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 19, 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	Principal 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
CSV	Comma-Separated Values

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, the investigation into the disparity between the these two kinds of languages overlaps with the broader debate regarding what constitutes a language. Central to this debate is the search for linguistic universals—

properties shared by all languages (Mairal and Gil 2006). The concept of universals in language is recognized as one of the most important areas of research in linguistics (Christiansen, Collins, and Edelman 2009) and has served as a foundation for much linguistic theory, especially in more recent history, stemming largely from the influential theories and works of Greenberg (Greenberg 1970) and Chomsky (Chomsky 1957; Cook and Newson 2007).

Analyzing their surface structures can reveal whether or not constructed languages adhere to the same linguistic conventions as their natural counterparts. If machine learning models fail to successfully distinguish between the two, it may reinforce the notion that these universals are present in all languages, regardless of origin. Conversely, the models succeeding may suggest the opposite. In short, the primary motivation behind this thesis is to contribute to this ongoing debate through the application of machine learning, and a desire to learn more about the fascinating genre of constructed languages.

1.1 Scope of Study & Research Question

The present work analyzes various linguistic features and seeks to successfully discriminate a language as being either natural or constructed based on these. More specifically, the scope of this study includes both binary classification and anomaly detection, with the models being trained on a set of selected features rather than raw text data.

Because of the wide-ranging nature of conducting such a broad analysis, there are of course 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 main focus for linguistic features relate to entropy, morphological complexity, and overall lexical diversity.

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, which is relevant to understanding the motivation behind the various computational approaches I employ in my experiments.

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 (though not always) 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. Overall, IALs can be viewed as an intended rival to natural languages, which is one reason why all of the constructed languages analyzed in the present work are IALs. A more detailed explanation of each is provided in Section 3.1

Lastly, 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 largely grew in the later part of the 20th century.

While all share the defining characteristic of having been purposefully cre-

ated, 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 meth-

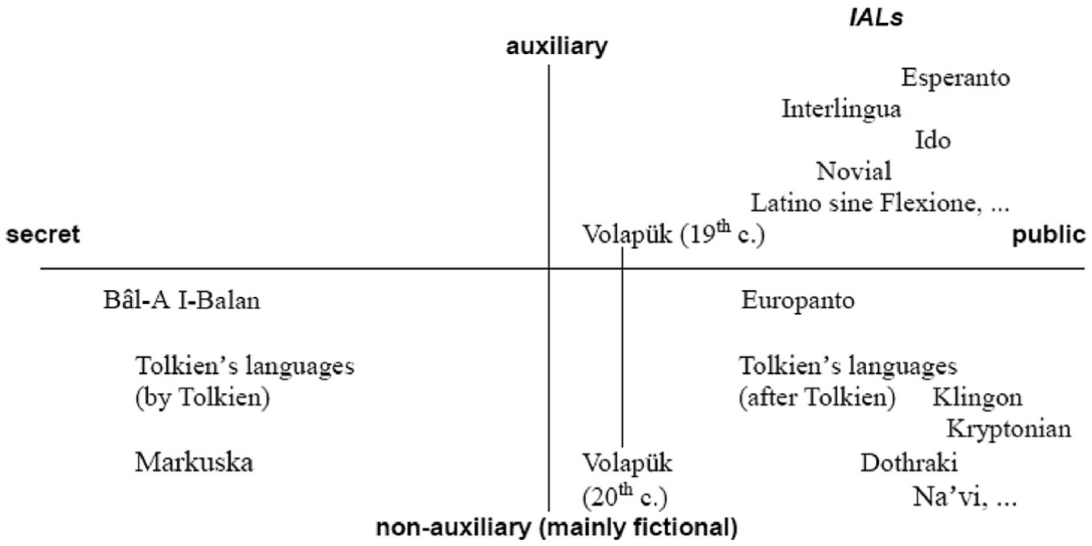

Figure 2.2: A taxonomy of constructed languages (Gobbo 2016)

ods to classify the two based on linguistic features are practically nonexistent. Consequently, the present study is a somewhat novel approach. However, 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.

Another feature examined is entropy. Originating from information science, entropy was introduced by Shannon (1949) as a measurement of uncertainty or surprisal for an event, with high surprisal being inversely proportionate to the amount of information conveyed by the event's occurrence. In NLP,

3 Methodology

In this section, I introduce the dataset for this paper and discuss the steps taken for preprocessing it, followed by discussing in detail the features examined along with the different methods involved in extracting them from the data, and finally the classification and anomaly detection models employed on the feature set. A brief description of the various APIs and libraries used is also included in 3.3.

