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, August 21, 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

1 Introduction

Constructed languages—also called artificial languages, invented languages, planned languages, glossopoeia, or more simply as "conlangs", a shortened neologism derived from combining the beginning letters of "constructed" and "language" (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). However, that is not to say constructed languages are characteristically opposite to natural languages. Consciously or even subconsciously, their creators take inspiration and influences from already existing languages, be it natural or other constructed languages (Oostendorp 2019).

The intended uses for constructed languages can range broadly. Some are created 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 (Quenya, Sindarin, Nandorin, etc.) found in the world of Middle-earth in his novels, 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, the most widespread and famous example (based on estimated number of speakers) being Esperanto, created in the 19th Century by L. L. Zamenhof. Typically, constructed languages can be further distinguished and categorized based on their various purposes for being created.

Despite their distinction in commonly accepted definitions, however, the separation between natural and constructed language is not always so clear when analyzed in greater detail (Goodall 2022). In fact, it can be argued that the question of what, if any, differences exist between the two is, at its core, a mere part of the larger debate surrounding what exactly constitutes a language.

1.1 Motivation

In many ways, the motivation behind this study relates to search for and defining of language universals is not only fundamental to the field of linguistic research, but also a topic of widespread debate. At present, many theoretical and foundational contributions exist, ranging from Greenberg's proposed universals (Greenberg 1970) to ideas about universal grammars, such as Chomsky

1.2 Scope of Study & Research Question

As previously noted, the specific boundaries which separate constructed from natural languages are not always clearly or consistently defined—even amongst linguists.

This study analyzes the linguistic features of constructed and natural languages and seeks to make a comparison on the differences, if any exist, between the two. Because of the wide-ranging nature of conducting such a broad analysis, there will of course be many features left unconsidered. 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 entropy, morphological complexity, and lexical diversity of each language, based on the corpora used. Ultimately, I seek to contribute to answering the age-old question of what defines a language.

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 studies delving into the intricacies of languages, both natural and constructed. This section will begin with a brief examination into constructed languages, defining what they are as well as their various purposes, as this is crucial to the study.

Finally, I will discuss the various computational methods implemented and provide some background information on how they work.

2.1 Constructed Languages

Constructed languages are actually a broad category of languages comprised of several sub-genres, differentiated according to their intended purpose for being created. These include philosophical languages, international auxiliary languages,

2.2 Prior 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-two languages are analyzed in this study. Four of these are constructed languages: Esperanto, Interlingua, Lingua Franca Nova, 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

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.

3.4.1 Type-Token Ratio

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

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

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

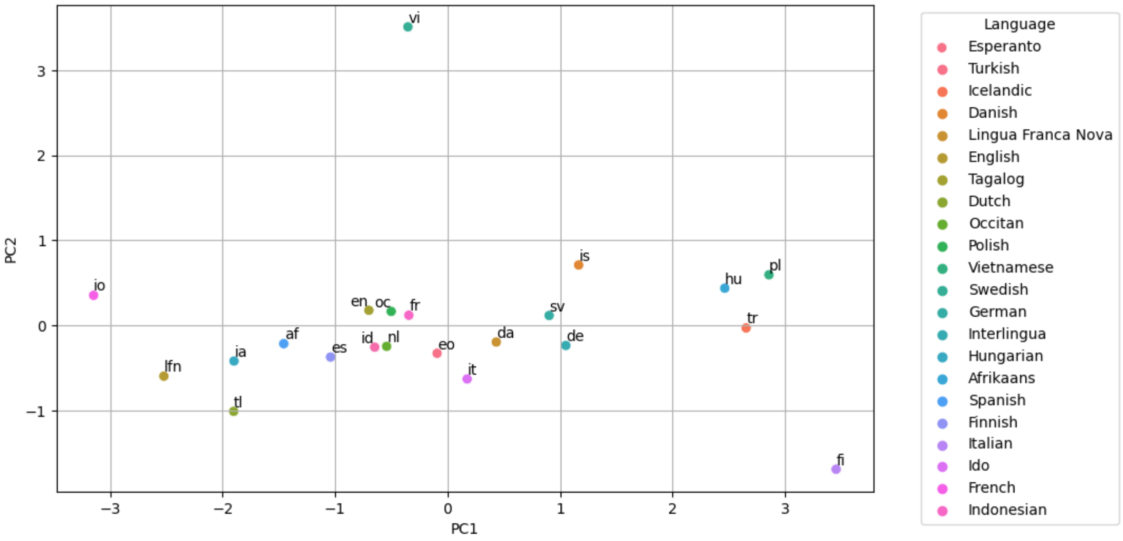

Figure 3.1: Principal Component Analysis on TTR, MATTR, Word and Char Distribution Entropy

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:

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

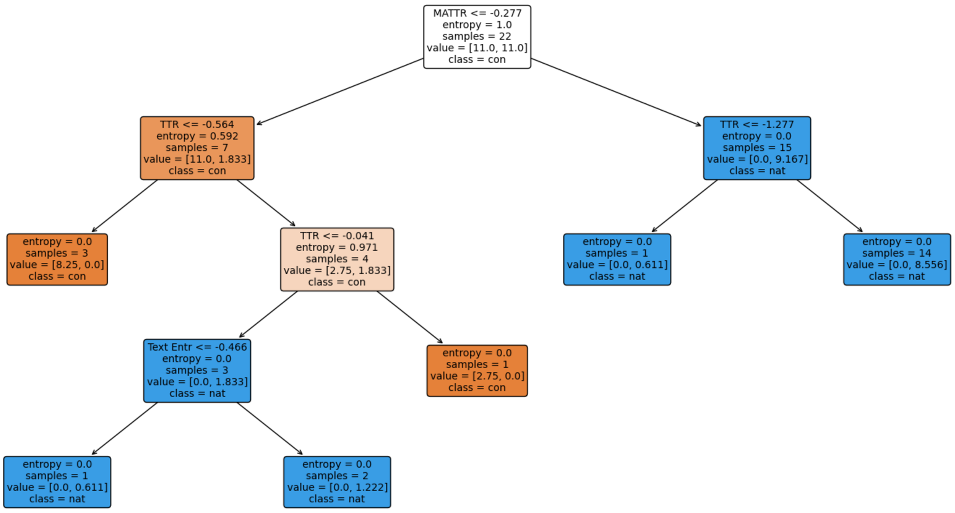

Figure 3.2: Decision Tree Classifier

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].

3.5 Classification Models

3.5.1 Decision Tree

Decision Tree Classifier...

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

3.5.2 Outlier Detection

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

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`.

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`.

Oostendorp, Marc van (2019). “11. Language contact and constructed languages”. In: *Volume 1*. Ed. by Jeroen Darquennes, Joseph C. Salmons, and Wim Vandenbussche. Berlin, Boston: De Gruyter Mouton, pp. 124–135. ISBN: 9783110435351. DOI: `doi:10.1515/9783110435351-011`. URL: `https://doi.org/10.1515/9783110435351-011`.

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