Properties of Constructed Language Phonological Inventories

Sara Blalock Ng
with Abigail Schwendiman

2 May 2017

Abstract

This paper considers the phonetic distributions of constructed language (conlangs) as evidence for their ability to reflect patterns of natural language. Ancillary to the aim of this direction of study was the creation of CLIPS, the ConLang Inventories of Phonological Segments, a small database of phoneme inventories of the 31 chosen conlangs. This interface allows for easy comparison between the inventories of natural languages and conlangs. We find that while conlangs as a set have encouraging similarities to natural language, they differ in important ways. The frequency with which certain phonemes occur in conlangs is similar to the frequency with which they appear in natural language. Conlang inventories do still contain segments not present (or even feasible) in natural language. Just over 6% of all segments in the set of conlang inventories were not present in any natural inventories. Furthermore, the set of conlangs had a much higher mean frequency index than natural languages. Based on this information, we conclude that conlangs may in fact be influenced by phonetic principles of natural language, but they are not representative of language in general, at least phonetically.

1 Introduction

For many linguists, both professional and amateur, the world of constructed languages (conlangs) is a creative outlet where they may explore their capacity for language invention. Outside the field, conlangs are ubiquitous with science-fiction worlds and far-flung dystopias, the brainchildren of unknown authors.

From an academic perspective, conlangs offer an unique opportunity to explore language creation. Unlike natural languages, conlangs have traceable sources, known authors, and well-defined purposes. The authors of these languages are not long-dead ancestors, but living language enthusiasts with e-mail addresses and personal websites. And yet, conlangs still seem to function like natural languages do. For example, Esperanto was created with a practical communicative purpose, and so it must withstand the rigors of standard language use in the same way as natural languages. And it claims native speakers. [3]

With this in mind, the obvious question is to what extent these conlangs actually mirror natural language. Given a language at random, would one be able to discern whether it was natural or constructed? The aim of this paper is to examine the phonemic inventories of constructed and natural languages, and to review the general phonemic characteristics of conlangs as a set.

2 Review of Literature

2.1 Purpose of Conlangs

Conlangs serve a variety of purposes for their authors and wider audience. The most common type of conlang is a subfield sometimes referred to as ficlangs (fictional languages) or artistic languages. [2] These languages are created for fictional use by invented communities. They often find home among the worlds of science-fiction, and some have faithful speakers in the real world. [1] However, the conlangs with the widest use are auxlangs or auxiliary languages. The purpose of these languages is to facilitate communication between different language communities. Esperanto and Interlingua are perhaps the most famous languages of this type.

Of special interest to linguists are the conlang type known as engelangs, or engineered languages. [2] These languages are specially designed for specific purposes. For example, Loglan was created with the intention of testing the Sapir-Whorf Hypothesis. They often have very specific and intentional properties, and in general do not have the same popularity in usage as the other conlang types.

2.2 The UPSID Database

To be able to compare the set of conlangs to the set of natural languages, we required existing information about the natural languages. Based on the available statistical information, we chose to use UPSID, the UCLA Phonological Segment Inventory Database. This database, created in 1984, contains the segment inventories for 451 natural languages and statistical information based on the languages, language classes, and information for the 919 individual phonological segments contained in all the inventories. [5] The original database was encoded in MS-DOS format, which made it difficult to access in its original form. We chose therefore to use the UPSID interface available through the University of Frankfurt. This program contains all of the information originally available in UPSID, accessible either through an HTML interface or for download as tab-delimited matrix. [7]

3 Interface

Language Name	Source
Atlantean	https://en.wikipedia.org/wiki/Atlantean_language
AUI	https://en.wikipedia.org/wiki/AUI_(constructed_language)
Barsoomian	https://www.datapacrat.com/True/lang/JAHENN∼1/barsoom.htm
Brithenig	http://steen.free.fr/brithenig/combinations.html
Dothraki	https://en.wikipedia.org/wiki/Dothraki_language
D'ni	https://en.wikibooks.org/wiki/D%27ni/Alphabet_and_Phonology
Draconic	ttp://celmin.pwcsite.com/conlang/dnd-draconic/grammar.html
Eskayan	https://en.wikipedia.org/wiki/Eskayan_language
Esperanto	https://en.wikipedia.org/wiki/Esperanto_phonology
Furbish	http://furbytoyshop.com/furby-language
Golic Volcan	http://www.omniglot.com/conscripts/vulcan.htm
Interlingua	https://en.wikipedia.org/wiki/Interlingua
Ithkuil	https://en.wikipedia.org/wiki/Ithkuil#cite_note-foer-2012-1
Klingon	https://en.wikipedia.org/wiki/Klingon_language#Phonology
Laadan	https://en.wikipedia.org/wiki/L%C3%A1adan
Loglan	http://www.loglan.org/Loglan1/chap2.html
Lojban	https://en.wikipedia.org/wiki/Lojban
Na'vi	https://en.wikipedia.org/wiki/Na%27vi_language
Quenya	https://en.wikipedia.org/wiki/Quenya
Sindarin	https://en.wikipedia.org/wiki/Sindarin#Phonology
Old Sindarin	http://folk.uib.no/hnohf/oldsind.htm
Syldavian	https://en.wikipedia.org/wiki/Syldavian
Talossan	http://talossan.com/phonology/
Teonaht	http://dedalvs.conlang.org/misc/lcc1sallyhandout.pdf
Toki Pona	https://en.wikipedia.org/wiki/Toki_Pona
Tsolyani	https://en.wikipedia.org/wiki/Tsoly%C3%A1ni_language
Valyrian	https://en.wikipedia.org/wiki/Valyrian_languages#Phonology
Verdurian	http://www.zompist.com/phonology.htm
Volapuk	https://en.wikipedia.org/wiki/Volap%C3%BCk
Vulcan	https://en.wikipedia.org/wiki/Vulcan_(Star_Trek)#Language
Wenedyk	https://en.wikipedia.org/wiki/Wenedyk

