we take $a\ b \rightsquigarrow \ulcorner a \urcorner(\ulcorner b \urcorner)$
- For $a \dot{\sim} \tau_1,\ b \dot{\sim} \tau_1 \to \tau_2$, we take $a\ b \rightsquigarrow \ulcorner b \urcorner(\ulcorner a \urcorner)$
- For $a \dot{\sim} \tau \to t$, $b \dot{\sim} \tau \to t$, we take $a\ b \rightsquigarrow \ulcorner a \urcorner \sqcap \ulcorner b \urcorner$, where the $\sqcap$ operator is just syntactic sugar for an 'and' operation, given by $P \sqcap Q := \lambda x.P(x) \wedge Q(x)$

The first two rules handle simple cases of function application, which is the most common type of rule since more constituents than not receive a function type. The third rule loosely deals with the composition of multiple adjectives, meaning we can represent intuitively that a "big red ball" is big and red and a ball, while keeping the translations of adjectives as simple lambdas, such as "red" $\rightsquigarrow \lambda x.\mathsf{red}'(x)$, instead of some more complicated combinatorial expression.

Although many constructs in ILL are easily recognizable to mathematicians and theoretical computer scientists, the iota expression is specific to formal semantics. It is typically read as "the unique x such that α", and refers to a specific entity in a given context. The syntactic constructs used in the ILL are given below in Figure 3.1:

Expression	**Description**
x	Linguistic Variable / Assignable
$\lambda x_{[\tau_1;\tau_2]}.\alpha$	Lambda expression
$\iota x_{[\tau]}.\alpha$	Iota expression
$\forall x_{[\tau]}.\alpha, \exists x_{[\tau]}.\alpha, \ldots$	Quantifiers[11]
$\alpha_1 \to \alpha_2$	Logical implication
$\alpha_1 \wedge \alpha_2$	Logical and
$\alpha_1 \vee \alpha_2$	Logical or
$\neg\alpha$	Logical negation
$\alpha_1(\alpha_2)$	Function application

Figure 3.1: Basic expressions in ILL, excluding model-defined constants

[7]Good luck trying to make a vector representation for the likes of Kay^{f}bop^t...

[8]or finite but infeasibly large

[9]and possibly divergence

[10]Some of you may be more familiar with the alternate notation for this type, $\langle\tau_1, \tau_2\rangle$.

[11]There are some extensions to the typical universal and existential quantifiers, such as the "for most" quantifier $\mathsf{W}x.\alpha$ and the "for few" quantifier $\exists x.\alpha$, which are useful syntactic sugar for more complex underlying expressions[9]. These tend to be harder to reason about than traditional logical quantifiers, but are useful in translating natural language quantifiers.

Symbol	Representation	Type
most	$\lambda P.\lambda Q.\mathsf{W}x.P(x) \rightarrow Q(x)$[12]	$(e \to t) \to (e \to t) \to t$
dog	$\lambda x_{[e;t]}.\mathsf{dog}'(x)$	$e \to t$
bark	$\lambda x_{[e;t]}.\mathsf{bark}'(x)$	$e \to t$
Jay	Jay'	e
see	$\lambda x_{[e;e\to t]}.\lambda y_{[e;t]}.\mathsf{see}'(x)(y)$	$e \to e \to t$
the	$\lambda P_{[e\to t;e]}.\iota x.P(x)$	$(e \to t) \to e$
rabbit	$\lambda x_{[e;t]}.\mathsf{rabbit}'(x)$	$e \to t$
small	$\lambda x_{[e;t]}.\mathsf{small}'(x)$	$e \to t$

Table 3.1: Table of ILL translations for the English examples

Most of the static typing rules are as expected: lambda expressions have function types, logical operators take expressions of type t, etc. The statics for iota expressions and arbitrary quantifiers are given below. We borrow notions of context Γ (for metalinguistic variables) and signature Σ (for linguistic assignables) from analogous constructs from descriptions of MA found in [12].

$$\frac{\Gamma \vdash_{\Sigma, x\sim\tau} \alpha \mathrel{\dot\sim} t}{\Gamma \vdash_{\Sigma} \iota x_{[\tau]}.\alpha \mathrel{\dot\sim} \tau} \qquad \frac{\Gamma \vdash_{\Sigma, x\sim\tau} \alpha \mathrel{\dot\sim} t}{\Gamma \vdash_{\Sigma} \forall x_{[\tau]}.\alpha \mathrel{\dot\sim} t}$$