Since this study involves many different experiments and elements being analyzed, I will begin by explaining an overview of what all was done. 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 thesis focuses mainly on empirical measurements relating to lexical diversity, morphological complexity, and entropy, along with some others measurements which are commonplace to calculating linguistic complexity, and thus seemed appropriate to also include. More specifically, the features investigated are average word length, average sentence length, type-token ratio (TTR) of morphemes, average number of segmentations in a word, average number of forms per lemma, lexical TTR and the related moving-average type-token ratio (MATTR), lexical entropy, text entropy, and character and word distribution entropies. Once the values of these were calculated for each language, the task became that of classification and anomaly detection with four machine learning models: a one-class support vector machine (SVM), random forest, isolation forest, and decision tree. Lastly, the methods for evaluating the performances of these models are discussed, as is the Principal Component Analysis (PCA) performed on the data.

3.1 Data

In total, twenty-four languages are analyzed in this study. Six of these are constructed languages: Esperanto, Ido, Interlingua, Lingua Franca Nova, Volapük, and Kotava. The remaining eighteen are natural: German, English, Spanish, Polish, Vietnamese, Indonesian, Turkish, Tagalog, Hungarian, French, Finnish, Italian, Dutch, Occitan, Danish, Swedish, Afrikaans, and Icelandic.

For consistency, only languages which are written using the Latin alphabet (including the use of diacritics) were chosen. This is mainly because the constructed languages in the dataset all use Latin alphabets, so the selection of natural languages followed the same criteria. Moreover, it allows for more uniform cross-linguistic analysis of features which are sensitive (e.g. in the case of character entropy) to writing systems.

3.1.1 Constructed Languages in the Dataset

All of the constructed languages in the dataset are IALs, with most of them resembling natural (particularly various European) languages. 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 both one another other and their natural counterparts.

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 constructed 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 scripts, though this dataset only contains the former.

The last constructed language used is Kotava. Created by Staren Fetcey in 1978, Kotava stands out in this dataset as being an attempt at creating a culturally neutral *a priori* language, free from any biases or influences of existing languages and based on a philosophy of linguistic egalitarianism. This intentionally designed uniqueness manifests in several of its linguistic systems, from morphology to syntax. For example, though word order in Kotava is not imposed, the most frequently used one is OSV, which is exceedingly rare in natural language. Other unique features include a 4^{th} person plural, object complements being introduced by a transitive preposition, and a lack of declension (Fetcey and Comité Linguistique Kotava 2013).

Finally, it is worth drawing attention to the fact that each of these languages were created based on various European languages, with the exception of Kotava. Consequently, this may influence the models used in the experiments and be visible in the results. This will be explored in greater detail later in Section 5.

Table 3.1 shows the dataset's constructed languages together with the main

Conlang	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.

source languages they draw from (with *N/A* for Kotava meaning *not applicable*). Note, however, that this is not an exhaustive list of all of their language influences.

3.1.2 Natural Languages in the Dataset

The natural languages included in this study represent a broad spectrum of linguistic diversity, comprising a variety of families, geographic regions, and typological features. Although this representation is not necessarily equal in distribution, it is meant to serve as a contrast to the constructed languages in the dataset, which lack a similar extent of variety due to overwhelmingly having the same source languages. However, rather than delving into the same level of details for each of the eighteen languages here as I did for the constructed ones, I will instead introduce them by focusing mostly on their collective significance and summarizing some of their relevant typological traits.

In total, five major language families are represented. The largest of these—based on number of speakers worldwide as well as the number of languages in the dataset (twelve)—is Indo-European, with several of its branches being included. English, German, Dutch, Afrikaans, Swedish, Icelandic, and Danish all belong to the Germanic branch. Similarly, Italian, Spanish, French, and Occitan are part of the Romance branch, all being descendants of Latin. Polish is an outlier as the only represented language from the Slavic branch.

The other four families span less representation in the dataset in comparison, but were nevertheless included so as to have more variety in linguistic features. These include the Uralic languages, consisting of Hungarian and Finnish, which are the most widely-spoken and thus representative of their group, as well as Austronesian (Tagalog and Indonesian), Austroasiatic (Vietnamese), and Turkic (Turkish).

