yod: A library for language generation

Daniel James

April 2017

Chapter 1

Introduction

yod is a library designed to aid in the construction and generation of artificial languages. Artificial languages (also known as constructed languages or 'conlangs') are probably most publicly well known through the popularisation of 'fictional languages' - that is, languages designed for use in works of fiction, usually for adding depth to a fictional world. For example, J. R. R. Tolkien devised the languages of *Sindarin* and *Quenya* for use in his *Lord of the Rings* series. Other popular fictional languages that have reached mainstream awareness include *Star Trek*'s *Klingon* and *Game of Thrones*' *Dothraki*.

The creation of these languages has become an indispensable tool for writers to add depth and character to their worlds. However, constructed languages do not exist solely for creative uses. The language *Esperanto* was developed by its creator L. L. Zamenhof with the goal of being a global language that was easy for people to learn[1]. Now it is reported that as many as 63,000 people worldwide can speak Esperanto to some degree[2]. The existence of Esperanto and other so-called "international auxiliary languages" (languages designed to simplify complication between people of different countries and cultures) shows that the usefulness of constructed languages extends into real-world applications as well as artistic uses.

yod was developed with a focus on artistic languages, and so the goal was to generate a large variety of languages which could suit many different worlds and fictional civilisations. However, due to the hierarchy of rules *yod* uses to build languages, the languages are regular, in contrast to most natural languages, which means generated languages with certain features could be treated as auxiliary languages too.

talk about structure

Chapter 2

Building Words

The 'naive' method of building words is by taking an alphabet (for example, the Latin alphabet) and concatenating random characters until the sequence reaches a random length between a minimum and maximum. For example, given $min = 3$ and $max = 8$ we can generate the following example text:

```
tqk qzsyjla msmnix jvxx wug sysrh cuepg snyow ptjo bcek
    arjdubw pfwpt nabgzk jmq taphh zewll dmpr uvpmx sfpfk
    uuo bdm vnjbq hahuj wstq kohvma irn fott axdut rlgg
    tawz wsol wigom psqwd tnv vlzgt lbcikk bof msmyg
    zkqgubb veht ukaznqn ixp rppfj eqllnko uyyp aot uowtn
    icv fgypx cenawnk hypq rruh eosgrf wmakeg hhweua gnbfh
     mkpzi ebtwbv cjwrxw ucky kqezcm ucme wmrk khsya
    llzbeqw uxwivpp pbao gkzu pda txdp iwl gkmfqn uxeupe
    atjxy vyul
```

We can immediately see several problems with this generated text. First, many of the words are difficult to pronounce and unrealistic with regards to their consonant clusters - for example, words like `jvxx` and `rppfj` are unlikely to exist in any natural languages. Generating words in this manner will also not produce very much variation in languages, as each letter has an equal probability to be picked.

We also quickly run into the problem of representing language in text. Written languages are based on spoken languages, so generating a written language first without basing it on a spoken one will lead to an unrealistic language. Furthermore, it is hard to say how our generated words are pronounced - we can apply English pronunciations to some of the words (for example `tawz` becomes /tɔːz/ and `axdut` becomes /ˈæks.dʌt/), but this results in a very Anglo-centric phonology, as we are biased to only use phonemes that exist in our own language while pronouncing unknown words.

cite

Because of these problems, it is obvious that merely stringing together random characters with no thought towards pronunciation is insufficient when it comes to creating realistic and varied languages. Therefore it is necessary to

first create a phonology on which to base all of our language's words.

2.1 Phonology

In its simplest form, a phonology is a list of every sound that is included in a language. English phonology, for example, contains around 24 consonants (with more or less depending on dialect) and anywhere between 7 and 14 vowels, again depending heavily on dialect. The phonology for the 'Received Pronunciation' dialect of English can be seen in 2.1 and 2.2.

cite (WALS)

Table 2.1: Consonant inventory in English phonology

		Labial	**Dental, Alveolar**	**Post-alveolar**	**Palatal**	**Velar**	**Glottal**
Nasal		m	n			ŋ	
Plosive, Affricate		p / b	t / d	t͡ʃ / d͡ʒ		k / g	
Fricative	**Sibilant**		s / z	ʃ / ʒ			
	Non-sibilant	f / v	θ / ð			x	h
Approximant			l	r	j	w	

Table 2.2: Vowel inventory in English phonology (Received Pronunciation)

	Front	Central	Back
Close	i / ɪ		u / ʊ
Mid	e	ɜ / ə	ɔ
Open	æ	ʌ	ɑ / ɒ

Other languages have different phonemic inventories, with varying numbers of consonants and vowels. Some, for example, have as few as 3 vowels, or as many as 84 or more consonants, depending on the method of counting[3]. A very large factor in the variety of languages generated is the phonology, as it restricts the type of sounds which often has a profound impact on how it is perceived. Therefore our first step towards a completed language should be a randomly generated, unique phonology.

The International Phonetic Alphabet (IPA) is an alphabet created by the International Phonetic Association for phonetic representation of speech and language[4]. It contains (or aims to contain) distinct symbols for each unique sound possible to create that is part of a language. As such, it provides a perfect way to convey the "end result" of generation process, since we can describe precisely how every word in the language is pronounced. The IPA also includes markers for syllable stress and other important factors which can be included in speech. In order to avoid the user from having to infer the pronunciation of a word from its orthography (how it is written), `yod` can produce an IPA

transcription of its output, which shows exactly how to pronounce it without ambiguity.

However, we can also use the IPA as a basis for *creating* a phonology.

2.2 Syllables

2.3 Stress and Long/Geminate Phonemes

2.4 Words

Chapter 3

Orthography

Chapter 4

Grammar

4.1 Lexicon

4.2 Phrase Structure Grammar

Bibliography

[1] L. L. Zamenhof. *Dr. Esperanto's International Language*. 1887.

[2] Svend Nielsen. *Per-country rates of Esperanto speakers*. 2016. URL: `https://svendvnielsen.wordpress.com/2016/12/10/percountry-rates-of-esperanto-speakers/`.

[3] Georges Dumézil and Tevfik Esenç. *Le verbe oubykh: études descriptives et comparatives*. 1975.

[4] International Phonetic Association. *Handbook of the International Phonetic Association*. 1999.