For an example, we'll look at the English sentences "Most dogs bark" and "Jay sees the small rabbit". The first of these can be broken down as a quantifier, a class, and an action, while the latter includes a named entity, a transitive relation, and a definite noun phrase with an adjunct modifier. By putting together the translated words as detailed in Table 3.1, this gives us the following translations:

$$\text{“Most dogs bark”} \rightsquigarrow$$
$$(\lambda P.\lambda Q.\mathsf{W}x.P(x) \rightarrow Q(x))(\lambda x.\mathsf{dog}'(x))(\lambda x.\mathsf{bark}'(x))$$

$$\text{“Jay sees the small rabbit”} \rightsquigarrow$$
$$(\lambda x.\lambda y.\mathsf{see}'(x)(y))(Jay')((\lambda P.\iota x.P(x))(\lambda x.\mathsf{rabbit}'(x) \sqcap \lambda x.\mathsf{small}'(x)))$$

Under eager evaluation, these can be further simplified respectively to

$$\mathsf{W}x.\mathsf{dog}'(x) \rightarrow \mathsf{bark}'(x)$$

$$\mathsf{see}'(Jay')(\iota x.\mathsf{rabbit}'(x) \wedge \mathsf{small}'(x))$$

ILL expressions are not themselves metalinguistic, as they must be parameterized by three terms before they can be meaningfully evaluated:

Expression	Description
$[\![\alpha]\!]^{\mathcal{M},g,w}$	Encapsulated denotation
x	Metalinguistic variable
$\mathcal{F}(\alpha)$	Intension
$\texttt{get}[g](x)$	Assignable lookup
$\top$	Literal true
$\bot$	Literal false
m_1 **and** m_2	Logical and
m_1 **or** m_2	Logical or
$\texttt{!}m$	Logical negation
$m_1 \stackrel{?}{=} m_2$	Equality test
$\texttt{all}(D\,;a.m)$	Universal quantification
$\texttt{some}(D\,;a.m)$	Existential quantification
$m_1(m_2)$	Function application

Figure 3.2: Constructs in MSML

- A language model, $\mathcal{M} = \langle \mathcal{F}, \Omega, W \rangle$, where $\mathcal{F}$ maps constant expressions in ILL to their metalanguage equivalents, W is the set of possible worlds, and Ω is the ontology, which is the set of possible entities and objects across all worlds[13].
- An assignment function, g, that indicates the current value of variable expressions, including, but not limited to, the antecedents of pronouns in a given context.
- A particular (potentially arbitrary) world $w \in W$, in which the truth conditions of an expression can be evaluated.

This brings us to MSML, which, as briefly mentioned earlier, encapsulates ILL expressions within denotation brackets and parameterizes them with specific $\mathcal{M}, g, w$, allowing for metalinguistic analysis of the encapsulated expression. Take note of the modal separation here: linguistic expressions have a linguistic type, $\phi \mathrel{\dot\sim} \tau$, while metalinguistic expressions have a metalinguistic type, $m : \tau$. In particular, types for encapsulated linguistic symbols are governed by the following static rules:

$$\frac{\Gamma \vdash_{\Sigma} \phi \mathrel{\dot\sim} \tau}{\Gamma \vdash_{\Sigma} \mathcal{F}(\phi) : s \to \tau} \qquad \frac{\Gamma \vdash_{\Sigma} x \mathrel{\dot\sim} \tau}{\Gamma \vdash_{\Sigma} \texttt{get}[g](x) : \tau}$$

$$\frac{\Gamma \vdash_{\Sigma} \phi \mathrel{\dot\sim} \tau \qquad \Gamma \vdash_{\Sigma} w : s}{\Gamma \vdash_{\Sigma} [\![\phi]\!]^{\mathcal{M},g,w} : \tau}$$

Modal separation allows us to reason about top-level denotation expressions regardless of the specific context, including information about variable/pronoun

[12] Type annotations omitted for brevity

assignments[12]. Therefore, expressions like "I saw them" can be evaluated to an expression that indicates the truth conditions relative to the assignment of the variables for $\ulcorner$I$\urcorner$ and $\ulcorner$them$\urcorner$. In theory, the modal separation present in MSML also allows the encapsulation of *any* (well-typed) linguistic expression in *any* language, although our focus will primarily be on different varieties of ILL. Instead of writing $[\![\text{cow}]\!]^{\mathcal{M},g,w}$, we will use $[\![\ulcorner\text{cow}\urcorner]\!]^{\mathcal{M},g,w}$ to indicate that we will look at the representation of the expression "cow" in ILL, rather than the expression itself. The semantics should follow no differently either way, but ILL is easier to reason about than untranslated linguistic symbols from an arbitrary language.

