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 10, 2024

_____

Firstname Surname

Contents

Abstract

Write the abstract here.

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
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
SOV	Subject-Object-Verb
IALA	International Auxiliary Language Association
LFN	Lingua Franca Nova

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. This will be discussed more comprehensively 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, 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, this section begins with a brief overview of their history and development, which provides some relevant context. 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; Schubert et al. 2001), with hundreds proposed for the purpose of being IALs in Europe alone (Schubert et al. 2001). Yet of these, only Esperanto is commonly considered to be successful in achieving its creator's intended goal of world-wide use as an auxiliary language (or rather that it is by far the most successful), with very few others even coming close, having a conservative estimation of two million speakers (Okrent 2009).

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 influ-

(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" in his philosophical language (Goodall 2022)

ential. Wilkins created a system of semantic categorization, cataloging all concepts in the universe (Okrent 2009), and then published his proposed language (Wilkins 1968). An example of this hierarchal categorization can be seen in Figure 2.1.

In the 19th and 20th centuries the focus for language construction, especially 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 the Indo-European language family (Novikov 2022; Goodall 2022), so such a description might not be apt. A more detailed explanation of each of the IALs used in this study is provided in Section 3.1

Finally, there exist constructed languages that have been made for experimental, artistic, literary, or fictional purposes. In contrast to IALs, these languages are not made with the intention of replacing existing languages for everyday communication. Instead, their creators want to push the boundaries of language, test scientific hypotheses like linguistic relativity, or create a world, as is the case for the fictional examples provided in Section 1. Some other examples in this category include Solresol, a language that uses musical notes; Láadan, a language designed to be inherently feminist (i.e. more capable of expressing the female experience); and Loglan, a self-described "logical" language whose morphology and syntax are based on predicate logic (Adelman 2014). Though it would be inaccurate to describe such languages as being only a recent invention, popularity in their conceptualization grew in the later part of the 20th century.

While all share the defining characteristic of having been purposefully created, the linguistic features of constructed languages (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 second-language acquisition, reflected in commonly assigned features such as SVO word orders, head-initial relative clauses, fronted *wh*-phrases, and morphological regularity (Goodall 2022; Gobbo 2016). Section 2.2 further examines research focused on linguistic features of these languages.

In addition to this classification based on their intended communicative functions, i.e. as philosophical or international auxiliary languages, there are also taxonomies based on other criterion. For example, another frequently used distinction is that of *a priori* and *a posteriori* (Schreyer and Adger 2021; Gobbo 2008; Schubert 1989; Schubert et al. 2001; Novikov 2022; Adelman 2014; Tonkin 2015). Languages described as being *a priori* are structurally entirely new (Tonkin 2015) and not based on existing languages, whereas so-called *a posteriori* languages are the opposite, drawing from aspects of specific natural languages (Schreyer and Adger 2021). Gobbo (2008) also proposed the dichotomy of *exoteric* (secret) and *esoteric* (public) languages, derived from Bausani (1974). Similar to critiques regarding the distinction between constructed and natural, such dichotomies for categorizing constructed languages are also argued by some linguists to be more accurately described as scales instead, with many languages falling somewhere in the middle (Novikov 2022). A final noteworthy classification scheme often cited by other linguists comes from BLANKE (1989) in the form of three classes: project, semi-planned, and planned. In short, these correspond to a set of steps that a constructed language must go through before it can be considered a "real" language (Schubert et al. 2001).

A two-dimensional taxonomy for constructed languages containing several notables examples is shown in Figure 2.2 (Gobbo 2016).

