Master's Thesis

in Computational Linguistics

Deconstructing Constructed Languages

Author:
Connor Kirberger

Supervisors:
Çağrı Çöltekin
Christian Bentz

Seminar für Sprachwissenschaft
Eberhard-Karls-Universität Tübingen

December 2023

Hiermit versichere ich, dass ich die Arbeit selbständig verfasst, keine anderen als die angegebenen Hilfsmittel und Quellen benutzt, alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe und dass die Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist und dass die Arbeit weder vollständig noch in wesentlichen Teilen bereits veröffentlicht wurde sowie dass das in Dateiform eingereichte Exemplar mit den eingereichten gebundenen Exemplaren übereinstimmt.

I hereby declare that this paper is the result of my own independent scholarly work. I have acknowledged all the other authors' ideas and referenced direct quotations from their work (in the form of books, articles, essays, dissertations, and on the internet). No material other than that listed has been used.

Tübingen, September 1, 2024

_____

Firstname Surname

Contents

Abstract

Write the abstract here.

List of Figures

List of Tables

List of Abbreviations

NLP	Natural Language Processing
PCA	Principle Component Analysis
TF-IDF	Term Frequency - Inverse Document Frequency
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
SVM	Support Vector Machine
XML	eXtensible Markup Language
TTR	Type-Token Ratio
MATTR	Moving-Average Type-Token Ratio
IAL	International Auxiliary Language
SVO	Subject-Verb-Object

1 Introduction & Motivation

Constructed languages—also called artificial languages, invented languages, planned languages, engineering languages, glossopoeia, or more simply as "conlangs" (Ball 2015)—are languages that are consciously and purposefully created for some intended use, usually being defined in antithesis to the spontaneous and organic method in which natural languages arise and develop (Sanders 2016). These variations of the term are often, but not always, used interchangeably, as linguists do not all agree upon a core term due to personal preferences (Adelman 2014), and there are sometimes differences in nuance depending on the context in which they appear. This thesis will mainly refer to them as constructed languages for simplicity.

The intended uses for which they are created can range broadly. Some are made specifically for fictional media, often seen in the genres of fantasy or science-fiction, with some more well-known examples being J. R. R. Tolkien's Elvish languages (e.g., Quenya, Sindarin, Nandorin) found in the world of Middle-earth in his writings, Marc Okrand's Klingon language from the Star Trek universe, and David J. Peter's Dothraki language used in George R. R. Martin's A Song of Ice and Fire novels along with their television adaptation, Game of Thrones (Jeffrey Punske (editor) 2020). Others are created to function as international auxiliary languages (IALs)—languages planned for the use of international and cross-cultural communication (Gobbo 2016). The most well-known example (based on estimated number of speakers) of these is Esperanto, created in the 19$^{\text{th}}$ Century by L. L. Zamenhof. Typically, constructed languages are distinguished and categorized based on these communicative functions, although other taxonomies also exist, as we will discuss further in section 2.

Despite being defined in contrast to one another, however, constructed and natural languages are not necessarily opposite to one another characteristically. Aside from their origins, the boundaries between the two are not always clear when analyzed in greater detail (Goodall 2022). For example, Schubert (1989) argues that some languages which are considered "natural" have some degree of artificiality, such as standardized written German and English differing from their spoken forms, and that the reverse is also true of some languages which are considered "artificial" because they draw from aspects of natural languages. As such, he believes human languages exist on a continuum of the two labels, rather than in the binary distinction—a view echoed by other linguists as well (Novikov 2022).

In many ways, it can be argued that investigation into the disparity between the two is at its core a mere part of the larger debate surrounding what exactly constitutes a language. The search for and defining of language univer-

sals is not only fundamental to the field of linguistic research, but also a topic of widespread debate. At present, many theoretical and foundational contributions to this discussion exist, ranging from Greenberg's proposed universals (Greenberg 1970) to ideas about universal grammars, such as Chomsky's.

Furthermore, while research on constructed languages is far from being novel (in fact it is a dedicated field of study called interlinguistics, which will be discussed in more detail further on), it is less common in computational linguistics. Thus, my motivation here was to try something relatively new, by approaching such an investigation using machine learning methods.