In order to evaluate the meaning of an expression, we need to evaluate the dynamics of the encapsulated expression in MSML. Conventionally, this is represented at the top level by $[\![\phi]\!]^{\mathcal{M},g,w} \stackrel{?}{=} \top$, which evaluates to the truth conditions of the linguistic expression ϕ in terms of metalinguistic representations of its constituent symbols. Whether lambdas are applied within ILL or only at the metalanguage level depends on your choice of lazy or eager evaluation. We will take $D_\tau \subseteq \Omega$ to be the universal set of type τ, which contains all instances $d : \tau$. The domains that are most frequently invoked are the domain of entities D_e and the domain of worlds D_s.

$$\frac{\phi \ \mathsf{const}}{[\![\phi]\!]^{\mathcal{M},g,w} \longmapsto \mathcal{F}(\phi)w} \qquad \frac{x \ \mathsf{var}}{[\![x]\!]^{\mathcal{M},g,w} \longmapsto \mathtt{get}[g](x)}$$

$$\frac{[\phi \ \mathsf{val}]}{[\![\forall x_{[\tau]}.\phi]\!]^{\mathcal{M},g,w} \longmapsto \mathtt{all}(D_\tau \,;\, a.[\![\phi]\!]^{\mathcal{M},g\otimes a/x,w})}$$

$$\frac{[\phi \ \mathsf{val}]}{[\![\exists x_{[\tau]}.\phi]\!]^{\mathcal{M},g,w} \longmapsto \mathtt{some}(D_\tau \,;\, a.[\![\phi]\!]^{\mathcal{M},g\otimes a/x,w})}$$

$$\frac{}{\mathtt{all}(\emptyset \,;\, -) \longmapsto \top} \qquad \frac{}{\mathtt{some}(\emptyset \,;\, -) \longmapsto \bot}$$

$$\frac{\exists d \in D, D' = D \setminus \{d\}}{\mathtt{all}(D \,;\, a.m) \longmapsto [d/a]m \ \textbf{and} \ \mathtt{all}(D' \,;\, a.m)}$$

$$\frac{\exists d \in D, D' = D \setminus \{d\}}{\mathtt{some}(D \,;\, a.m) \longmapsto [d/a]m \ \textbf{or} \ \mathtt{some}(D' \,;\, a.m)}$$

The full static and dynamic rules for MSML are omitted for brevity, but are broadly discussed in [11], albeit in non-computational terms.

In order to handle hypernym-hyponym relationships and entity-class relationships, a model is generally evaluated with a set of meaning postulates: short tatuological expressions in ILL that represent logical relationships between constants. These may be a quantified expression over all entities, such as $\forall x_{[e]}.\mathsf{rabbit}'(x) \rightarrow \mathsf{animal}'(x)$ representing the concept "All rabbits are animals", or an expression exclusively in terms of constants, like $\mathsf{person}'(Alex')$ which represents "Alex is a person". We generally assume that named entities are uniquely identifiable[13], so any basic information about an entity that is true over all worlds may be included as a meaning postulate, whereas other judgements, like $\mathsf{student}'(Alex')$ only hold in a subset of possible worlds, namely those where Alex is a student.

Although there are a number of varieties of MSML, including Event Semantics[14] and SML of New Jersey[15], we will focus on a simple version of MSML based on a simplified variety of ILL. We will call this language SMSML for Simplified MSML, or $(\text{SM})^2$L for short. $(\text{SM})^2$L is built using the ILL flavor used in the previous example (see Table 3.1), where nouns, adjectives and intransitive verbs are assigned constants that are similar to the indicator function of a set, that is, $[\![\mathsf{dog}'(x)]\!]^{\mathcal{M},g,w}$ is true if and only if x belongs to the set of all dogs, and $[\![\mathsf{bark}'(x)]\!]^{\mathcal{M},g,w}$ is true if and only if x belongs to the set of all things that bark (or are bark, as it were). Transitive verbs are represented as a relation between two entities, which is more in line with intuition. Information about plurality, tense, and aspect is ignored.

3.2 Translation as Implication

With these tools in mind, we can define some relations to get a handle on semantic complexity. We will make use of the implication expression in ILL, which is often used by semanticists to evaluate entailments within a language. For a pair of languages A and B, let s be a sentence in Language A, and $s \rightsquigarrow s'$ be a translation of s into Language B.

Definition 3.1 (Lossless Translation)**.** A translation, $s \rightsquigarrow s'$, is **lossless** if there exists a context C consisting of $O(|s| + |s'|)$ Boolean judgements such that $[\![\ulcorner s'\urcorner \wedge C \rightarrow \ulcorner s\urcorner]\!]^{\mathcal{M},g,w} \stackrel{?}{=} \top$ is a tautology for all possible g, w under a compatible model $\mathcal{M}$.