One notable typological feature of all of these languages which relates to the scope of this study is that of their morphology. Traditionally, classification of morphological systems follows two paradigmatic axes: inflection versus derivation and agglutination versus flexion.

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 data is available in a considerable number of different languages—including several constructed ones. This allows for composing a set of corpora which is adequately parallel to each other and from the same domain. Additional constructed languages which are also available from the 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, so 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 files with the use of WikiExtractor[2], a Python script (Attardi 2015) that I

[1] `https://dumps.wikimedia.org/backup-index.html`
[2] GitHub repository for WikiExtractor: `https://github.com/attardi/wikiextractor`

adapted 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 a much easier format to clean and work with.

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 particular 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. To give an example of this, there is no letter 'h' in Kotava, but this would sometimes be used in proper nouns such as 'Hiroshima'. After the text is cleaned, the remaining word left behind 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 for each language, with every line in the file being a single sentence. The corpora with the smallest and largest number of words are 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 full breakdown of these size for each language's text after pre-processing, refer to Table 8.1.

3.3 Libraries and APIs

Several libraries and APIs were used in both the feature engineering and classification steps of the experiment, and the most important of these will be briefly introduced here. In the field of machine learning, two of the most popular model frameworks used are PyTorch and TensorFlow, which are interacted with via the torch and Keras APIs, respectively. Aside from some relatively minor differences (e.g., the syntax of their code and performance optimization), they share a lot of similarities and are typically used according to personal preference. In the context of this paper, these frameworks are

used for calculating some of the entropy values from the corpora. In addition to libraries used for constructing the model architectures, there are also ones used for the data itself. Arguably the most fundamental for this is NumPy, which stands for Numerical Python and is used to accomplish extremely fast and efficient computation of arrays. Other essential libraries include Pandas, used for data manipulation and analysis, and Matplotlib, used for visualizing data and plotting model results. Lastly, scikit-learn provides a wide range of tools for machine learning algorithms, data preprocessing, and model evaluation—as well as computation, thanks to it being built on top of NumPy. The models I used for classification (i.e., One-Class SVM, Random Forest, Decision Tree) as well as PCA come from this library. Altogether, these libraries are often used in tandem in NLP tasks due to integrating so well with one another.

3.4 Feature Engineering

Before classification or anomaly detection can be done with the data, an initial step of feature engineering is performed. Put simply, feature engineering is the process of transforming raw text data into a more structured and comprehensible format for machine learning models through the specific selection of its most informative and relevant features, thereby increasing the model's effectiveness.

In the scope of this paper, this process involved several dimensions of cross-linguistic analysis and measurement.

The following features are calculated for each language in the dataset.

3.4.1 Lexical Diversity

A common way of measuring the lexical diversity of a text is with TTR, with a high value indicating that a given text contains a large amount of lexical variation. This is calculated using the formula:

$$\text{TTR} = \frac{|\text{V}|}{|\text{N}|}$$

where |V| denotes the vocabulary size as the number of unique words, or types, and |N| denotes the text length as the total number of words, or tokens. I then multiply this by 100 to get a percentage.

A big issue with TTR, though, is that it does not always provide an accurate assessment due to its sensitivity to text length; the longer a particular text, the higher the likelihood of repetition in words occurring, consequently affecting the calculation. To remedy this, I also calculate the MATTR, a variation of TTR proposed by Covington and McFall (2010) that uses a sliding

window of a fixed-length over the text and calculates the TTR at each length of the window, which is then averaged together. This is denoted by the formula:

$$\text{MATTR}_i = \frac{1}{N-i+1} \sum_{j=1}^{N-i+1} \frac{|\text{Types}_{j,j+i-1}|}{|\text{Tokens}_{j,j+i-1}|}$$

where N is the total number of tokens in the text, i is the window size, $|\text{Types}_{j,j+i-1}|$ is the number of unique words (types) in the window, and $|\text{Tokens}_{j,j+i-1}|$ is the total number of words (tokens) in the window.

While resistant to variation in text length, calculations for MATTR vary based on the window size, and deciding which value to use depends on the task. Here, a length of $i = 100$ tokens was used.

3.4.2 Morphological Complexity

The morphological systems of each language were analyzed using Morfessor 2.0.