1.1 Scope of Study & Research Question

The present work analyzes various linguistic features of both types of languages and seeks to make a comparison on the differences, if any exist, between them. More specifically, this study presents a binary classification task using decision tree and outlier detection models to determine if they can be distinguished from one another, based on the specific features examined and parallel Wikipedia data.

Because of the wide-ranging nature of conducting such a broad analysis, there will of course be many features left unconsidered or excluded, intentionally or otherwise. With this in mind and following the precedent set by other related research on this topic, the focus in this particular study is mainly on features such as the entropy, morphological complexity, and lexical diversity of each language, based on the selected corpora.

The following is a breakdown of the structure of this thesis from here onward: the next section provides relevant background information, including an overview on constructed languages and a comprehensive review of related literature that examines the prior theoretical groundwork laid for exploring linguistic similarities and differences between constructed and natural languages; section 3 covers in detail the methodology taken in this research, from an explanation of the data used to the various experiments performed; section 4 presents the results of the study and discussion of these follows in section 5; lastly, section 6 consists of a conclusion as well as elaboration for possible future work.

2 Background

The vast landscape of linguistic research comprises a myriad of literature delving into the intricacies of languages, both natural and constructed. As this paper is concerned with constructed languages in particular and possible distinctive properties they may have, a brief discussion of their history and development is of relevance here. Following this is an overview of some related literature, after which I will discuss the various computational methods implemented in this study and provide some background information on how they work.

2.1 History of Constructed Languages

Okrent (2009) states, "The history of invented languages is, for the most part, a history of failure." She may be justified in saying this, depending on one's definition of failure in this context. From past to present, the total number of constructed languages may be as high as a thousand (Libert 2016; Schubert 1989), with hundreds proposed for the purpose of being IALs in Europe alone (Schubert et al. 2001). Yet of these, Esperanto is widely considered the most successful, with very few others even coming close.

While the construction of languages is possibly as old as human history, they typically were not written down and were limited to in-group communication (Gobbo 2016). The first documented endeavors came out of religious contexts and were likely used as secret languages, intentionally obscured and incomprehensible to lay people. In the 12^{th} century, abbess Hildegard of Bingen described and recorded a lexicon for Lingua Ignota, a Latin name meaning "unknown language". While extensive documentation of it (i.e., a grammar) was never found, it possessed a semiotic system based on Latin, German, and Greek. Later in the 14^{th} century, a group of Sufi mystics created Balaibalan, a language written in the Ottoman Turkish alphabet and which incorporated features of Persian, Turkish, and Arabic languages (Novikov 2022).

Interest in creating such languages picked up in the 17^{th} century with the rise of so-called philosophical languages. In contrast to the last two, these languages were made to be more precise, less ambiguous, and better allow for philosophical reasoning (compared to natural language), such as by organizing world knowledge into hierarchies (Goodall 2022). Notable figures involved in making these include Francis Lodwick, Gottfried Leibniz, and John Wilkins, the latter of whose being arguably the most well-known and influential. Wilkins created a system of semantic categorization, cataloging all concepts in the universe (Okrent 2009). An example can be seen in figure 2.1.

In the 19^{th} and 20^{th} centuries the focus for language construction, espe-

(1) special > creature > distributively > substances > animate > species > sensitive > sanguineous > beasts > viviparous > clawed > rapacious > oblong-headed > European > terrestrial > big > docile

Figure 2.1: Wilson's expression of "dog" (Goodall 2022)

cially in Europe, shifted to that of making international auxiliary languages (IALs) intended to better enable communication across language barriers, i.e., people who do not share a similar language (Goodall 2022). Notably, this means they were generally designed to resemble natural language, with choice exceptions being the simplification of certain linguistic features. The surge in need for IALs correlated with the increase in prevalence and accessibility regarding international travel and communication at the time. Such languages were also described as "neutral" (Large 1985), in the sense that individual advantages amongst speakers and learners would, theoretically, not exist due to IALs being second languages to everyone (Gobbo 2016). That being said, many of the most prominent examples (e.g., Volapük, Interlingua, Esperanto, Ido) are derived from European languages (Novikov 2022; Goodall 2022).