Table 1: Language Sources

While descriptions of individual conlangs are readily available on online forums and through the personal websites of their authors (see Table 1), there does not exist a centralized source for phonological information on conlangs. As this posed a barrier to our research, we endeavored to create for our own use a collection of the available information about conlangs and their phonemic inventories.

3.1 Design

The thirty-one languages in Table 1 were selected based on the availability of their robust phonological descriptions. This set is fairly representative of the most common conlangs, and the different types of conlangs (auxlangs, artlangs, engineered languages). Complete phonemic inventories were collected for each language, as well as the conlang type and native language(s) of its author(s).

Following the precedent of UPSID, language information was encoded, and an interface was created analogous to the Frankfurt program. As the original UPSID data was encoded before the advent of Unicode IPA symbols, segment inventories are stored in the ASCII format. So that the conlang interface could be easily compared to the data in UPSID, our encoding was also in ASCII, following the guidelines in Moran 2012. [6] This choice limited in some ways the kind of segmental features encoded, for example ASCII does not offer any convention for noting whether a segment may be syllabic (which is achieved using a diacritic marker in IPA).

We called our program CLIPS, the ConLang Inventories of Phonological Segments, and source code can be found at github.com/Sara
BlalockNg/fake-upsid. Table 2 shows the basic capabilities of this new interface. The first six capabilities in this table are identical to the information available through the Frankfurt interface. [7]

In addition, to these functions, CLIPS also allows for the direct comparison between the inventory of the conlangs with the inventories of the native languages of their authors. For facilitate this functionality, the inventories for English, Yiddish, German, Dutch, and Boholano-Visayan (Cebuano) were encoded from external sources. These languages were not originally in UPSID (although Russian and French, which are among the authors' native languages, were).

4 Analysis

The following sections present some of the more interesting properties found in CLIPS. The inventories of the 31 conlangs contained 214 unique segments, 6.62% of which did not appear in any of the inventories in UPSID. We posit that the cause of this phenomenon is two-fold: First, the phonologies of conlangs are not constrained in the same ways as natural language. Many artistic conlangs are designed to be spoken by alien races with physiologies very different to speakers of natural languages. Thus, some segments which are difficult or even impossible to be vocalized by humans may appear more readily in a constructed language. Second, the set of surveyed natural languages, while representative of the set of all natural languages, only actually account for a small proportion of all existing languages.

Question in Interface	Associated Page Display Contains:
Do you want to... get information about a language?	- full inventory for language - number of segments - author's native language(s) - source
sort languages by the number of sounds?	List of languages and inventory size, ordered from least number of phonemes to greatest
sort languages by their frequency index?	Frequency index, number of segments, and language, sorted by frequency index from least to greatest
get information about a language class?	List of the language contained in a selected class (artistic, auxiliary, or engineered)
find certain sounds and languages that have them?	- languages containing segments matching selected features - the specific sounds in the inventories that match the set of features - the percent of sounds in each matching inventory that meet the criterion
compare two languages?	The common segments between two selected languages, or among a language class
compare a conlang to the native language of its author?	- full inventory for language - number of segments - author's native language(s) - native language inventory - percent of segments shared by conlang and parent language - list of segments unique to the conlang (segments not found in the parent inventory) - source

Table 2: Interface Capabilities

It may well be that some of the 6.62% are actually present in some natural language, just not in UPSID.

One concern in our analysis was the classification of conlangs. In UPSID, languages are separated into classes based on geography and etymological similarities. For most conlangs, this dichotomy is impractical. We therefore divided the conlangs into classes based on their intended purposes, as described in Section 2.1. There were 19 languages in the Artistic class, four in the auxiliary class, and seven in the engineered class. A breakdown of these classes is provided in Table 3.