After the text files were segmented,

3.4.3 Entropy

In total, five values of entropy were measured: text, lexical, reversed lexical, character distribution, and word distribution. The latter two were calculated using the formula for Shannon's entropy with a given text:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Where X refers to the random variable (e.g. the word or character distribution of a text), n is the number of distinct values (i.e. types) in the distribution, $p(x_i)$ shows the probability p of each type x_i occurring and is calculated by its relative frequency, and $-\log_2 p(x_i)$ represents the self-information for each type. Put together, $-p(x_i) \log_2 p(x_i)$ represents the mathematical expectation for a given type, and the sum of all of these is the entropy.

Text entropy was calculated using a character-level Long Short-Term Memory (LSTM) model built with PyTorch for text generation.[3] LSTMs are a type of Recurrent Neural Network (RNN) that are typically better at handling long-term dependencies in sequential data due to their gating mechanisms for retaining only the information deemed useful, consequently making them better suited for generative tasks.

[3]Model adapted from `https://github.com/LeanManager/NLP-PyTorch/blob/master/Character-Level\%20LSTM\%20with\%20PyTorch.ipynb`

Using this model required first creating a dictionary of the input text's characters, using this dictionary to map the characters in each text to their corresponding integers, then one-hot encoding them (converting them into binary vector representations). Once the text data was encoded, it was split into mini-batches to be used for training the model. A mini-batch is essentially a sliding window in the shape of an $N * M$ array of characters, where N is the number of sequences (equivalent to a batch size) and M is the number of steps, or length of the window. Additionally, this means the input must be divisible into full batches, so remainder text that is insufficient in size is discarded. These mini-batches are then fed to the model.

LSTM For Calculating Text Entropy	
Parameters	**Values**
Number of epochs	20
Number of sequences (`n_seqs`)	128
Number of steps (`n_steps`)	100
Number of hidden layers (`n_layers`)	2
Number of hidden units (`n_hidden`)	256
Learning rate	0.001
Dropout rate	0.5
Optimizer	Adam
Criterion	Cross-Entropy Loss
Fraction of data for validation	0.1
Gradient clipping	5

Table 3.2: Parameters of PyTorch LSTM used to Calculate Text Entropy

Table 3.2 shows the model's parameters alongside their associated values as used in this paper. Its architecture comprises 2 hidden layers (`n_layers`), a dropout layer, and a fully-connected output layer. Both `n_layers` contain 256 hidden units (`n_hidden`), which are used to store information from the input and essentially act as the memory of the LSTM. Default values for learning rate, dropout rate, and gradient clipping were used, and training was done over 20 epochs. Minimal finetuning was performed overall, and primarily just for the `n_seqs` and `n_steps` hyperparameters.

Additionally, a fraction (0.1) of the initial input data is set aside to be used for validation, from which the model's perplexity—a measure related to entropy—is derived. To arrive at this measurement though, the validation loss is first calculated using (multi-class) cross-entropy. Cross-entropy is an extension of Shannon's entropy, but measures instead the difference between a model's predicted probability distribution and the true one. More formally, this is represented as:

$$H(p, q) = -\sum_{x=i} p(x) \log_q(x)$$

Where p and q represent the discrete predicted and true probability distributions, respectively, and with the natural logarithm $\log_e$, as commonly used by machine learning libraries such as PyTorch. As the predicted distribution gets closer to the true one, the resulting cross-entropy becomes lower. Entropy $H(q)$ is thus a lower bound of cross-entropy $H(p, q)$.

The validation loss was repeatedly assessed throughout training, and the mean of these was used to calculate the perplexity. Expressed mathematically, perplexity PP is simply an exponentiation of cross-entropy and can be denoted by the equation:

$$\begin{aligned} PP &= \mathrm{e}^{H(p,q)} \\ &= \mathrm{e}^{-\sum_{x=i} p(x) \log_q(x)} \end{aligned} \tag{1}$$

For calculating both lexical and reversed lexical entropies, an RNN built using TensorFlow was used. Early stopping was implemented based on the validation loss with `patience` set to 3.

Table 3.3 shows the most relevant parameters.

