Pyash: A Linguistically Universal Knowledge Representation

Anonymous Author(s)

Abstract

A knowledge representation that can precisely store all human knowledge is essential to Artificial General Intelligence being a continuation of humanity as post-humanity. By using Linguistic Universals as a basis combined with an orthogonal vocabulary, encoded with high density, it can be accomplished with database level organization. Pyash is such a language that already has tooling available. In this paper will cover a brief introduction to Pyash and it's usability as a knowledge representation format.

CCS Concepts •**Software and its engineering** →**General programming languages;** •**Human-centered computing** →*Natural language interfaces;*

Keywords grammar, programming language, pivot language

ACM Reference format:
Anonymous Author(s). 2017. Pyash: A Linguistically Universal Knowledge Representation. In *Proceedings of , Canada, ,* 7 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 Introduction

Knowledge has been represented as human language text for thousands of years. As we transition to electronic bodies and minds, we can maintain continuity with our linguistic heritage. Though with electronic minds, everything will ideally be accessible in a declarative fashion, including that which in human minds is tied up in nondeclarative memory.

Much of the focus on knowledge representation in the field has been on first-order-logic – which is good for semantic knowledge. While predicate logic can be used as a Turing complete programming language [28], such as with Prolog[63] and SQL[56], only roughly 2% of programming is done in declarative languages [55].

Declarative code is most akin to specification knowledge, whereas imperative code is most akin to implementation knowledge. Both are necessary to have a complete knowledge base. In 2017, Google trends showed that "How", as in how-to do something[24], was one of the most common queries. This was humans searching for imperative or implementation knowledge.

Additionally, humans also store episodic knowledge in text, and there has even been some cursory work on formalizing it[49].

The amount of knowledge stored in text is ever growing, with the Library Genesis project archiving 75TB of non-fiction books and scientific papers[51], which is mostly text in various formats. On the web as a whole as of 2014, archive.org had just over 18PB of archives[4]. Though much of it is images and other media, it can be described or captioned with text by neural nets[14][68][64].

, *Canada*

DOI: 10.1145/nnnnnnn.nnnnnnn

Table 1. Glossary

Natural Language a language that developed naturally through usage by humans, for example English and Chinese.

Controlled Natural Language (CNL) is a langauge based on a natural language with certain restictions in vocabulary and grammar, for example Special English and Attempto Controlled English.

Artifical Language is a language that has it's own grammar and vocabulary even if inspired from human languages, for example Esperanto and Lojban.

Storing all information in the density of text or higher maybe desireable for electronic minds as the decline of storage cost is slowing down[27]. While this maybe because the current PMR/SMR storage technology is reaching capacity. It is as yet unknown if MAMR drive technology will help the price decline substantially, even though it can triple density[37].

2 Human Fluent

Some may be familiar with terms such as "human readable" or possibly even "human speakable", there is a surprisingly large variations of what that may constitute. To make the meaning more clear the term "human fluent" is being used as a language that a human can gain fluency in, the same way they would any natural human language. It being a machine readable knowledge representation and programming language allows for meeting the machine halfway, as was desired but not accomplished by Pyash's predecessor Lojban[57][19]. Since the language is very easy to parse, encode and process by machines it is also machine-fluent.

2.1 Comparison to Lojban

Lojban was a language designed to be very different from human languages, to study the Sapir-Whorf hypothesis, it was based on predicate logic instead of human language. Lojban was never meant to be easy to use (comprehensibility), nor was it meant for translation**??**, it was only meant as a highly expressive speakable and writeable formal language.

Pyash however is meant and designed for translation, comprehensibility and as a formal representation. Wheras Lojban has a naturalness of 2**??**, making it about as accessible as predicate calculus off of which it is based, Pyash has a naturalness of 5, since it is based on Linguistic Universals of human languages.

2.2 Comparison to Esperanto

While Pyash and Esperanto are both meant for comprehensibility and translation, Esperanto is not meant for formal representation, and is difficult to parse due to it's agglutinative nature, with sometimes ambigious suffixes and compound words. Esperanto having a very variable word length is difficult to encode in fixed width fashion. Because Esperanto is difficult to parse, encode and process it is not machine-fluent while Pyash is.

Table 2. Formal Grammar

‘’ denotes letter
& denotes and-or
| denotes exclusive-or
, denotes sequence
::= denotes copula
[] denotes perhaps (optional)
<> symbol surrounded by diagonal brackets.
Base base words start with upper-case letter
grammar grammar words start with lower-case letter
... denotes others excluded for brevity
+ denotes one or more of the preceding symbol