As the constructed languages examined and used in the dataset of the present work are all IALs, it would be beneficial to introduce them in more detail here. Volapük was made in 1879 by Catholic German priest Johann Martin Schleyer, who believed it was given to him by God. Argued to be the first successful constructed language due to amassing so many supporters (Gobbo 2016), it soon died out in favor of Esperanto, which Ludwik Lejzer Zamenhof published in 1887.

While they all share the defining feature of having been purposefully created, their other features (e.g., phonetic, morphological, syntactical, lexical, orthographic) can vary immensely depending on factors such as, for example, their intended purpose for use or the other languages they draw from. An example of this was observed by Gobbo (2016) in secret languages, specifically their tendency to have more complicated features, such as morphological irregularities, "in order to preserve their secrecy." Contrast to this are IALs, which have the opposite tendency for the sake of ease of communication and language learning, reflected in commonly assigned features such as SVO word orders, head-initial relative clauses, fronted *wh*-phrases, and morphological regularity (Goodall 2022). Section 2.2 discusses this more.

Thus far we have classified constructed languages according to their communicative functions as intended by their creators. This is not the only taxonomy used when analyzing and discussing these, however. Another commonly accepted distinction is that of *a priori* and *a posteriori* languages (Schreyer and Adger 2021; Gobbo 2008; Schubert 1989; Schubert et al. 2001; Novikov 2022).

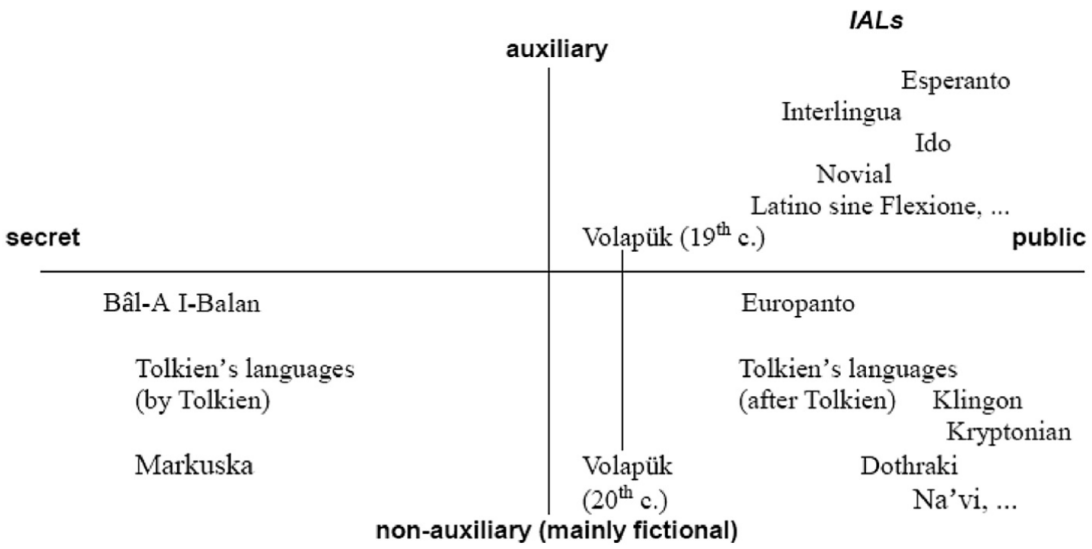

Figure 2.2: A taxonomy of constructed languages (Gobbo 2016)

Figure 2.2 shows one such taxonomy.

2.2 Prior Studies

In contrast to the abundance in literature and cross-linguistic analyses done on natural languages, similar research regarding constructed languages is relatively scarce. In particular, while there is some research that analyzes, to an extent, specific instances of linguistic differences between certain natural and constructed languages, larger-scale cross-linguistic studies which utilize computational methods are practically nonexistent. Consequently, the present study is a somewhat novel approach.