RNN For Calculating Lexical & Reverse Lexical Entropy	
Parameters	**Values**
Number of epochs	100
Batch size	32
Learning rate	0.001
Optimizer	Adam
Criterion	Sparse Categorical Cross-Entropy Loss
Fraction of data for validation	0.2

Table 3.3: Parameters of TensorFlow RNN used to Calculate Lexical and Reverse Lexical Entropy

3.4.4 PCA

PCA is an unsupervised method of dimensionality reduction, used for

3.5 Classification & Anomaly Detection

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

4 Results

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

4.1 Results of Lexical Diversity

4.2 Results of Morphological Segmentation

4.3 Results of Feature Engineering

Corpus	Type	Avg Word Length	Avg Sentence Length	TTR	MATTR	Morpheme TTR	Avg Segs Per Word	Avg Forms Per Lemma	Char Dist Entr	Word Dist Entr	Text Entr	Lex Entr	Rev Lex Entr
Volapük	con	5.072	11.266	2.455	0.622	0.145	2.175	3.254	4.256	7.666	1.192	2.086	2.135
Ido	con	4.594	14.484	3.433	0.557	0.103	2.238	4.402	4.077	8.055	1.157	1.985	2.069
Dutch	nat	5.419	18.194	8.559	0.694	0.107	2.169	4.383	4.117	10.593	3.813	1.811	1.866
Afrikaans	nat	5.067	20.496	6.987	0.645	0.124	2.098	3.941	4.072	9.993	4.088	1.839	1.914
Turkish	nat	6.63	14.458	14.097	0.828	0.09	2.023	5.573	4.386	13.151	4.114	1.562	1.656
Esperanto	con	5.175	18.909	10.708	0.692	0.096	2.127	4.982	4.164	10.923	3.858	1.801	1.893
Hungarian	nat	6.242	15.782	16.234	0.776	0.079	2.169	5.999	4.543	12.443	4.423	1.67	1.727
Tagalog	nat	5.119	21.102	7.593	0.611	0.103	2.16	4.59	3.895	9.991	3.824	1.884	1.917
Italian	nat	5.455	25.727	8.505	0.764	0.088	2.152	5.352	4.029	11.308	4.003	1.672	1.786
Swedish	nat	5.597	17.322	11.031	0.756	0.091	2.217	5.028	4.294	11.488	4.17	1.775	1.836
Icelandic	nat	5.375	15.055	11.727	0.747	0.091	2.16	5.181	4.468	11.512	4.643	1.728	1.796
Kotava	con	5.06	12.824	8.153	0.582	0.091	2.195	5.115	4.186	10.287	3.085	2.011	2.066
Danish	nat	5.346	16.466	10.517	0.737	0.099	2.18	4.727	4.197	11.274	4.342	1.808	1.87
Spanish	nat	4.978	25.315	7.085	0.674	0.113	2.089	4.285	4.046	10.327	3.502	1.759	1.864
Lingua Franca Nova	con	4.221	19.532	5.063	0.601	0.11	2.149	4.274	3.912	9.316	3.936	2.027	2.114
English	nat	5.087	21.301	6.079	0.697	0.116	2.136	4.091	4.167	10.673	4.116	1.926	1.981
Finnish	nat	7.874	11.969	20.409	0.841	0.083	2.172	5.727	4.144	13.729	3.915	1.547	1.631
Polish	nat	6.248	14.951	14.89	0.825	0.093	2.132	5.166	4.553	12.905	4.316	1.651	1.685
French	nat	5.16	23.12	7.461	0.721	0.11	2.167	4.269	4.179	10.711	3.497	1.793	1.865
Indonesian	nat	6.173	18.164	5.782	0.699	0.097	2.254	4.656	4.072	11.142	3.518	1.956	1.976
Vietnamese	nat	3.498	29.835	1.749	0.732	0.167	2.026	3.033	4.855	9.717	4.001	2.421	2.387
Occitan	nat	5.215	18.66	7.185	0.715	0.109	2.162	4.336	4.173	10.546	2.963	1.871	1.934
Interlingua	con	5.05	19.547	6.88	0.607	0.105	2.144	4.492	4.032	10.005	3.336	1.821	1.906
German	nat	6.206	16.907	12.128	0.771	0.094	2.207	4.961	4.23	11.601	3.965	1.608	1.666

Table 4.1: Feature set

4.4 Results of PCA

A script was used to increase readability of the text in the graph[4].