Definition 3.2 (Lossy Translation)**.** A translation, $s \rightsquigarrow s'$, is **lossy** if there exists a context C^* of ar-

[13]Implementation varies, but I'm partial to `uuidgen` myself

bitrary size such that $[\![\ulcorner s' \urcorner \wedge C^* \rightarrow \ulcorner s \urcorner]\!]^{\mathcal{M},g,w} \stackrel{?}{=} \top$ is a tautology for all possible g, w under compatible model $\mathcal{M}$, and there is no way to express C^* as a linear number of Boolean judgments.

Recall that translations occur on the level of expressions, whereas completeness is defined on the level of entire languages. To say that language A is B-complete is to say that for *all* possible expressions $s \in B$, it is possible to generate a lossless translation $s \rightsquigarrow s' \in A$. If there is even a single expression in B for which no suitable translation $s' \in A$ and linear size context C can be constructed that makes the expression $[\![\ulcorner s' \urcorner \wedge C \rightarrow \ulcorner s \urcorner]\!]^{\mathcal{M},g,w} \stackrel{?}{=} \top$ a tautology, then completeness does not hold, ie. $A \not\geq_S B$.

4 Tap to add section title

We propose the hypothesis that a majority of extant human languages are Toki Pona-complete, and we will present the case for English specifically in this section, building up a rigorous semantic model using the tools discussed in Section 3. Although Toki Pona is probably not (and was never designed to be) a perfect NSM, its nature as a minimal language for human communication can act as a stepping stone to augment ongoing research into semantic primes to search for an NSM.

4.1 Modelling Toki Pona in (SM)²L

The translation of content words is fairly straightforward, but somewhat nontrivial. Content words pattern nicely with the type $e \to t$, but there is a clear hierarchy in meaning for any modifiers attached to the same head. For instance, "jan telo" and "telo jan" mean roughly "sailor" and "bodily fluid" respectively, but naively translating content words to the same constant in (SM)^{2}L would result in both of these becoming equivalent to $\lambda x.\mathsf{jan}'(x) \wedge \mathsf{telo}'(x)$, which is a problem. To remedy this, we will make two constants for each content word: when used as a head, we translate "telo" $\rightsquigarrow \lambda x.\mathsf{telo}'_H(x)$, but when used as a modifier, we translate it as "telo" $\rightsquigarrow \lambda x.\mathsf{telo}'_M(x)$.

Another issue is that of quantification. English has distinct definite and indefinite markers, whereas Toki Pona does not have either of these general-case quantifiers. There are contexts in which an iota expression may be closer to the intended meaning, but it may be reasonable to analyze the indefinite as the default and use an existential quantifier, since the existential construction is an entailment of the iota construction, ie. $Q(\iota x.P(x))$ implies $\exists x.P(x) \wedge Q(x)$ for all P, $Q \mathrel{\dot{\sim}} e \to t$, but the converse does not hold. Based on the design principles of Toki Pona as discussed in [2], and the brief semantic analysis presented in [3], the broader meaning seems like the best representation. When an explicit quantifier like 'ni' or 'ale' is used, we will use the corresponding ILL quantifier (see Figure 4.1).

$$
\begin{array}{rcl}
\text{ale} & \rightsquigarrow & \lambda P.\forall x.P(x) \\
\text{mute} & \rightsquigarrow & \lambda P.\mathbb{W}x.P(x) \\
\text{ni} & \rightsquigarrow & \lambda P.\iota x.P(x) \\
\text{li} & \rightsquigarrow & \lambda P.\lambda x.P(x) \\
\text{e} & \rightsquigarrow & \lambda y.\lambda P.\lambda x.P(x) \wedge \mathsf{e}'(x)(y) \\
\text{pi} & \rightsquigarrow & \lambda P.\lambda x.\exists y.P(y) \wedge \mathsf{pi}'(x)(y) \\
\text{ala} & \rightsquigarrow & \lambda P.\forall x.\neg P(x)
\end{array}
$$

Figure 4.1: Proposed ILL translations of some particles in Toki Pona. These may not be accurate translations in all circumstances.

Named entities in Toki Pona have obligatory marking for class, including a head and optionally some additional modifiers. This means that the name of a person, like Sonja Lang, is rendered in Toki Pona as "jan Sonja" and not just *"Sonja", and the name of a place, such as Finland[14], is rendered as "ma Sumi" rather than *"Sumi". The upshot of this is that the meaning postulates needed to handle named entities in most other languages are pre-encoded into any Toki Pona sentence already! Where a Model for English needs meaning postulates to represent "Sonja Lang is a person", and append an ILL representation like $\mathsf{person}'(Sonja')$ to the model, Toki Pona might simply translate the phrase "jan Sonja" as $\lambda P.P(Sonja') \wedge \mathsf{jan}'_H(Sonja')$ to capture the implicit meaning postulate. Alternatively, we may construct a simple meaning postulate that captures that all entities can be assumed to belong to the class with which they are introduced.