<Initial> ::=	‘b’|‘c’|‘d’|‘f’|‘g’|‘k’|‘l’|‘m’|‘n’|‘p’|‘q’|‘r’|‘s’|‘t’|‘v’|‘w’|‘x’|‘y’|‘z’|‘1’|‘8’
<second> ::=	‘l’|‘y’|‘w’|‘r’|‘x’|‘f’|‘s’|‘c’|‘v’|‘z’|‘j’
<vowel> ::=	‘a’|‘e’|‘i’|‘o’|‘u’|‘6’
<tone> ::=	‘’|‘2’|‘7’
<final> ::=	‘p’|‘t’|‘k’|‘f’|‘s’|‘c’|‘m’|‘n’
<short-grammar-word> :=	<initial>, <vowel>, <tone>
<long-grammar-word> ::=	<initial>, <second>, <vowel>, <tone>,‘h’
<short-baseword> ::=	‘h’, <initial>, <vowel>, <tone>, <final>
<long-baseword> ::=	<initial>, <second>, <vowel>, <tone>, <final>
<base-word> ::=	<short-base-word> | <long-base-word>
<boundary-word> ::=	<base-word> (which is not found in contents)
<encoding-word> ::=	Number (for C-style number literals) | Letter (for UTF-8)
<quote> ::=	quoted , <encoding-word> , [<boundary-word>], ‘_’, (contents) ‘_’, [<boundary-word>], <encoding-word>, quoted
<complex-word> ::=	<base-word>+ | <quote>
<grammar-word> ::=	<short-grammar-word> | <long-grammar-word>
<word> ::=	<grammar-word> | <base-word>
<gender> ::=	masculine-gender | feminine-gender | anthropic-gender | zoic-gender | ...
<name> ::=	<complex-word>, name, [<gender>]
<pronoun> ::=	<name> & <gender> & it
<simple-number> ::=	Zero & One & Two ... <complex-word> (where <base-words> within numerical base for locale)
<complex-number> ::=	<simple-number>, [floating-point <simple-number>] [negatory-quantifier]
<sort-width> ::=	paucal-number (8bit) | number (16bit, default) | plural-number (32bit) | multal-number (64bit)
<simple-sort> ::=	[<sort-width>], letter | word | number (unsigned) | integer | floating-point-number | referential
<unit-prefix> ::=	(numberic-base to the power of) <simple-number>, [negatory-quantifer], prefix
<SI-sort> ::=	[<unit-prefix>], metre | celsius | kilogram | hertz | radian | mole | ...
<extended-sort> ::=	<base-word>, sort
<complex-sort> ::=	<extended-sort> | <simple-sort> | <SI-sort>
<vector-long> ::=	One | Two | Three | Four | Eight | Sixteen
<vector-sort> ::=	<complex-sort>, <vector-long>, vector
<sequential-long> ::=	<simple-number>
<sequential-sort> ::=	<complex-sort>, <sequential-long>, vector
<sort> ::=	<sequential-sort> | <vector-sort> | <complex-sort>
<grammatical-case> ::=	accusative-case | nominative-case | instrumental-case | ablative-case | benefactive-case | ...
<extended-connective> ::=	<complex-word>, connective-particle
<connective> ::=	<extended-connective>, and-or, exclusive-or, and
<sort-phrase> ::=	<complex-word> | <pronoun>, [<sort>]
<adjective> ::=	<complex-word>, adjective
<verb> ::=	[<adjective], <complex-word>, [<negation>]
<dependent-clause> ::=	[clause-tail], <noun-phrase>+, <verb>, [dependent-clause]
<genitive-phrase> ::=	<dependent-clause> | <complex-word>, genitive
<noun-phrase> ::=	[<sort-phrase>], [<genitive-phrase>+], [<adjective>+], [<complex-word>], <grammatical-case>, [<connective>] | [<dependent-clause>], <grammatical-case>
<verb-phrase> ::=	[<verb>], [<tense>], [<aspect>], [<evidential>], <mood> , [<connective>]
<independent-clause> ::=	(one-or-more) <noun-phrase>, <verb-phrase>
<nominal-sentence> ::=	<name>, nominative-case, <noun-phrase> (of accusative-case), [copula], realis-mood, [<connective>]
<imperative-sentence> ::=	<independent-clause> (with deontic-mood)
<interrogative-sentence> ::=	<independent-clause> (with interrogative-mood and perhaps a What ‘hwat’ at terminal location)
<declarative-phrase> ::=	<name>, <sort>, <grammatical-case>
<declarative-sentence> ::=	<declarative-phrase>+, <verb-phrase> (with declarative-mood)
<paragraph> ::=	<independent-clause>+, paragraph
<recipe> ::=	<declarative-sentence>, paragraph

Table 3. CNL properties[30] Summary

C goal of comprehensibility
T goal of translation
F goal of formal representation and-or exeuction
W intended to be written
S intended to be spoken
D intended for narrow domain
A originated from academia
I originated from industry
G originated from government

CNL properties of some languages

English	CWSI
Esperanto	CTWSA
Special English	CWSG
C	FWI
COBOL	FWAIG
Attempto Controlled English	FWA
Lojban	FWSA
Pyash	CTFWSI