2.2 Prior Studies

In contrast to the abundance in literature and cross-linguistic analyses done on natural languages, similar research which also includes constructed languages is relatively sparse. In particular, while there is research that analyzes specific instances of linguistic differences between certain natural and constructed languages, large-scale cross-linguistic studies which utilize computational methods to classify the two based on linguistic features are practically nonexistent. Consequently, the present study is a somewhat novel approach. However,

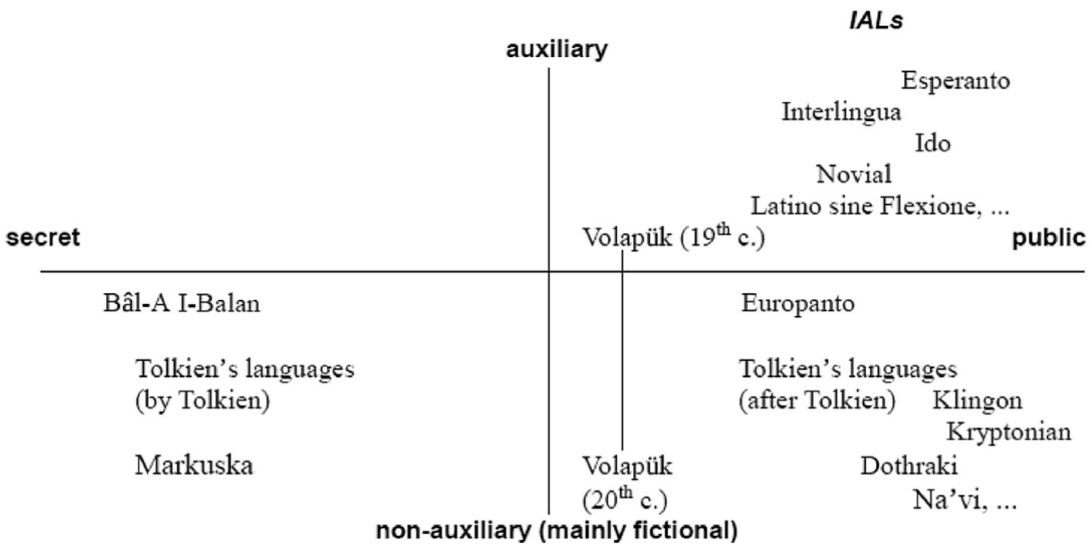

Figure 2.2: A taxonomy of constructed languages (Gobbo 2016)

there is precedent for this research and the specific features examined, as well as computational approaches used, which this section will describe.

As noted in the previous section, the creation of IALs often involved the intentional simplification of particular linguistic features to facilitate language acquisition, for instance having more regularity in their morphological systems. Intuitively, then, one would assume this translates to measurable differences in various aspects of linguistic complexity when compared to natural languages, which often have irregularities as a result of their development and evolution. When comparing Volapük and English, **GObbo2016** concluded that

Much of the literature on constructed languages focuses on Esperanto specifically.

3 Methodology

In this section, I introduce my data and discuss the steps taken for preprocessing it. I then discuss the linguistic features examined along with the different methods involved in extracting them from the data, and subsequently the classification approaches and models employed on the feature set. The number of possible features and measurements of linguistic complexity which could be analyzed in such a study is extensive to say the least; however, the scope of this particular thesis focuses mainly on various empirical measurements of morphological complexity and entropy. More specifically, the features I investigate are morphological complexity, type-token ratio (TTR), moving-average type-token ratio (MATTR), lexical entropy, text entropy, and character and word distribution entropies. Once these values were calculated for each language, the task became that of binary classification using these values with two machine learning models: a one-class support vector machine (SVM) and a decision tree. Finally, I also include a brief description of the various APIs and libraries used.

3.1 Data

In total, twenty-four languages are analyzed in this study.

3.1.1 Constructed Languages

The six constructed languages used for this study are Esperanto, Ido, Interlingua, Lingua Franca Nova, Volapük, and Kotava. All are IALs and were chosen primarily due to their availability as parallel texts, as well as having data which is comparable in size. I will briefly introduce each of them in this section, explaining where they come from, some notable typological features they have, and how they compare to each other.