The final step is to observe the syntax of Toki Pona. Most of the content words will combine with the modifier concatenation rule, $a\ b \rightsquigarrow \ulcorner a \urcorner \sqcap \ulcorner b \urcorner$, so all we have left are the particles. Being primarily head-inital, we find that quantifier-like particles in Toki Pona tend to come after the head of a phrase, while prepositions, case-like particles, and verbal auxiliaries

[14]This claim is disputed

tend to come before the head. This means that both directions of function application are likely to be seen given the right contexts. With that, we can construct translations for the most common particles, which are listed in Figure 4.1.

4.2 Bilingual Models

Montaguean semantic models are typically used to ascertain the meaning of expressions within a given language, rather than to compare the denotations of expressions in two or more different languages. Since we need to work with multiple languages simultaneously in our use case, we need to create a notion of multilingual models for our ILL and MSML structures to work as expected.

If we have two language models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we can define their union, $\mathcal{M}_1 \oplus \mathcal{M}_2$, to be the tuple $\langle \mathcal{F}_1 \sqcup \mathcal{F}_2, \Omega, W \rangle$, where $\mathcal{F}_1 \sqcup \mathcal{F}_2$ maps ILL constants to the corresponding metalanguage expressions for constants in both $\mathcal{F}_1$ and $\mathcal{F}_2$. We will take as a given that the input domains of $\mathcal{F}_1$ and $\mathcal{F}_2$ do not overlap as long as they are from models representing different languages: any overlap can be handled trivially by explicitly marking the constants by the model from which they originate[15]. Following the intuition that a model of a given language should reflect the judgements and understanding of a native speaker, we can think of a compound model as a reflection of the judgements of someone who is a native speaker of both languages[16]. In order to handle this, the union model $\mathcal{M}_1 \oplus \mathcal{M}_2$ will also have to contain meaning postulates to describe the relationships between expressions in each language. For example, if a bilingual speaker of English and Toki Pona would say that something described as 'a book' in English would be 'lipu' in Toki Pona, then we can represent this relation with the meaning postulate $\forall x.\mathsf{book}'_{\mathsf{eng}}(x) \rightarrow \mathsf{lipu}'_{H,\mathsf{tok}}(x)$. We can't do this for all pairs of overlapping words in general due to polysemy, although we can combine them with other judgements found in the context, and create a meaning postulate based on the 20-questions test like so: $\forall x.\mathsf{bank}'_{\mathsf{eng}}(x) \wedge \neg \text{EDGE.WATER}'(x) \rightarrow \text{银行}'_{\mathsf{zho}}(x)$ where $\text{EDGE.WATER}' \mathrel{\dot{\sim}} e \rightarrow t$ is an ILL translation of a metalinguistic judgement of whether x is the edge of a body of water, and we represent the Chinese word for 'bank, financial institution' (seen in Section 2.1) as $\ulcorner \text{银行} \urcorner = \lambda x.\text{银行}'_{\mathsf{zho}}(x)$.

[15] eg. by a subscript ISO 639-3 code, our convention of choice.

[16] Strictly speaking there are no native speakers of Toki Pona, but for the sake of our model, we can approximate this by imagining the model to reflect a speaker with advanced or fluent level of speaking. About 16% of respondents to the 2021 Toki Pona Census claim these levels of proficiency[16].

(7) soweli lili li kalama
animal small COP sound
'The little dog barked'

Figure 4.2: Gloss for our worked example.

4.3 Proving Losslessness

To begin, let's consider the Toki Pona sentence presented in Figure 4.2, gloss (7) as s. This sentence can be translated into English in a number of ways: 'The little dog barked', 'A small cat is purring', 'The mice sqeak', 'The rabbit barks', 'Small animals are sound waves', are all valid translations of this sentence. It doesn't make much difference which one we choose – in fact all of these English sentences are (in a sense[17]) precizations of the original sentence in Toki Pona – so for simplicity and familiarity, we can take 'The little dog barked' to be our s'.

We can construct some meaning postulates to reflect the relationships between relevant English and Toki Pona concepts. In particular, we will utilize the following:

1. $\forall x.\mathsf{dog}'_{\mathsf{eng}}(x) \rightarrow \mathsf{soweli}'_{H,\mathsf{tok}}(x)$
2. $\forall x.\mathsf{little}'_{\mathsf{eng}}(x) \rightarrow \mathsf{lili}'_{M,\mathsf{tok}}(x)$[18]
3. $\forall x.\mathsf{bark}'_{\mathsf{eng}}(x) \wedge \neg \text{TREE.PART}'(x) \rightarrow \mathsf{kalama}'_{H,\mathsf{tok}}(x)$

Toki Pona doesn't have any significant inflectional grammar that we need to account for in our translation, so the trivial context, $\top$, works fine for the grammar of the sentence. We do need a judgement for the word 'bark', since this could refer to either an animal sound or a part of a tree[19], so we assert that $\text{TREE.PART}(x)$ is false for our context.