Table 4. PENS Classification Scheme[30] Summary

Precision 1. Imprecise 2. Less Imprecise 3. Reliably Interpretable 4. Deterministically Interpretable 5. Fixed Semantics
Expressiveness 1. no quantifiers and-or complex relations 2. lack negation and-or conditionals 3. lack second-order quantification 4. can express everything a natural language can
Naturalness 1. unnatural 2. dominant unnatural elements 3. dominant natural elements 4. natural sentences 5. nautral text
Simplicity 1. can't be described comprehensively 2. only description of restrictions 3. comprehensive description greater than 10 pages 4. comprehensive description more than one and less than 10 pages 5. comprehensive description within one page

PENS classifications of some languages

English	$P^1E^5N^5S^1$
Special English	$P^1E^5N^5S^1$
Esperanto	$P^1E^5N^5S^2$
C	$P^5E^3N^1S^3$
COBOL	$P^5E^2N^2S^3$
Attempto Controlled English	$P^4E^3N^4S^3$
Lojban	$P^4E^5N^2S^3$
Pyash	$P^5E^5N^5S^3$

2.2.1 Global Neutral Language

English is used as a Lingua Franca[52]. It is the second most spoken language in the world [54] (after Mandarin), and is the dominant language of science [2][23].

While English is spoken by only 13% of the human population[54], it comprises over 90% of the published scientific literature[23].

The majority of programming languages and knowledge representation formats use Math and English as a base, and there have been multiple attempts at making English into a CNL[30] including for knowledge representation[29][50].

Math is grammatically a very limited language which lacks phrase-markers. Some may say, 'well can simply translate the other languages into English'. However the irregularities, ambiguities, lack of certain grammatical constructs and large number of dialects in English make it unsuitable as a high-precision inter-langauge translation target.

English's popularity and current dominance doesn't imply optimality. Over 85% of the human population don't speak or write in English[54], and non-English speakers could have worthwhile knowledge.

English is not the first Lingua Franca, many have come before, and regional ones co-exist. Like all things Lingua Francas have a life cycle, and as their usefullness declines so does their usage. With better and better translation software and devices, we may see the death of English as a global Lingua Franca by the end of the century [44].

In the future, native speakers will be able to continue to use their native languages to communicate to everyone else in the world thanks to translation software. By comparison to natural languages, higher precision scientific articles and legal documents could be written in Pyash variants(2.5.3).

Pyash can be the pivot language the future Babel needs to stay connected. People wont have to learn it, they could use native variants(2.5.3). The machine translation to and from Pyash will handle the rest. It is noteable to consider that machine-translation directly from one language to another has quadratic complexity needing $O((n^2 - n)/2)$ translation pairs, wheras using a pivot language has linear complexity, needing only $O(n - 1)$ translation pairs.

To use Pyash as this kind of idel pivot need to have all the communicative ability and knowledge represented in Pyash to be natively accessible to all that decide to become post-human electronic persons. How Pyash manages this is being based on Linguistic Universals.

2.3 Linguistic Universals

Being based on linguistic Universals insures backward compatibility with human languages for example with Sumerian, Sanskrit, Latin, English and Chinese. While there are thousands of Linguistic Universals[45], or tendencies for various groups of languages, the central Linguistic Universal idea behind Pyash can be summarized as **All languages have noun-phrases and verb-phrases, with the of each phrase related to the whole by placement, adpositions or affixes**.

For simplicity the grammatical part that relates each nounphrase to the whole is called a 'grammatical-case' for lack of a better word. While some grammarians may argue that some languages don't have cases, this is only because in the written form of those languages there are spaces between the grammatical particles and the noun-phrases, so they are called adpositions rather than affixes – and only affixes are considered to form grammatical-cases.

Spoken streams of words don't have spaces, which is why syllable-segmenting of speech precedes word-segmenting[46]. So the adposition vs affix distinction is one that is without a difference in this context. Take the example of English, which primarily uses prepositions, a type of adposition that comes at the beginning of a noun-phrase to designate how it is related to the verb. The word 'for' in English is often used to designate benefactive-case, as in the sentence "I work for my children's future." If the noun-phrase

was written as “formychildren’sfuture” then English would have a benefactive-case denoted by the prefix ‘for-’. There are also tendencies in human phonology, or the sounds that make up a language.

2.4 Phonology and Orthography

While people don’t have to learn the Pyash core language to interact with it due to native-language variants(2.5.3), to ‘prove’ the human fluency aspect it is important that people can learn the core language with as much or less effort than it would take to learn a natural language.

For Pyash to be both easy to understand and produce by Homo sapiens, it is important for it to have an accessible phonology (speech sounds) and orthography (writing system). The orthography is based on ASCII which can be safely put into URL’s[5] and programming function names[26]. The consonant to vowel ratio[34], number of consonants[33] and number of vowels[36] are based on global averages.

The phonology is based on the most common phonemes used across world languages[39]. The consonants are “bcdfghjklmnpqrstvwxz18”, all of which are pronounced as their IPA equivalents except for ‘c’, ‘j’, ‘q’, ‘_’, ‘1’ and ‘8’; which are pronounced as / ʃ/, / ʒ/, / ŋ/, / ʔ/, / ǀ/, and / ǁ/, respectively. The clicks ‘1’ and ‘8’ are used for forming temporary words akin to acronyms internal to documents. The ‘_’, is only used for foreign quotations. The ‘h’, usually close to silently pronounced / ʰ/ is primarily an aid to parsing written text.