Esperanto, the most widely-spoken constructed language and considered by many to be the most successful (Gobbo 2008), was created in 1887 by Polish ophthalmologist L. L. Zamenhof. Zamenhof's goal was to create a neutral, easy-to-learn language that would facilitate international communication. Esperanto is a highly regular language, with consistent grammar and a simplified, phonetic spelling system. It draws its lexical roots and syntax primarily from Romance, Germanic, and Slavic languages (Gobbo 2008; Gobbo 2011), making it recognizable and familiar to speakers of many European languages, while also intentionally being made to have a comparatively simpler grammar that avoids some complexities found in natural languages, such as irregular verbs or noun cases. It also has a strong global community with speakers around

the world, an array of written literature, and even a number of native speakers who learn it from birth—a distinguishing trait which sets it apart from other constructed languages (Goodall 2022). As a result of its success, Esperanto also serves as a direct influence for many other constructed languages that have come after it, one being Ido.

Ido is a reform of Esperanto that was proposed in 1907 by a group of linguists led by Louis Couturat, a French philosopher and mathematician, and in fact is an Esperanto word meaning "offspring" (Schubert et al. 2001). Its creators sought to address what they saw as imperfections in Esperanto, particularly those related to orthography and morphology. For instance, Ido avoids the use of the accusative case and reforms some Esperanto words to make them more universally recognizable. Overall, though, Ido still retains much of Esperanto's vocabulary and basic structure, and the two are mutually intelligible to a large extend (Goodall 2022; Schubert et al. 2001). Like most of the remaining constructed languages to be discussed in this section—with the exception of Volapük—Ido has small but a dedicated community of speakers and enthusiasts.

Interlingua was developed by the International Auxiliary Language Association (IALA) with the assistance of linguist Alexander Gode, officially being published in 1951. The idea behind Interlingua for it to most recognizable to the greatest number of people without requiring prior study (Goodall 2022), with most attention having been spent on its lexicon. The IALA's stated goal was to not so much create a new international language, but rather present a standardized international vocabulary (Large 1985) ("international" here basically referring to Western Europe). It is largely derived from and resembles Romance languages (with lesser influence from Greek and Germanic languages) (Schubert et al. 2001). In fact, this intentional resemblance extends even to morphological irregularities such as allomorphy, with other irregularities also being introduced to the language to make it appear more natural (Goodall 2022), a contrast to other IALs like Esperanto.

Volapük was created in 1879 by Johann Martin Schleyer, a German Catholic priest who believed the language had been given to him by God. It features highly agglutinative structure and regular, yet complex, morphology. While being derived mainly from English, German, and Latin, roots in Volapük differ significantly to the point of being unrecognizable to speakers of these languages (Goodall 2022). Despite being argued to be the first successful constructed language due to its rise in popularity, having amassed a large number of supporters worldwide along with the formation of clubs and societies (Gobbo 2016), various issues regarding its complexity led to a rapid decline and eventual fall from usage in favor of Esperanto.

Lingua Franca Nova, also abbreviated as LFN, is a relatively recent con-

structed language created by linguist C. George Boeree in 1998. Its lexicon is based mainly on Romance languages, specifically French, Italian, Portuguese, Spanish, and Catalan, while its grammar is based on Romance creole languages (Pawlas and Paradowski 2020). In particular, inspiration came from the similarly-named Mediterranean Lingua Franca, a pidgin that developed for trade in the Mediterranean basin and was used from the 11^{th} to 18^{th} centuries, as well as from other creoles, such as Haitian Creole. It can be written in both Latin and Cyrillic script.

The last constructed language used is Kotava. Created by Staren Fetcey in 1978, Kotava stands out in this dataset as being a unique attempt at creating a culturally neutral *a priori* language, free from any biases or influences of existing languages and based on the philosophy of linguistic egalitarianism, especially in regards to its lexicon. Its morphology, syntax, and phonetic system are all intentionally made to be simplified and regular.

