Properties of Constructed Language Phonological Inventories*

Sara Ng, Abigail Schwendiman Sleight

Department of Linguistics, University of Washington, Box 354340, Seattle, WA, 98195-4340
Lendio, 2901 West Coast Hwy Ste. 200, Newport Beach, CA, 92663
sbng@uw.edu, abigailsleight@gmail.com

Abstract: This paper considers the phonetic distributions of constructed languages (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, a small database of phoneme inventories sampled from documented 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. We find that frequency with which certain phonemes occur in conlangs is similar to the frequency with which they appear in natural language. However, we also find that Conlang inventories do still contain segments not present (or even feasible) in natural language. Furthermore, we find that conlangs have 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.

Keywords: phonology, constructed language, typology

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 research 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 (Fiedler 2012).

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 Background

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

*Our thanks to Aaron Kaplan for his mentorship and feedback.

(Destruel 2016). 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 (Adams 2011). 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 (Destruel 2016). 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 (Maddieson & Precoda 2011). 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 (Reetz 1999).

3 Interface

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). 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 https://github.com/SaraBlalockNg/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 (Reetz 1999).

Language Name	Source
Atlantean	Ager n.d.(a)
aUI	Cosmic Communication Foundation n.d.
Barsoomian	*Baroomsian* n.d.
Brithenig	Smith 2007
Dothraki	*DothraWiki: Phonology* n.d.
D'ni	Ager n.d.(b)
Draconic	van Steenbergen 2015
Eskayan	Kelly 2006
Esperanto	Wennergren 2016
Furbish	*Furby Toy Shop* 2010
Golic Vulcan	Ager n.d.(c)
Interlingua	Gode & Blair 1951
Ithkuil	Quijada 2011
Klingon	Okrand 1992
Láadan	Elgin 1985
Loglan	Brown 1989
Lojban	Cowan 1997
Na'vi	Zimmer 2009
Quenya	Fauskanger n.d.(b)
Sindarin	Fauskanger n.d.(c)
Old Sindarin	Fauskanger n.d.(a)
Syldavian	Rosenfelder n.d.
Talossan	Association of Talossan Language Organisations n.d.
Teonaht	Caves n.d.
Toki Pona	Lang 2014
Tsolyani	Barker 1978
Valyrian	Peterson 2013
Verdurian	Rosenfelder 2004
Volapük	Sprague 1887
Vulcan	*Vulcan (Star Trek) Language* n.d.
Wenedyk	van Steenbergen 2006

Table 1: The source from which the CLIPS inventories were derived. Note that for some inventories (e.g. Dothraki), the most complete source was multi-user-generated.

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. The inventories for English, Yiddish, Dutch, and Boholano-Visayan (Cebuano) were encoded from various sources in order to compare their inventories to the conlangs created by their native speakers. Additionally the inventories of French, German, and Russian were ported from the UPSID web interface. While it would be preferable to match the encoded language varieties to the native varieties of the authors, this is infeasible at scale due to both a lack of reliable biographical information and ambiguity in resolving parent varieties of languages with multiple authors.

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)
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) - percent of segments shared by conlang and parent language - list of segments unique to the conlang (segments not found in the parent inventory)

Table 2: Interface Capabilities

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.

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

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

Table 3: Languages sorted into Artistic, Auxiliary, and Engineered Classes.

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 in UPSID (Reetz 1999). This means that conlang inventories contained on average seven more segments than natural inventories.

Parent Language	Inventory Size
Boholano-Visayan[1]	39
Dutch[2]	42
English[3]	40
French[4]	37
German[4]	41
Russian[4]	38
Yiddish[5]	47

Table 4: Parent inventory sizes

One possible reason for the large inventory size is that the native languages of the conlang authors are all higher than the average. Inventory sizes for the native inventories 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.

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 [aː] appears in 22.8% of sampled constructed languages. Thus, its frequency

[1] Kelly et al. 2015
[2] Gussenhoven 1992
[3] Ladefoged 1999
[4] Maddieson & Precoda 2011
[5] Kleine 2003

index in the conlang set is 0.228. In contrast, the same segment has a frequency index of 0.0754 in UPSID (Reetz 1999). 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 (Reetz 1999). 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 in UPSID was 0.391 (Reetz 1999). 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, especially native languages of the conlang authors in CLIPS, have large populations of speakers actively using and diachronically 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(s). In fact, it appears that conlangs take much of their inventories from their parent languages. On average, 63.82% of the segments in the conlang inventories were also present in their parent inventories. The lower bound of shared percentages was 34.783% (Wenedyk). Klingon, a language which was designed to sound foreign or alien, still shared 45.455% 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.

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

Adams, Michael. 2011. *From Elvish to Klingon: exploring invented languages*. Oxford University Press.

Ager, Simon. N.d.(a). *Atlantean (DIG adlantisag)*. https://omniglot.com/conscripts/atlantean.htm. Accessed: 2017-05-17.

Ager, Simon. N.d.(b). *D'ni alphabet)*. https://omniglot.com/conscripts/dni.htm. Accessed: 2017-05-17.

Ager, Simon. N.d.(c). *Golic Volcan*. https://www.omniglot.com/conscripts/vulcan.htm. Accessed: 2017-05-17.

Association of Talossan Language Organisations. N.d. *El Glhe Talossan Phonology*. http://talossan.com/phonology. Accessed: 2017-05-17.