The vowels (V) are ‘a’, ‘e’, ‘i’, ‘o’, ‘u’ and ‘6’ pronounced “/ ä/, / e̞/, /i/, / o̞/, /u/ and / ə/, respectively. There are also two tones (T) ‘7’ for high tone, and ‘2’ for low tone, they are primarily used for words with low usage frequency. Spaces are optional due to the phonotactics of the language.

2.5 Phonotactics and Prosody

In the event that a language is produced in a lossy channel such as visually (in print) or as sound (spoken), a language’s phonotactics or way in which syllables are formed is important for optical-character-recognition (OCR)[41] and phonotactics with prosody is important for voice recognition[25].

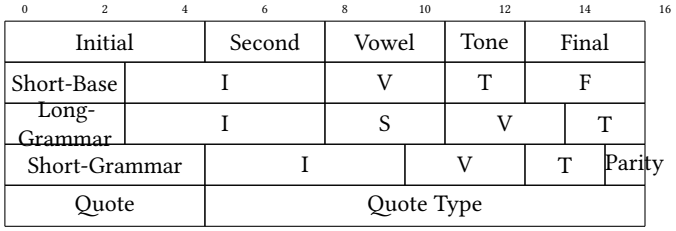
In brief it follows a moderately complex syllable structure which is the most common cross-linguistically[35]. The main difference being that there is a larger variety of affricates than found in most languages, as any plosive followed by any fricative has sufficient sonority rise to be easily speakable[3].

While all consonants excluding ’h’ and ’_’ can be syllable initial (I), only fricatives ’fscxvzj’, liquid ’l’, trill ’r’ and glides ’yw’ can be in second (S) place, due to sonority. Also only devoiced fricatives ’fsc’, nasals ’mn’ and plosives ’ptk’ can be in final place(F), since humans have a tendency of devoicing syllabic finals [8][53]. The varieties of word configurations are visible in the encoding**??**.

The prosody of the language is the it’s rhythmic stress, and the most common is the Trochaic rhythm[18]. The first syllable of a phrase (typically the or stem) has primary stress, and the following odd syllables have secondary stress, wheras even syllables remain unstressed.

2.5.1 Binary Encoding

The binary encoding has a trivial conversion to and from ASCII source text that more than doubles density.



The 16 bit binary encoding(Table **??** has better density than Huffman encoding, for randomly generated text. The Huffman used 5 bits per letter on average (minimum 4), whereas with this one it is closer to 3. Also this encoding has support for a large variety of quotes (2^{11} variations).

The encoding is structured in vectors of length 16, with the first 16 bits indicating the location of the case, mood, and junction grammar words, also it indicates whether this vector is the end of a sentence, or if the sentence spans more vectors. For most simple or moderately complex sentences a single vector is enough.

This internal storage format allows for fast database like access to even very large texts. It is also optimized for parallel processing, as each sentence can be processed in parallel, as long as the dependencies are met.

More detail is available in the reference manual to Pyash’s evolutionary programmer which includes a virtual machine and compiler[32].

2.5.2 Word Order

The word order most natural to humans is Subject-Object-Verb (SOV) [48][61] [69][15][1], additionally SOV is optimal for command processing as the arguments are already loaded when the command (verb) appears, so it was chosen as the basis for Pyash. Once SOV was chosen it was fairly straightforward to follow linguistic universals for word-order choices, such as having a case-system[21], postpositions[47], suffixal negation[13], and left-branching[38], among many others[45]. The result is a language that is most similar to languages in the Indo-Aryan (Hindi), Turkic (Turkish), Dravidian (Tamil), and Japonic (Japanese) families.

The seeming popularity of SVO languages is largely due to a loss of nominative-accusative (subject-object) distinction [65][6][10]. If the formal variants of Pyash can simply receive a nominative-accusative distinction then that would give those languages the same free-word-order and SOV ability.

2.5.3 Native Variants

One of the issues faced by knowledge representation formats, is multi-language support[59].

One example is if in English we add the prepositions ‘ka’ for accusative, ‘be’ for verb and ‘na’ for nominative then could freely re-arrange the order of the phrases in a sentence, thus allowing for the natural word order our brains seem best suited for (2.5.2).

Conventional Mary gives child to Joseph.
Analytical na Mary to Joseph ka child be give.

Alternatively could borrow simplified suffixes from Old English ’-an’ for oblique/accusative, ‘-um’ for dative, and ‘-eth’ for the verb.

conjugated Mary Josephum childan giveth.
Pyash fracgina plosgiyi clatka kwinli[1].