Table 3.1 shows these languages together with the main source languages they draw from (if any). Note, however, that this is not an exhaustive list of all of their language influences.

Constructed Language	**Source Languages/Families**
Esperanto	Romance, Germanic, Slavic
Interlingua	Romance
Lingua Franca Nova	Romance
Volapük	Germanic
Kotava	N/A
Ido	Romance, Germanic, Slavic

Table 3.1: Constructed languages used in the study, together with their main respective source languages from which they were designed.

It is worth drawing attention to the fact that each of these languages were constructed mainly from various European languages, with the exception of Kotava. Consequently, this may influence models performing classification and be visible in the results. This will be explored later in Section 5.

3.1.2 Natural Languages

Eighteen natural languages belonging to five different language families were used: German, English, Spanish, Polish, Vietnamese, Indonesian, Turkish, Tagalog, Hungarian, French, Finnish, Italian, Dutch, Occitan, Danish, Swedish, Afrikaans, and Icelandic. The language families represented are Austroasiatic, Austronesian, Turkic, Uralic, and Indo-European, with the latter comprising the bulk of the natural languages used. This diversity was intentional,

3.1.3 Wikimedia

The data for this thesis comes from Wikimedia dump files. Wikimedia is a global movement and community founded on shared values, whose goal is to provide free and openly accessible information to everyone in the form of massive collaborative projects (which include, among others, the widely-used Wikipedia and Wiktionary). For a large, cross-linguistic study, massive databases with open-access make for an ideal source for corpora. Most importantly, the projects are multilingual, meaning the databases are available in a considerable number of different languages—including several constructed languages. This allows for composing parallel corpora. Additional constructed languages which are also available from these dumps but were not included in the present study due to having a much smaller amount of data are Novial, Interlingue, and Lojban.

The dump files provide detailed, archived snapshots of the content from Wiki repositories for a specified point in time and are available in different formats. All dumps used were XML-formatted and from the 2024-07-01 archive, containing articles together with their metadata[1]. It is also worth mentioning here that there are some drawbacks to using these dumps for the present study. The files sizes vary considerably depending on the language, with the largest being roughly 22 gigabytes (English) and the smallest around 4 megabytes (Lingua Franca Nova), meaning all files do not contain the exact same articles. Additionally, the open and collaborative nature of Wikimedia means the articles are often authored by a multitude of different people, which can result in inconsistencies in the texts, such as with writing style. Similarly, it may also produce an imbalance in the amount of information provided across languages, with the same article in one language being considerably more detailed than in another, and inconsistent or low-quality translations, as Novikov (2022) noted to be the case for Wikipedia articles in Volapük. Thus, while Wikimedia was decided as the best available option for the task at hand, there are some unfavorable aspects of using it which may influence the results; this will be discussed more in Section 5.

3.2 Data Preprocessing

Preprocessing text data is essential for natural language processing (NLP) tasks, and since I essentially had to compile my own corpora for this study, meticulous effort was made to thoroughly clean all of the texts and obtain as close to a set of parallel data as possible.

Text data was first extracted from the Wikimedia XML-formatted dump

[1] https://dumps.wikimedia.org/backup-index.html

files with the use of WikiExtractor, a Python script (Attardi 2015) that I adapted from the original by adding a limit to the number of articles in order to make extraction of the largest of the files (English in particular) less demanding and quicker. The output is a simple text file, which is easier to clean.

I then used several regular expressions to remove general, unnecessary text from each file such as page titles, section headers, links, fragments, HTML tags, braces, and all other non-alphabet symbols aside from periods. This also includes the removal of parantheses and their contents. The text was then made all lowercase and split by the periods—while also attempting to account for abbreviations—to make separate sentences. This was done mainly to enable more accurate measurement of entropy later.