Previous studies

2.3 Computational Methods

3 Methodology

In this section, I explain in greater detail the dataset, APIs, and libraries used, as well as the approaches and steps taken to arrive at my results. Because of the broad nature of this study, several different experiments are done to test various linguistic features. Namely, these features are morphological complexity, character and word frequency distributions, and character entropy. Once these were calculated for each language, the task became that of classification through the use of 2 classifier models: outlier detection and a decision tree.

3.1 Data

Because this study is so broad in scope, the corpora used must be adequately sized as well. In total, twenty-four languages are analyzed in this study. Six of these are constructed languages: Esperanto, Interlingua, Lingua Franca Nova, Volapük, Kotava, and Ido. The remaining eighteen are natural languages: German, English, Spanish, Polish, Vietnamese, Indonesian, Turkish, Tagalog, Hungarian, French, Finnish, Italian, Dutch, Occitan, Danish, Swedish, Afrikaans, and Icelandic.

To briefly introduce the constructed languages used here, Esperanto is the most widely spoken constructed language

The final constructed language used in this study is Lingua Franca Nova, created by Dr. C. George Boeree. Compared to the others discussed so far, it is much more recent in its creation, having first appeared in 1998 online. As a result of being both newer and more niche, there is considerably less existing research related to it. Its inclusion in my dataset is primarily due to it being used on Wikipedia as an available source language, and thus also having a Wikimedia dump file.

As this study is cross-linguistic in nature, it would naturally be ideal to use parallel corpora, as this would enable more conclusive comparative analysis. However, creating or finding parallel corpora that also includes the aforementioned constructed languages is rather challenging due to the limited availability of resources for many of them. Therefore, I instead opted for a more practical approach of using

3.1.1 Wikimedia

The corpora were compiled using Wikimedia database dumps—large files containing Wikipedia articles for a given language which are formatted in XML. The corpus sizes were constricted to be comparable in length based on number of words, while still maintaining complete sentences. The smallest dump was

Lingua Franca Nova, so its size was the minimum value used to shorten the others.

3.2 Data Preprocessing

Preprocessing text data is essential for NLP tasks. As this study covers a broad range of different languages and includes constructed languages, which often have less resources available, meticulous effort was made to obtain as close to a parallel set of corpora as possible and to clean the text thoroughly.

After extracting the text from the dump files through the use of a Python script[1], regular expressions were used to remove page titles, links, headers, fragments, and other extraneous symbols. The text was then made all lowercase and split by the periods—while also attempting to account for abbreviations—to make separate sentences, and all characters that were not part of the language's writing system were removed in an attempt to have only that language's words, without any foreign words that occasionally appear in the scraped Wikipedia texts. Remaining punctuation and numbers were also removed. The end result was a single corpus file corresponding to each language, with each line of the files being a single sentence.

3.3 Libraries and APIs

3.3.1 Keras

3.3.2 PyTorch

3.3.3 NumPy, Pandas, and Matplotlib

3.3.4 scikit-learn

3.4 Feature Extraction

Once the data had been preprocessed, some initial values were calculated to be used as a starting point for our investigation in comparing natural and constructed languages. These were Zipf's law of abbreviation, type-token ratio (TTR), moving-average type-token ratio (MATTR), and character and word distribution entropies.

[1] https://github.com/apertium/WikiExtractor/tree/master

3.4.1 Type-Token Ratio

TTR is a way of measuring lexical diversity. It is calculated using the following formula:

$$TTR = \frac{\sum_{i=1}^{n} \delta(w_i)}{n}$$

3.4.2 Moving-Average Type-Token Ratio

A glaring issue with TTR, however, is that it can vary widely based on a text's length. The longer a particular text, the higher the likelihood of repetition occurring. There have been several solutions proposed to address this issue, one being MATTR. MATTR is given in the formula

$$MATTR_i = \frac{TTR_1 + TTR_2 + ... + TTR_i}{i}$$

3.4.3 Morphological Complexity

The morphological systems of each language was also investigated and analyzed, specifically their complexities. Morfessor,

3.4.4 Zipfian Distribution

3.4.5 Entropy

In information science, entropy means... In linguistics, entropy refers to... Surprisal...