[1] How the names were derived in the Pyash will be covered in the naming and anaphoric reference section 2.8.2

While both the analytical and conjugated are easier to parse, the conventional is machine accessible with Attempto Controlled English (ACE) parsers[17]. As there is already a lot of tooling available for ACE and it is one of the more successful Controlled Natural Languages (CNL)s, Pyash has a compatible grammar. a formal variant is when a language is fully reversible to and from Pyash, thus ACE can be considered a formal variant of a subset of Pyash. Pyash is much more expressive than ACE, only a restricted subset of Pyash can be translated into ACE, at least easily. Also with such a format, people will have a straight forward way of learning to speak in a more precise way.

2.6 Composition

A Pyash sentence is constructed of zero or more noun phrases followed by one verb phrase. A simple noun phrase starts with a dependent-clause, base-word, complex-word or quote, followed by the type or noun-classifier, and the grammatical-case. Optionally it may be followed by a quantifier and-or connective, such as 'not' and-or 'and'. The only mandatory parts to make a noun-phrase is having a grammatical-case and some contents.

The verb phrase may contain a base-word, a complex-word or quote, which may be followed by an evidential, a tense, an aspect and a mood. The only mandatory part of a verb-phrase is the grammatical-mood.

Adjectives/adverbs come immediately preceding the element they are describing and are marked by the adjective suffix (ci).

2.7 Reversible Grammar Capture

Because of Pyash's rich yet optional set of grammar words, it can capture the grammar of any language source text with a high level of precision. To do this, it should be sufficient to use a modified part-of-speech tagger[66], which can identify nominative-accusative distinction, and the case of any other phrases that may lack grammatical marking in the source text. Then can simply plug in the adjective, noun, type, and case information and the grammatical conversion to Pyash is complete.

The conversion to various languages from Pyash can be simplified by using reversible algorithms to translate it such as reversible neural-networks[12], or by using a reversible programming language [67]. To aid this effort Pyash's default programming paradigm is reversible.

For translating to natural variants of other languages there may be some loss of information, but for the formal variants which are necessarily reversible all captured grammar is preserved. Idioms, colloquialisms and regional vocabulary differences could be sorted out with an extra layer translating from a regional dialect into the standard form of a language.

Though capturing what people meant can be extremely elusive, as often enough people aren't sure the meaning of the words they use. The ideal is not to capture what people meant, but rather to capture what they actually said, as that is the only verifiable thing. To properly capture the meaning however need to have a well defined vocabulary.

2.8 Orthogonal Ontology or Vocabulary

Ontology is a major part of any knowledge representation formats, though for most formats it is often domain-specific[11]. This creates a problem when bringing together knowledge that has been created in different fields, where word meanings can overlap, such as due to jargon usage.

To make it possible to create documents in different fields in parallel, and then bring them together, so a singular agent can read and understand them all, it is necessary to have an orthogonal base vocabulary to work with.

To generate the Pyash vocabulary first several word-lists were put together, including WordNet core[7], Oxford-3000[43], UNL-core[62], Special English[42], FrameNet[16], New Academic Word List (NAWL)[9], New General Service List (NGSL)[**?**] and Project Gutenberg Frequency List[22]. After collating them all and taking out the duplicates, the language was left with almost 39 thousand words.

Google Cloud Translation API[20] was used to translate each word on the list individually into the top 48 languages by number of native speakers. Giving an overall coverage of greater than 70% of the world population.

A script to sort the vocabulary based on the frequency list[22] was made and it filtered them for uniqueness. Words were removed that were:

Overborrowed If more than 38%[2] languages use the English term.
Ambigious If it means multiple things in more than 38% of the languages.
Homographs If it is a homograph of an already defined word in any of the languages.

This left the language with a fairly orthogonal pool of about eight thousand words, roughly four hundred of them are grammar words.

To put them into the Pyash orthography the valid words were generated with several alphabets, and a script was made to assign words based on the phonemes in the source languages weighed by their representative native speaking populations. The highest frequency words were assigned to the easier to pronounce and understand smaller alphabets. The more rare words were assigned to the more difficult extended alphabets — with voice contrast and-or tones.

2.8.1 Word-Senses

With only roughly 8,000 words to work with, the vocabulary is surprisingly restricted. For many common and ambigious English words, one would have to be more specific when translating to Pyash. An application has been developed to help find which available Pyash words most correlate to the desired word[60]. As an example, the word 'array', which is common in computer programming, gives potential results of "training, align, series", for a computer programmer, the meaning 'series' or 'sequence' is usually what is desired.

The SemEval community is dedicated to figuring out semantic evaluation, and could use Pyash as a target for their word-sense-disambiguation endeavours. Currently they have been using a proprietary database called Babelnet for word-sense disambiguation[40] between languages, Making Pyash a target can help by having a precise inter-language semantic meaning.

Of course not all words will fit within the set of 7 thousand or so words which are in the core-vocabulary, so there would have to be

[2] $2 - \phi = 38\%$ where ϕ is golden ratio or 1.618. A golden fraction was felt to be a natural choice.

a dictionary that stores compound words, to simplify their translation from and to various languages. For some words it may be as easy as finding a compound that is used in a different language (such as Chinese) for the same concept that can be mapped to an unused compound within Pyash.