The last step before plugging in our values into the definition of lossless translation is producing the ILL representations. We will assume eager evaluation and

[17] Rigorously defining precizations in a bilingual setting is difficult since they rely on entailments, which we already had to augment to deal with translations. We use "precization" loosely to reflect the judgments of a bilingual speaker. Modulo issues with cross-linguistic comparisons, it is reasonable to say that English 'dog' is a hyponym of Toki Pona 'soweli'[9].

[18] The head interpretation of 'lili' would also work, but isn't used in this derivation.

[19] or something similar to either of these, technically

skip the intermediary expanded form, only showing the result after lambda evaluation.

$$\ulcorner s \urcorner = \exists x.\mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x)$$

$$\ulcorner s' \urcorner = \mathsf{bark}'_{\mathsf{eng}}(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x))$$

$$C = \neg\text{TREE.PART}'(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x))$$

Finally, we can plug these expressions into the MSML expression given in Definition 3.1 and evaluate the results. Taking the union of simple models for English and Toki Pona gives us $\mathcal{M} = \mathcal{M}^S_{\mathsf{eng}} \oplus \mathcal{M}^S_{\mathsf{tok}}$ as our model. Full evaluation is shown in Figure 4.3.

Continuing in this manner has allowed us to show that all Toki Pona sentences that can be represented in $(\text{SM})^2\text{L}$ have a lossless translation into English, indicating that English is Toki Pona-complete, as we hypothesized. The proof of this is left as an exercise for the reader.

5 Conclusions

Our research suggests the possibility of semantic complexity classes to create a well-defined notion of completeness as a relation between two languages, particularly a natural language and a constructed minimal language. Such classes can be formally analyzed and evaluated with the formal semantic definitions of translatability, namely evaluating whether $[\![\ulcorner s' \urcorner \wedge C \rightarrow \ulcorner s \urcorner]\!]^{\mathcal{M},g,w}$ is a tautology for some context C of $O(n)$ size, or only for a larger context. This reflects intuitive judgements, such as the 20-questions test to include simple inflection and the most basic cases of polysemy or homonymy when translating individual symbols, and expands them to apply to strings and other compound expressions via formal semantic compositionality. If for all sentences $s \in B$, there exist translations $s' \in A$ such that a context C of $O(n)$ length can be constructed to make $[\![\ulcorner s' \urcorner \wedge C \rightarrow \ulcorner s \urcorner]\!]^{\mathcal{M},g,w}$ a tautology, then we have that $A \geq_S B$, meaning that A is semantically complete with respect to B, or simply B-complete. If for some sentence in B, there is no sentence in A for which a linear-size context can be constructed to form a tautology, then A is not B-complete.

We specifically explored the example of Toki Pona as a basis for semantic completeness and laid out the beginnings of a proof to show that English is Toki Pona-complete. These results provide a significant first step towards evaluating the semantic complexity of human languages as a whole. Hopefully, future research can determine whether other languages are Toki Pona-complete, and perhaps evaluate the completeness of natural languages with respect to other minimal languages aside from Toki Pona. The goal is to ultimately guide semanticists to the holy grail of natural semantic metalanguage, or perhaps discover that NSM is impossible. It is still too early to say how this will turn out.

The current study was limited in that only $(\text{SM})^2\text{L}$ was considered in the construction of translatability metrics, and other varieties of MSML were not evaluated. $(\text{SM})^2\text{L}$ alone may not be sufficient to fully analyze Toki Pona and its translatability into other languages, so future research may consider expanding to other varieties of MSML such as Event Semantics, which may be a more accurate representation of the semantic space of Toki Pona. 'moku' used as a verb may not be accurately translated as describing the subject $\lambda x.\mathsf{moku}'_H(x)$ as in $(\text{SM})^2\text{L}$, but rather as describing an event in which the subject participated: maybe $\lambda x.\lambda e.\mathsf{moku}'_H(e) \wedge \text{AGENT}(e)(x)$ to represent that the event is 'moku' while the subject takes the theta-role of agent for that event[14]. This may provide for a deeper analysis than $(\text{SM})^2\text{L}$, which we utilized due to space and time limitations.

In summary, our research serves as a proof-of-concept that a minimal, semantically complete language is possible, as Toki Pona fulfills these properties. Future research can perform the necessary minimaxing to search for a sufficiently robust NSM, and discover any potential limitations to NSM, which will undoubtedly provide unique insight into the fundamental limits of human language an understanding. Unless of course GPT 5 in its infinite wisdom just gives us a perfect, general-purpose NSM from a black box.