Following this, foreign symbols (i.e., characters not part of a language's alphabet) were removed for each text/language, as occasionally proper nouns, loanwords, etc. would appear in the text, which would also affect measurements of entropy, in addition to morphological segmentation and analysis. For example, there is no letter 'h' in Kotava, but this would sometimes be found in the original extracted text in loanwords such as 'Hiroshima'. After being cleaned, the remaining word is 'irosima'.

Finally, each text file was truncated according to the file size of the smallest corpus, LFN, so as to have similar lengths. This was calculated based on number of words, with the limit being 630000 (since this is roughly the number of words remaining in the LFN text file after cleaning), and while preserving complete sentences. Sentences containing only one word were also removed. The end result of pre-processing was a single text file corresponding to each language, with each line in the file being a single sentence. The corpora with the smallest and largest number of words is Kotava and Danish/Volapük at 617400 and 629999 words, respectively. For number of sentences, the smallest and largest corpora are Vietnamese and Volapük at 21115 and 55920 sentences, respectively. For a breakdown of these size for each language's text after pre-processing, refer to Table 8.1.

3.3 Libraries and APIs

3.4 Feature Extraction

Since this study uses models for classification, it is necessary to first extract the features...

3.4.1 TTR & MATTR

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

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

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

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

3.4.2 Morphological Complexity

The morphological systems of each language were analyzed using Morfessor,

3.4.3 Entropy

In information science, entropy means...

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

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

3.4.4 Additional Features

3.4.5 PCA

Principal Component Analysis was performed for dimensionality reduction.

3.5 Classification

3.5.1 Decision Tree

Decision Tree Classifier...

3.5.2 Random Forest

3.5.3 One-Class SVM

One-Class SVMs are a special kind of SVMs used in the domain of anomaly detection. While similar to one another, classic SVMs separate two classes using a hyperplane with the largest possible margin. In contrast to this, One-Class SVMs...Rather than training on an entire dataset containing two classes as with classic SVMs, these models train exclusively on the majority class, also called the "normal" class. After training,

Standardization is given by the formula

$$X' = \frac{X - \mu}{\sigma}$$

where μ is mean and σ is the standard deviation.

3.6 Evaluation of Classifiers

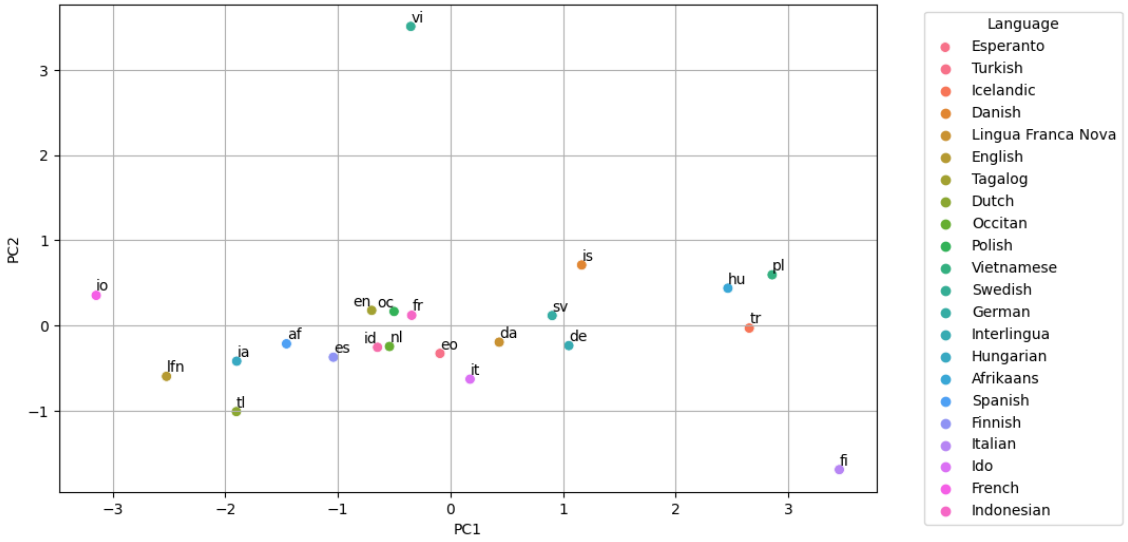

Figure 4.1: Principal Component Analysis on TTR, MATTR, Word and Char Distribution Entropy

4 Results

This section reports the results of each of the methods implemented.