Barker, M. A. R. 1978. *The Tsolyani language*. Vol. 1. Imperium Publishing Company.

Baroomsian. N.d. https://www.datapacrat.com/True/lang/JAHENN%E2%88%BC1/barsoom.htm. Accessed: 2017-05-17.

Brown, James Cooke. 1989. *Loglan 1: a logical language*. Loglan Institute. Chap. 2.

Caves, Sallh. N.d. *Teonaht*. https://dedalvs.com/misc/lcc1sallyhandout.pdf. Accessed: 2017-05-17.

Cosmic Communication Foundation. N.d. *Elements of Meaning*. https://auilanguage.org/elements-of-meaning. Accessed: 2017-05-17.

Cowan, John Woldemar. 1997. *The complete Lojban language*. Vol. 15. Logical Language Group.

Destruel, Mathieu. 2016. *Reality in fantasy: linguistic analysis of fictional languages*. Boston College dissertation.

DothraWiki: Phonology. N.d. https://wiki.dothraki.org/phonology. Accessed: 2017-05-17.

Elgin, Suzette Haden. 1985. *A first dictionary and grammar of Laadan*. Society for the Furtherance, Study of Fantasy & Science Fiction.

Fauskanger, Helge. N.d.(a). *Old Sindarin - between Primitive Elvish and Grey-elven*. https://folk.uib.no/hnohf/oldsind.htm. Accessed: 2017-05-17.

Fauskanger, Helge. N.d.(b). *Quenya - the Ancient Tongue*. https://folk.uib.no/hnohf/quenya.htm#Heading6. Accessed: 2017-05-17.

Fauskanger, Helge. N.d.(c). *Sindarin - the Noble Tongue*. https://folk.uib.no/hnohf/sindarin.htm#Heading6. Accessed: 2017-05-17.

Fiedler, Sabine. 2012. 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). 69–84.

Furby Toy Shop. 2010. http://furbytoyshop.com/furby-language. Accessed: 2017-05-17.

Gode, Alexander & Hugh E Blair. 1951. *Interlingua: a grammar of the international language*. Storm Publishers.

Gussenhoven, Carlos. 1992. Dutch. *Journal of the International Phonetic Association* 22(1-2). 45–47.

Kelly, Piers. 2006. The classification of the Eskayan language of Bohol. *Tagbilaran, Bohol: National Commission on Indigenous Peoples*.

Kelly, Piers et al. 2015. A comparative analysis of Eskayan and Boholano-Visayan (Cebuano) phonotactics: implications for the origins of Eskayan lexemes.

Kleine, Ane. 2003. Standard Yiddish. *Journal of the International Phonetic Association* 33(2). 261–265.

Ladefoged, Peter. 1999. American English. *Handbook of the international phonetic association* 4144.

Lang, Sonja. 2014. *Toki Pona: the language of good*. Sonja Lang.

Maddieson, Ian & Kristin Precoda. 2011. *UCLA Phonological Segment Inventory Database (UPSID)*. http://www.linguistics.ucla.edu/faciliti/sales/software.htm. Accessed: 2017-05-17.

Moran, Steven. 2012. Using linked data to create a typological knowledge base. *Linked data in linguistics: Representing and connecting language data and language metadata*. 129–138.

Okrand, Marc. 1992. *The Klingon dictionary: the official guide to Klingon words and phrases*. Simon & Schuster.

Peterson, David J. 2013. *Tīkuni Zōbrī, Udra Zōbriar*. https://dothraki.com/2013/04/tikuni-zobri-udra-zobriar. Accessed: 2017-05-17.

Quijada, John. 2011. *A Grammar of New Ithkuil*. http://ithkuil.net/newithkuil_01_phonology.htm.

Reetz, Henning. 1999. *Web interface to UPSID*. http://web.phonetik.uni-frankfurt.de/upsid_info.html. Accessed: 2017-05-17.

Rosenfelder, Mark. 2004. *Verdurian reference grammar*. https://www.zompist.com/vergram.html. Accessed: 2017-05-17.

Rosenfelder, Mark. N.d. *Hergé's Syldavian: A grammar*. http://www.zompist.com/syldavian.html. Accessed: 2017-05-17.

Smith, Andrew. 2007. *Brithenig*. http://steen.free.fr/brithenig/combinations.html. Accessed: 2017-05-17.

Sprague, Charles Ezra. 1887. *Hand-book of Volapük*. Office Company.

van Steenbergen, Jan. 2006. *Wenedyk*. http://steen.free.fr/wenedyk/alphabet.html. Accessed: 2017-05-17.

van Steenbergen, Jan. 2015. *The Draconic Language of D&D*. http://celmin.pwcsite.com/conlang/dnd-draconic/grammar.html. Accessed: 2017-05-17.

Vulcan (Star Trek) Language. N.d. https://en.wikipedia.org/wiki/Vulcan_(Star_Trek)#Language. Accessed: 2017-05-17.

Wennergren, Bertilo. 2016. *Plena manlibro de Esperanto gramatiko*. Accessed: 2017-05-17. Esperanto-Logo por Norda Ameriko. Chap. 1.2, 25–36.

Zimmer, Ben. 2009. *Some highlights of Na'vi*. https://languagelog.ldc.upenn.edu/nll/?p=1977. Accessed: 2017-05-17.