4.1 Inventory Size

One feature by which conlangs and natural languages differ is in the size of their segment inventories. The average size of the conlang inventories was 37.74, but the average for natural languages was 30.96. This means that conlang inventories contained on average seven more

Artistic		Auxiliary	Engineered
Atlantean	Quenya	AUI	Brithenig
Barsoomian	Sindarin	Eskayan	Ithkuil
Dothraki	Syldavian	Esperanto	Laadan
D'ni	Talossan	Interlingua	Loglan
Draconic	Teonaht		Lojban
Furbish	Tsolyani		Toki Pona
Golic Volcan	Valyrian		Wenedyk
Klingon	Verdurian		
Old Sindarin	Vulcan		
Na'vi			

Table 3: Language Classes

segments than natural inventories. This difference is statistically significant (p=0.0017).

One possible reason for the large inventory size is that the native languages of the conlang authors are all higher than the average. Exact inventory sizes can be found in Table 4. It may be that the large inventory sizes of the parent languages sets a precedent for the created languages that descend from them.

Parent Language	Inventory Size
Boholano-Visayan	39
Dutch	55
English	56
French	37
German	44
Russian	37
Yiddish	44

Table 4: Parent Inventory Sizes

4.2 Frequency Indices

The frequency index of a segment is the percentage of inventories in which it appears in the set. For example, the segment [a:] appears in 22.8% of constructed languages. Thus, its frequency index in the conlang set is 0.228. In contrast, the same segment has a frequency index of 0.0754 in UPSID. [7] In general, segments with high frequency indices in UPSID had relatively high frequency indices in CLIPS. One notable difference was the set of long vowels (like the example). Long vowels had much higher frequency among conlangs than among the natural languages in UPSID.

The frequency index of a language is the arithmetic mean of the frequency indices of the segments in its inventory. There was a statistically significant difference between the average frequency index of conlangs and the frequency index of natural languages (p = 0.0001). The average index of conlangs was 0.584, while the average for natural languages was 0.391. This means that conlang inventories as a set reuse popular segments more often than natural languages do.

One possible reason for the high relative frequency of segments in CLIPS is the lack of diversity in authorship. Some of the languages were created by the same author (Sindarin and Quenya, for instance, were both the creation of J.R.R. Tolkien). Even when they had different authors, many of the languages were inspired by one another in some way. Natural languages often have millions of speakers actively using and changing their phonological inventory; the creative pool from which the conlangs were devised cannot compete with this diversity of thought.

4.3 Comparison to Parent Languages

Many of our suppositions about the cause of the observed distributions in CLIPS rely on the relationship between a conlang and the native language(s) of its author. In fact, it appears that conlangs take much of their inventories from their parent languages. On average, 62.42% of the segments in the conlang inventories were also present in their parent inventories.

The lower bound of shared percentages was 42.03%. Klingon, a language which was designed to sound foreign or alien, still shared 42.42% with its parent language (English).

5 Conclusion

While it seems that some patterns of conlangs' segment inventories do follow the patterns observed in the set of natural languages, they differ in important ways. The average inventory size is much larger for constructed languages. In addition, the set of conlangs tends to use popular segments, like long vowels, much more often. In contrast, natural languages are more likely to use 'rare' segments in their inventories.

5.1 Moving Forward

It is our hope that the creation of CLIPS will facilitate any future research on the phonological properties of constructed languages. As we believe this database to be the first of its kind, we hope that CLIPS can serve as a centralized source of information for conlangs, and that the interface's capabilities will expand as a result of future collaboration.

In addition, it would be of interest to further investigate the phonemic distributions across language classes, i.e., to examine whether special phonemic properties exist in conlangs because of their express purpose.

In general, differences between the sets of phonemic inventories speak to inherent difference between conlangs and natural languages. There are important factors separating conlangs from 'real' language. The difference in makeup of their phonological inventories show how they may be influenced by their authors' parent languages and the intent of their creation.

References

[1] Adams, M. *From Elvish to Klingon: exploring invented languages*. Oxford University Press, 2011.

[2] Destruel, M. *Reality in Fantasy: Linguistic analysis of fictional languages*. PhD thesis, Boston College, 2016.

[3] Fiedler, S. The Esperanto denaskulo: The status of the native speaker of Esperanto within and beyond the planned language community. *Language Problems and Language Planning 36*, 1 (2012), 69–84.

[4] Kelly, P., et al. A comparative analysis of Eskayan and Boholano-Visayan (Cebuano) phonotactics: Implications for the origins of Eskayan lexemes.

[5] Maddieson, I., and Precoda, K. The UCLA Phonological Segment Inventory Database (UPSID).

[6] Moran, S. Using Linked Data to create a typological knowledge base. In *Linked Data in Linguistics*. Springer, 2012, pp. 129–138.

[7] Reetz, H. Web interface to UPSID, 1999.