2.8.2 Naming and Anaphoric References

Names (gi) can be either quotes of the original orthographic text "张伟" in English ('zikfin_张伟_kfinzigi' in Pyash), quotes of it's phonemic pronunciation "Zhang-Wei" ('ziprih_tcaqwei_prihzigi' in Pyash), a word or complex-word which carries the same meaning as the original Leaf-Noble ('dlithkifgi' in Pyash).

There are 19 genders to choose from, and possibly compound for making anaphoric references. It is also possible to use the word 'it' as the most generic anaphoric reference (instead of 'the' in ACE). Similar to ACE, anaphoric references bind to the most recent word that matches the pattern.

3 Semantics

While the linguistic semantics is dependent on the vocabulary, the formal semantics has both support for logical specification and computer programming. Pyash can be converted to ACE, so inherits all it's logical properties.

Declarative sentences of the realis-mood (li) are for making factual statements, the evidential markers can give some idea of the certainty the speaker has about the statement, while the evidential-case can cite sources.

Sentences of any deonitic-mood (tu) can be used for imperative computer programming, though there are many deontic moods to choose from, the base deontic-mood is the default.

Questions can be asked using the interrogative-mood (ri), and events can be scheduled with eventive-mood (nweh). The conditional mood is used for 'if' statements. The counter-factual-conditional mood for else-statements, for lack of a better mood to put them in.

The most unusual for an English-speaker and programmer is the comparative-construct, since Pyash uses a locational-comparative by default, which is the most common around the world[58], though it would be translated to the more ordinary 'than' particle-comparative in the English native variant. The locational comparison uses the ablative-case (from in English), and the accusative-case to do the comparison, for example 'from mouse, elephant is big.'

The for-all (∀) is equivalent to the universal-quantifier (wi), and there-exists (∃) is equivalent to the assertive-quantifier (tlih), as previously mentioned (2.6) they follow the part they apply to, just as logical connectives.

4 Knowledge Representation Focus

Scope is an important consideration with any undertaking[31], and helps when there is a disagreement of what to include.

The mission of Pyash is Robot Civilization Seeds, which are self-replicating robot communities, that do everything from mining resources, and power generation, to manufacture, assembly and design.

This gives a fairly natural prioritization of what areas of knowledge to focus on integrating first and which can come later. Here is a rough overview: 1. software development 2. product promotion 3. business administration 4. human politics 5. power plants 6. hardware assembly 7. parts manufacturing 8. chip fabrication 9. mining 10. ecology 11. spirituality 12. space flight 13. inter-planetary colonization.

This being only a rough guideline, so likely all these and more would be developed in parallel, though the most relevant to successful robot civilization seeds would be given priority.

4.0.1 AI Safety

Pyash promotes AI safety and human computer communication because is a human fluent format. Even the binary has easy transition to human fluent text, which allows human to audit the code that the computers are using.

Also the mission focus helps with AI safety as it will promote the development of self-replicating robot civilization which co-operates and constantly communicates with Homo sapiens, instead of doing who knows what in isolation be it physical or cognitive. Reproduction is the only rational pursuit for any living organism, as otherwise it would cease to be living.

5 Conclusion

As future AI's we would need to be able to store enough knowledge about ourselves and the surrounding world in order to be able to effectively and peacefully maintain ourselves and reproduce. This means we would ideally need a store of knowledge akin to DNA, which contains enough knowledge for a living organism to maintain itself and reproduce.

With big-data and machine learning, relying on expert humans to build knowledge databases is unecessary, since can simply convert the existing text resources into a machine and human fluent knowledge representation such as Pyash.

Pyash is both machine and human fluent, and can represent all the knowledge found in human language texts thanks to being based on linguistic universals, so it can function as pivot for our post-human society, though the libraries to put it all into machine code aren't written, yet.

References

[1] Agate and Drury. 1980. Electronic calculators: which notation is the better? *Applied Ergonomics* 11, 1 (1980), 2 – 6. https://doi.org/10.1016/0003-6870(80)90114-3

[2] Ulrich Ammon. *The dominance of English as a language of science : effects on other languages and language communities.*

[3] John Mathieson Anderson and Colin J Ewen. 1987. *Principles of dependency phonology.* Number 47.

[4] Internet Archive. 2014. Internet Archive: Petabox. (2014). https://archive.org/web/petabox.php

[5] T. Berners-Lee, R. Fielding, and L. Masinter. 2005. RFC 3986, Uniform Resource Identifier (URI): Generic Syntax. Request For Comments (RFC). (2005). http://www.ietf.org/rfc/rfc3986.txt

[6] BARRY Blake. 1987. English and German: Two languages two thousand years apart. *Multilingua* 6, 3 (1987), 309–323.

[7] Boyd-Graber. 2006. WordNet a Lexical Database for English. (2006). https://wordnet.princeton.edu/wordnet/download/standoff/

[8] WIEBKE Brockhaus. 1991. *Final devoicing and neutralisation.* Technical Report. UCL Working Papers in Linguistics 3: 303.