The entropy was calculated for the character and word distributions in each of the corpora, given by the following formula:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

3.4.6 Perplexity

Perplexity is related to entropy... Cross-entropy as the loss function...

3.4.7 PCA

Principal Component Analysis was performed for dimensionality reduction.

A script was used to increase readability of the text in the graph[2].

[2]https://github.com/Phlya/adjustText

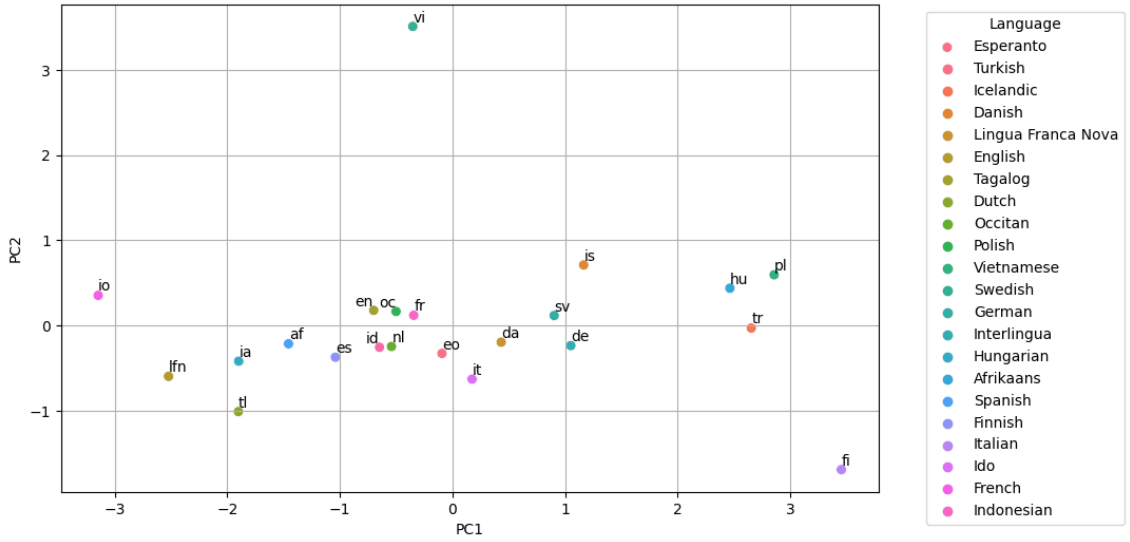

Figure 3.1: Principal Component Analysis on TTR, MATTR, Word and Char Distribution Entropy

3.5 Classification Models

3.5.1 Decision Tree

Decision Tree Classifier...

3.5.2 Outlier Detection

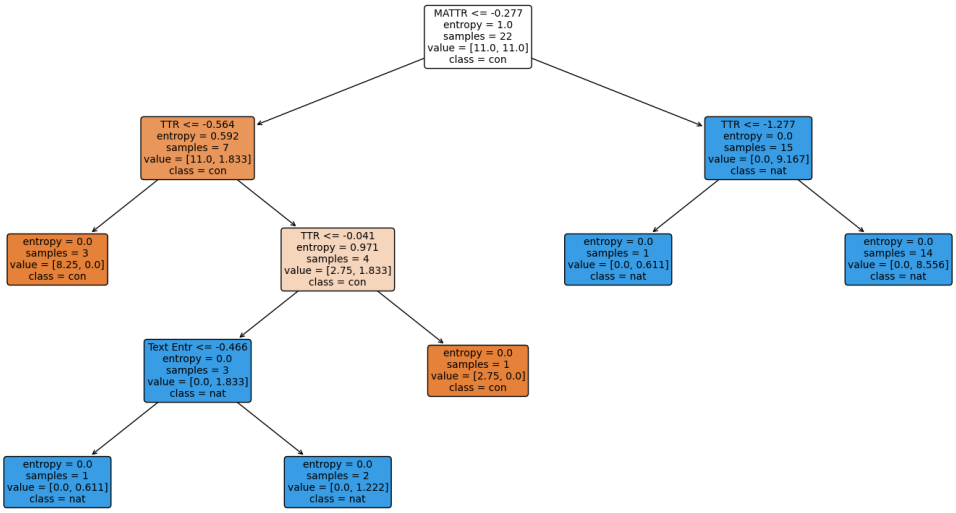

Figure 3.2: Decision Tree Classifier

4 Results

5 Discussion

6 Conclusion

6.1 Future Work

The research presented in this thesis is far from encompassing all there is to the topic of defining language, and distinguishing between constructed and natural language. At present, this is an area of research with ample room for potential development.