4.1 Results of TTR & MATTR

4.2 Results of Morphological Segmentation

4.3 Results of Feature Extraction

4.4 Results of PCA

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

4.5 Results of One-Class SVM

4.6 Results of Decision Tree

4.7 Results of Random Forest

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

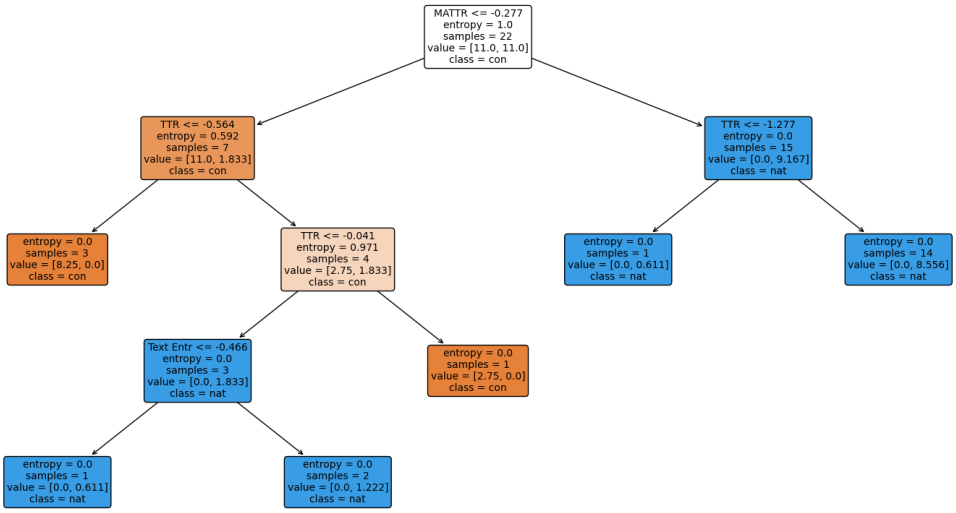

Figure 4.2: Decision Tree Classifier

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.

Attardi, Giuseppe (2015). *WikiExtractor*. https://github.com/attardi/wikiextractor.

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.

Bausani, A. (1974). *Le lingue inventate: Linguaggi artificiali, linguaggi segreti, linguaggi universali*. Collana di studi umanistici 'Ulisse'. Ubaldini. ISBN: 9788834003879. URL: https://books.google.de/books?id=z4GAngEACAAJ.

BLANKE, DETLEV (1989). "Planned languages – a survey of some of the main problems". In: *Aspects of the Science of Planned Languages*. Ed. by Klaus Schubert. Berlin, New York: De Gruyter Mouton, pp. 63–88. ISBN: 9783110886115. DOI: doi:10.1515/9783110886115.63. URL: https://doi.org/10.1515/9783110886115.63.

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

– (Sept. 2011). "The Case of Correlatives: A Comparison between Natural and Planned Languages". In: *Journal of Universal Language* 12, p. 34. DOI: 10.22425/jul.2011.12.2.45.

– (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 Uni-

versity 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.

Pawlas, Elżbieta and Michał B. Paradowski (Jan. 2020). "Misunderstandings in communicating in English as a lingua franca: Causes, prevention, and remediation strategies". In: pp. 101–122. ISBN: 978-83-66666-28-3. DOI: 10.48226/978-83-66666-28-3.

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*.