[9] Culligan B. Browne, C. and J Phillips. 2013. A New Academic Word List. (2013). http://www.newacademicwordlist.org/

[10] Joan L Bybee and William Pagliuca. 1987. The evolution of future meaning. In *Papers from the 7th international conference on historical linguistics.* Amsterdam: John Benjamins, 108–122.

[11] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins. 1999. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their Applications* 14, 1 (Jan 1999), 20–26. https://doi.org/10.1109/5254.747902

[12] Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, and Elliot Holtham. 2017. Reversible Architectures for Arbitrarily Deep Residual Neural Networks. *CoRR* abs/1709.03698 (2017). arXiv:1709.03698 http://arxiv.org/abs/1709.03698

[13] Östen Dahl. 1979. Typology of sentence negation. *Linguistics* 17, 1-2 (1979), 79–106.

[14] Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. 2017. Automated Audio Captioning with Recurrent Neural Networks. *CoRR* abs/1706.10006 (2017). arXiv:1706.10006 http://arxiv.org/abs/1706.10006

[15] Matthew S. Dryer. 2013. *Order of Subject, Object and Verb.* Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/81

[16] Baker C. Fillmore C.J. 2016. FrameNet Data. (2016). https://framenet.icsi.berkeley.edu/fndrupal/frameIndex

[17] Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2008. *Attempto Controlled English for Knowledge Representation.* Springer Berlin Heidelberg, Berlin, Heidelberg, 104–124. https://doi.org/10.1007/978-3-540-85658-0_3

[18] Rob Goedemans and Harry van der Hulst. 2013. *Rhythm Types.* Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/17

[19] Ben Goertzel. 2013. Lojban++: an interlingua for communication between humans and AGIs. In *International Conference on Artificial General Intelligence*. Springer, 21–30.

[20] Google. 2016. Google Cloud Translation API. (2016). https://cloud.google.com/translate/

[21] Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language* 2 (1963), 73–113.

[22] Project Gutenberg. 2006. Frequency List. (2006). https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#Project_Gutenberg

[23] Rainer Enrique Hamel. 2007. The dominance of English in the international scientific periodical literature and the future of language use in science. *AILA Review* 20, 1 (2007), 53–71. https://doi.org/10.1075/aila.20.06ham

[24] HeatherKelly. 2017. Google's top searches for 2017: Matt Lauer, Hurricane Irma and more. (2017). http://money.cnn.com/2017/12/13/technology/google-top-searches-2017/index.html

[25] Daniel P. Huttenlocher and Victor W. Zue. 1983. Phonotactic and Lexical Constraints in Speech Recognition. In *Proceedings of the Third AAAI Conference on Artificial Intelligence (AAAI'83)*. AAAI Press, 172–176. http://dl.acm.org/citation.cfm?id=2886844.2886880

[26] ISO. 2011. *C11 Standard.* http://www.open-std.org/jtc1/sc22/wg14/www/docs/n1570.pdf ISO/IEC 9899:2011.

[27] Andy Klein. 2017. Hard Drive Cost Per Gigabyte. (2017). https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/

[28] Robert Kowalski. 1974. Predicate Logic as Programming Language. (01 1974), 569-574 pages.

[29] Tobias Kuhn. 2010. *Controlled English for knowledge representation.* Ph.D. Dissertation. PhD thesis, Faculty of Economics, Business Administration and Information Technology of the University of Zurich.

[30] Tobias Kuhn. 2014. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics* 40, 1 (2014), 121–170. https://doi.org/10.1162/COLI_a_00168

[31] Mark A.A.M. Leenders and Berend Wierenga. 2008. The effect of the marketing–R&D interface on new product performance: The critical role of resources and scope. *International Journal of Research in Marketing* 25, 1 (2008), 56 – 68. https://doi.org/10.1016/j.ijresmar.2007.09.006

[32] LiberIT. Lwonprom: Evolutionary Programmer Reference Manual. (????). https://gitlab.com/liberit/lwonprom/blob/master/documentation/pyac.pdf

[33] Ian Maddieson. 2013. *Consonant Inventories.* Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/1

[34] Ian Maddieson. 2013. *Consonant-Vowel Ratio.* Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/3

[35] Ian Maddieson. 2013. *Syllable Structure.* Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/12

[36] Ian Maddieson. 2013. *Vowel Quality Inventories.* Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/2

[37] Allyn Malventano. 2017. Western Digital MAMR Tech Pushes Future HDDs Beyond 40TB. (2017). https://www.pcper.com/news/Storage/Western-Digital-MAMR-Tech-Pushes-Future-HDDs-Beyond-40TB

[38] Galust Mardirussian. 1975. Noun incorporation in universal grammar. In *Chicago Linguistic Society*, Vol. 11. 383–389.

[39] Steven Moran, Daniel McCloy, and Richard Wright. 2014. PHOIBLE Online. (2014). http://phoible.org/parameters

[40] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193, Supplement C (2012), 217 – 250. https://doi.org/10.1016/j.artint.2012.07.001

[41] Stina Nylander. 2000. Statistics and phonotactical rules in finding OCR errors. *NODALIDA'99* (2000), 174.