Let g, w be arbitrary under $\mathcal{M} = \mathcal{M}^S_{\mathsf{eng}} \oplus \mathcal{M}^S_{\mathsf{tok}}$.

$$\llbracket \ulcorner s' \urcorner \wedge C \rightarrow \ulcorner s \urcorner \rrbracket^{\mathcal{M},g,w} \stackrel{?}{=} \top$$

$$\longmapsto \llbracket (\mathsf{bark}'_{\mathsf{eng}}(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x))) \wedge (\neg \text{TREE.PART}'(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x))) \rightarrow (\exists x.\mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x)) \rrbracket^{\mathcal{M},g,w} \stackrel{?}{=} \top$$

$$\longmapsto (\llbracket \exists x.\mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g,w} \textbf{ or } \,!\, \llbracket (\mathsf{bark}'_{\mathsf{eng}}(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x))) \wedge (\neg \text{TREE.PART}'(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x))) \rrbracket^{\mathcal{M},g,w}) \stackrel{?}{=} \top$$

$$\longmapsto \llbracket \exists x.\mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g,w} \stackrel{?}{=} \top \textbf{ or } (\,!\, \llbracket \mathsf{bark}'_{\mathsf{eng}}(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x)) \rrbracket^{\mathcal{M},g,w} \textbf{ or } \,!\, \llbracket \neg \text{TREE.PART}'(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x)) \rrbracket^{\mathcal{M},g,w}) \stackrel{?}{=} \top$$

$$\longmapsto \llbracket \exists x.\mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g,w} \stackrel{?}{=} \top \textbf{ or } \llbracket \mathsf{bark}'_{\mathsf{eng}}(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x)) \rrbracket^{\mathcal{M},g,w} \stackrel{?}{=} \bot \textbf{ or } \llbracket \neg \text{TREE.PART}'(\iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x)) \rrbracket^{\mathcal{M},g,w} \stackrel{?}{=} \bot$$

$$\longmapsto \mathtt{some}(D_e\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } \llbracket \mathsf{bark}'_{\mathsf{eng}} \rrbracket^{\mathcal{M},g,w}(\llbracket \iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x) \rrbracket^{\mathcal{M},g,w}) \stackrel{?}{=} \bot \textbf{ or } \llbracket \text{TREE.PART}' \rrbracket^{\mathcal{M},g,w}(\llbracket \iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x) \rrbracket^{\mathcal{M},g,w}) \stackrel{?}{=} \top$$

Let $\text{DOG} = \llbracket \iota x.\mathsf{dog}'_{\mathsf{eng}}(x) \wedge \mathsf{little}'_{\mathsf{eng}}(x) \rrbracket^{\mathcal{M},g,w}$.

$$\longmapsto \mathtt{some}(D_e\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } \llbracket \mathsf{bark}'_{\mathsf{eng}} \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \bot \textbf{ or } \llbracket \text{TREE.PART}' \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \top$$

$$\longmapsto \mathtt{some}(D_e \setminus \{\text{DOG}\}\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } \llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w} \stackrel{?}{=} \top \textbf{ or } \llbracket \mathsf{bark}'_{\mathsf{eng}} \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \bot \textbf{ or } \llbracket \text{TREE.PART}' \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \top$$

$$\longmapsto \mathtt{some}(D_e \setminus \{\text{DOG}\}\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } \llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w} \textbf{ and } \llbracket \mathsf{lili}'_{M,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w} \textbf{ and } \llbracket \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w} \stackrel{?}{=} \top \textbf{ or } \llbracket \mathsf{bark}'_{\mathsf{eng}} \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \bot \textbf{ or } \llbracket \text{TREE.PART}' \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \top$$

$$\longmapsto \mathtt{some}(D_e \setminus \{\text{DOG}\}\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } (\llbracket \mathsf{soweli}'_{H,\mathsf{tok}} \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w}(\mathtt{get}[g \otimes \text{DOG}/x](x)) \stackrel{?}{=} \top \textbf{ and } \llbracket \mathsf{lili}'_{M,\mathsf{tok}} \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w}(\mathtt{get}[g \otimes \text{DOG}/x](x)) \stackrel{?}{=} \top \textbf{ and } \llbracket \mathsf{kalama}'_{H,\mathsf{tok}} \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w}(\mathtt{get}[g \otimes \text{DOG}/x](x)) \stackrel{?}{=} \top) \textbf{ or } \llbracket \mathsf{bark}'_{\mathsf{eng}} \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \bot \textbf{ or } \llbracket \text{TREE.PART}' \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \top$$