Tonkin, Humphrey (Apr. 2015). "Language Planning and Planned Languages: How Can Planned Languages Inform Language Planning?" In: *Interdisciplinary Description of Complex Systems* 13, pp. 193–199. DOI: 10.7906/indecs.13.2.1.

Wilkins, John S. (1968). "An essay towards a real character, and a philosophical language, 1668". In: URL: https://api.semanticscholar.org/CorpusID:161991811.

8 Appendices

Here I...

Corpus	Type	Avg Word Length	Avg Sentence Length	TTR	MATTR	Char Dist Entr	Word Dist Entr	Text Entr	Lex Entr	Rev Lex Entr
id	nat	6.173	18.164	5.782	0.699	4.072	11.142	3.518	1.956	1.976
tl	nat	5.119	21.102	7.593	0.611	3.895	9.991	3.824	1.884	1.917
tr	nat	6.63	14.458	14.097	0.828	4.386	13.151	4.114	1.562	1.656
en	nat	5.087	21.301	6.079	0.697	4.167	10.673	4.116	1.926	1.981
de	nat	6.206	16.907	12.128	0.771	4.23	11.601	3.965	1.608	1.666
fr	nat	5.16	23.12	7.461	0.721	4.179	10.711	3.497	1.793	1.865
eo	con	5.175	18.909	10.708	0.692	4.164	10.923	3.858	1.801	1.893
lfn	con	4.221	19.532	5.063	0.601	3.912	9.316	3.936	2.027	2.114
ia	con	5.05	19.547	6.88	0.607	4.032	10.005	3.336	1.821	1.906
io	con	4.594	14.484	3.433	0.557	4.077	8.055	1.157	1.985	2.069
pl	nat	6.248	14.951	14.89	0.825	4.553	12.905	4.316	1.651	1.685
vi	nat	3.498	29.835	1.749	0.732	4.855	9.717	4.001	2.421	2.387
fi	nat	7.874	11.969	20.409	0.841	4.144	13.729	3.915	1.547	1.631
it	nat	5.455	25.727	8.505	0.764	4.029	11.308	4.003	1.672	1.786
af	nat	5.067	20.496	6.987	0.645	4.072	9.993	4.088	1.839	1.914
nl	nat	5.419	18.194	8.559	0.694	4.117	10.593	3.813	1.811	1.866
es	nat	4.978	25.315	7.085	0.674	4.046	10.327	3.502	1.759	1.864
oc	nat	5.215	18.66	7.185	0.715	4.173	10.546	2.963	1.871	1.934
da	nat	5.346	16.466	10.517	0.737	4.197	11.274	4.342	1.808	1.87
sv	nat	5.597	17.322	11.031	0.756	4.294	11.488	4.17	1.775	1.836
is	nat	5.375	15.055	11.727	0.747	4.468	11.512	4.643	1.728	1.796
hu	nat	6.242	15.782	16.234	0.776	4.543	12.443	4.423	1.67	1.727
vo	con	5.072	11.266	2.455	0.622	4.256	7.666	1.192	2.086	2.135
avk	con	5.06	12.824	8.153	0.582	4.186	10.287	3.085	2.011	2.066

Table 4.1: Feature set

Language	**Number of Words**	**Number of sentences**
Icelandic	629995	41847
German	629987	37261
Polish	629997	42138
Ido	629990	43496
Afrikaans	629994	30737
Kotava	617400	48145
Hungarian	629946	39916
Lingua Franca Nova	628683	32188
Danish	629999	38260
Spanish	629978	24886
Interlingua	629996	32229
French	629983	27248
Occitan	629998	33762
Esperanto	629994	33317
Dutch	629997	34627
Turkish	629995	43573
English	629958	29574
Tagalog	629989	29855
Swedish	629998	36370
Vietnamese	629958	21115
Italian	629987	24487
Volapük	629999	55920
Indonesian	629997	34683
Finnish	629994	52637

Table 8.1: Lengths of each language's text after pre-processing.