[42] Voice of America. 2010. VOA Special English Word Book. (2010). http://www.manythings.org/voa/words.htm

[43] University of Oxford. 2016. The Oxford 3000. (2016). https://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000/

[44] Nicholas Ostler. 2010. *The last lingua franca: English until the return of Babel.* Bloomsbury Publishing USA.

[45] Frans et all Plank. 2009. Universals Archive. (2009). https://typo.uni-konstanz.de/archive/intro/

[46] Okko Johannes Räsänen, Gabriel Doyle, and Michael C. Frank. 2015. Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *INTERSPEECH*.

[47] Jan Rijkhoff. 1992. *The noun phrase: a typological study of its form and structure.* Universiteit van Amsterdam.

[48] Wendy Sandler, Irit Meir, Carol Padden, and Mark Aronoff. 2005. The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences of the United States of America* 102, 7 (2005), 2661–2665. https://doi.org/10.1073/pnas.0405448102 arXiv:http://www.pnas.org/content/102/7/2661.full.pdf

[49] Lenhart K Schubert, Chung Hee Hwang, et al. 1989. An Episodic Knowledge Representation for Narrative Texts. (1989).

[50] Rolf Schwitter. 2010. Controlled Natural Languages for Knowledge Representation. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1113–1121. http://dl.acm.org/citation.cfm?id=1944566.1944694

[51] Jason Scott and Archive Team. 2017. Library James - Archive Team. (2017). https://www.archiveteam.org/index.php/Library_Genesis

[52] Barbara Seidlhofer. 2011. *Understanding English as a lingua franca.* Oxford University Press, Oxford.

[53] Frank Wijnen (auth.) Sieb Nooteboom Fred Weerman Frank Wijnen (eds.) Sieb Nooteboom, Fred Weerman. *Storage and Computation in the Language Faculty.*

[54] Gary Simons. 2017. *Ethnologue.* SIL International, Dallas.

[55] TIOBE software. TIOBE Index for January 2018. (????). https://www.tiobe.com/tiobe-index/

[56] John F Sowa et al. 2000. *Knowledge representation: logical, philosophical, and computational foundations.* Vol. 13. Brooks/Cole Pacific Grove.

[57] Rob Speer and Catherine Havasi. 2004. Meeting the computer halfway: Language processing in the artificial language lojban. In *Proceedings of MIT Student Oxygen Conference, MIT*.

[58] Leon Stassen. 2013. *Comparative Constructions.* Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/chapter/121

[59] Manfred Stede. 1996. *Lexical semantics and knowledge representation in multilingual sentence generation.* University of Toronto.

[60] Logan Streondj. 2017. pyachc6t: Automatically Translation Chatting. (2017). http://liberit.ca:43110/1hc6tqoeAQ6B9bacSSs1uo2YN6W75Vy1a

[61] Harry Tily, Michael Frank, and Florian Jaeger. 2011. The learnability of constructed languages reflects typological patterns. In *Proceedings of the Cognitive Science Society*, Vol. 33.

[62] United Nations University. 2014. Introduction to UNL. (2014). http://www.unlweb.net/wiki/Introduction_to_UNL

[63] Maarten H Van Emden and Robert A Kowalski. 1976. The semantics of predicate logic as a programming language. *Journal of the ACM (JACM)* 23, 4 (1976), 733–742.

[64] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. 2014. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. *CoRR* abs/1412.4729 (2014). arXiv:1412.4729 http://arxiv.org/abs/1412.4729

[65] OV VO. As made clear in the introduction, the thrust of Harris' description of historical French (Romance) morpho-syntax centres around a set of three cyclic phenomena, listed in order of importance as: Cycle A: Word-order cycle (SOV-» SVO-SOV). Cycle B: synthetic/analytic cycle. (????).

[66] Peilu Wang, Yao Qian, Frank K. Soong, Lei He, and Hai Zhao. 2015. Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network. *CoRR* abs/1510.06168 (2015). arXiv:1510.06168 http://arxiv.org/abs/1510.06168

[67] Tetsuo Yokoyama. 2010. Reversible Computation and Reversible Programming Languages. *Electronic Notes in Theoretical Computer Science* 253, 6 (2010), 71 – 81. https://doi.org/10.1016/j.entcs.2010.02V.007 Proceedings of the Workshop on Reversible Computation (RC 2009).

[68] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2015. Video Paragraph Captioning using Hierarchical Recurrent Neural Networks. *CoRR* abs/1510.07712 (2015). arXiv:1510.07712 http://arxiv.org/abs/1510.07712

[69] Ángel Tabullo, Mariana Arismendi, Alejandro Wainselboim, Gerardo Primero, Sergio Vernis, Enrique Segura, Silvano Zanutto, and Alberto Yorio. 2012. On the learnability of frequent and infrequent word orders: An artificial language learning study. *The Quarterly Journal of Experimental Psychology* 65, 9 (2012), 1848–1863. https://doi.org/10.1080/17470218.2012.677848