$$\longmapsto \mathtt{some}(D_e \setminus \{\text{DOG}\}\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } (\top \textbf{ and } \top \textbf{ and } \llbracket \mathsf{kalama}'_{H,\mathsf{tok}} \rrbracket^{\mathcal{M},g\otimes \text{DOG}/x,w}(\text{DOG}) \stackrel{?}{=} \top) \textbf{ or } \llbracket \mathsf{bark}'_{\mathsf{eng}} \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \bot \textbf{ or } \llbracket \text{TREE.PART}' \rrbracket^{\mathcal{M},g,w}(\text{DOG}) \stackrel{?}{=} \top \qquad \text{(Meaning postulates 1 and 2 on DOG)}$$

$$\longmapsto \mathtt{some}(D_e \setminus \{\text{DOG}\}\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } (\mathcal{F}(\mathsf{kalama}'_{H,\mathsf{tok}})w)(\text{DOG}) \stackrel{?}{=} \top \textbf{ or } (\mathcal{F}(\mathsf{bark}'_{\mathsf{eng}})w)(\text{DOG}) \stackrel{?}{=} \bot \textbf{ or } (\mathcal{F}(\text{TREE.PART}')w)(\text{DOG}) \stackrel{?}{=} \top$$

$$\longmapsto \mathtt{some}(D_e \setminus \{\text{DOG}\}\,;a.\llbracket \mathsf{soweli}'_{H,\mathsf{tok}}(x) \wedge \mathsf{lili}'_{M,\mathsf{tok}}(x) \wedge \mathsf{kalama}'_{H,\mathsf{tok}}(x) \rrbracket^{\mathcal{M},g\otimes a/x,w}) \stackrel{?}{=} \top \textbf{ or } \top \qquad \text{(Meaning postulate 3 on DOG)}$$

$$\longmapsto \top$$

The tautology holds, therefore the translation is lossless. □

Figure 4.3: Evaluation of lossless translation formula for the sentence in Gloss (7).

References

[1] C. Goddard and A. Wierzbicka, *Meaning and universal grammar : Theory and Empirical Findings /*. Amsterdam: J Benjamins Pub, XV, 334 Seiten. ISBN: 1588112640.

[2] S. Lang, *Toki Pona: The Language of Good*. Tawhid, 2014.

[3] jan Misali, *What is toki pona? (toki pona lesson one)*, Dec. 2021. [Online]. Available: `https://www.youtube.com/watch?v=2EZihKCB9iw`.

[4] u/pisceyo, "Toki pona analysis: Parts of speech," *r/Tokipona - Reddit*, 2019. [Online]. Available: `https://www.reddit.com/r/tokipona/comments/df0zbi/toki_pona_analysis_parts_of_speech/`.

[5] Wikipedia Contributors, *Sisu — Wikipedia, the free encyclopedia*, Jan. 2024. [Online]. Available: `https://en.wikipedia.org/wiki/Sisu`.

[6] T. Scott, *The sentences computers can't understand, but humans can*, Feb. 2020. [Online]. Available: `https://www.youtube.com/watch?v=m3vIEKWrP9Q`.

[7] D. Swanson, *Kay(f)bop(t) babel text*, Dec. 2015. [Online]. Available: `https://crazyninjageeks.wordpress.com/2015/12/29/kayfbopt-babel-text/`.

[8] D. Swanson, *Kay(f)bop(t) conjugations*, Nov. 2015. [Online]. Available: `https://crazyninjageeks.wordpress.com/2015/11/29/kayfbopt-conjugations/`.

[9] This was once revealed to me in a dream.

[10] jan Misali, *Conlang critic episode eight: Kay(f)bop(t)*, Nov. 2016. [Online]. Available: `https://www.youtube.com/watch?v=XhxjI3t6Ht4`.

[11] R. Montague and R. H. Thomason, *Formal philosophy: Selected papers of Richard Montague*. Yale University Press, 1974.

[12] R. Harper, "Chapter 35: Assignable references," in *Practical Foundations for programming languages*. Cambridge University Press, 2016, pp. 323–332.

[13] T. M. V. Janssen and T. E. Zimmermann, *Montague semantics*, Apr. 2021. [Online]. Available: `https://plato.stanford.edu/entries/montague-semantics/`.

[14] T. Werner, "Theta roles and event semantics," Class lecture from 80-381: Meaning in Language, Nov. 2023.

[15] M. Erdmann and D. Kaynar, *Functional programming, spring 2024*, P. Morphism, Ed., Jan. 2024. [Online]. Available: `https://www.cs.cmu.edu/~15150/`.

[16] jan Tamalu, *Results of the 2021 toki pona census*, Jun. 2021. [Online]. Available: `https://tokiponacensus.github.io/results/`.