Limiting factors: number of languages and which languages/language families, lack of real parallel corpora, problems associated with low-resource languages, relatively narrow scope of experimentation,

7 Acknowledgments

I would like to thank ....

References

Adelman, Michael J. (2014). "Constructed Languages and Copyright: A Brief History and Proposal for Divorce". In: *Harvard Journal of Law & Technology* 27, p. 543. URL: https://api.semanticscholar.org/CorpusID:58553165.

Ball, Douglas (2015). "Constructed languages". In: *The Routledge Handbook of Language and Creativity*. Ed. by Rodney H Jones. Routledge. Chap. 8. DOI: 10.4324/9781315694566.ch8.

Gobbo, Federico (Jan. 2008). "Planned languages and language planning: The contribution of interlinguistics to cross-cultural communication". In: *Multilingualism and Applied Comparative Linguistics* 2.

– (Oct. 2016). "Are planned languages less complex than natural languages?" In: *Language Sciences* 60. DOI: 10.1016/j.langsci.2016.10.003.

Goodall, Grant (Sept. 2022). "Constructed Languages". In: *Annual Review of Linguistics* 9. DOI: 10.1146/annurev-linguistics-030421-064707.

Greenberg, Joseph H. (1970). "Language Universals". In: *Theoretical Foundations*. Berlin, Boston: De Gruyter Mouton, pp. 61–112. ISBN: 9783110814644. DOI: doi:10.1515/9783110814644-003. URL: https://doi.org/10.1515/9783110814644-003.

Jeffrey Punske (editor) Nathan Sanders (editor), Amy V. Fountain (editor) (2020). *Language Invention in Linguistics Pedagogy*. Oxford University Press. ISBN: 0198829876,9780198829874. URL: http://gen.lib.rus.ec/book/index.php?md5=0B5CF2BFC00DCB569EBA10BD96AD68D4.

Large, J.A. (1985). *The Artificial Language Movement*. Language library. B. Blackwell. ISBN: 9780631144977. URL: https://books.google.de/books?id=xaeCQgAACAAJ.

Libert, Alan (Jan. 2016). "On Pragmemes in Artificial Languages". In: pp. 375–389. ISBN: 978-3-319-43490-2. DOI: 10.1007/978-3-319-43491-9_20.

Novikov, Philipp (July 2022). "Constructed Languages as Semantic and Semiotic Systems". In: *RUDN Journal of Language Studies, Semiotics and Semantics* 13, pp. 455–467. DOI: 10.22363/2313-2299-2022-13-2-455-467.

Okrent, Arika (2009). *In the Land of Invented Languages: Esperanto Rock Stars, Klingon Poets, Loglan Lovers, and the Mad Dreamers who Tried*

to Build a Perfect Language. Spiegel & Grau. ISBN: 9780385527880. URL: https://books.google.de/books?id=E3UE9IoW27AC.

Sanders, Nathan (Sept. 2016). "Constructed languages in the classroom". In: *Language* 92, e192–e204. DOI: 10.1353/lan.2016.0055.

Schreyer, Christine and David Adger (Mar. 2021). "Comparing prehistoric constructed languages: World-building and its role in understanding prehistoric languages". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 376. DOI: 10.1098/rstb.2020.0201.

Schubert, Klaus (Jan. 1989). "Interlinguistics – Its Aims, Its Achievements, and Its Place in Language Science". In: pp. 7–44. ISBN: 9783110886115. DOI: 10.1515/9783110886115.7.

Schubert, Klaus et al. (Jan. 2001). *Planned Languages: From Concept to Reality*.