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

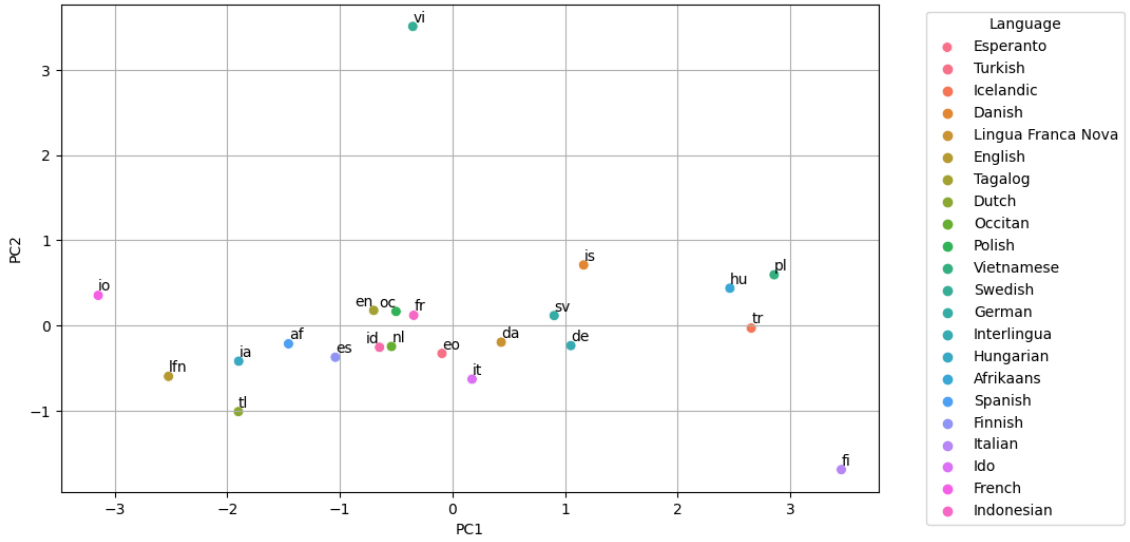

Figure 4.1: Principal Component Analysis on TTR, MATTR, Word and Char Distribution Entropy

4.5 Results of One-Class SVM

4.6 Results of Decision Tree

4.7 Results of Random Forest

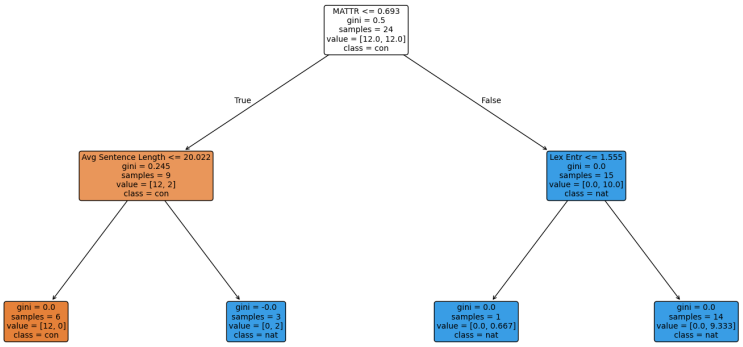

Figure 4.2: Decision Tree Classifier

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.

Chomsky, Noam (1957). *Syntactic Structures*. Berlin, Boston: De Gruyter Mouton. ISBN: 9783112316009. DOI: doi:10.1515/9783112316009. URL: https://doi.org/10.1515/9783112316009.

Christiansen, M.H., C. Collins, and S. Edelman (2009). *Language Universals*. Oxford University Press. ISBN: 9780190294113. URL: https://books.google.de/books?id=bCLiBwAAQBAJ.

Cook, V.J. and M. Newson (2007). *Chomsky's Universal Grammar: An Introduction*. Wiley. ISBN: 9781405111874. URL: https://books.google.de/books?id=mguunu3sI-YC.

Covington, Michael and Joe McFall (May 2010). "Cutting the Gordian knot: The moving-average type-token ratio (MATTR)". In: *Journal of Quantitative Linguistics* 17, pp. 94–100. DOI: 10.1080/09296171003643098.

Fetcey, S. and le Comité Linguistique Kotava (2013). *Kotava: grammaire officielle complète (version III.14)*. URL: ttp://www.kotava.org/fr/fr_pulviropa_000.pdf.

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

Gobbo, Federico (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 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.

Mairal, Ricardo and Juana Gil (Jan. 2006). *Linguistic Universals*. ISBN: 9780521837095. DOI: 10.1017/CBO9780511618215.

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*.

Shannon, C. E. (1949). *A Mathematical Theory of Communication*. Vol. 27, pp. 379–423.

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...

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.