University of Zurich[UZH]

English Seminar

Masterarbeit der Philosophischen Fakultät der Universität Zürich

Automatic Article Correction in Academic Texts

Contrasting Rule-Based and Machine Learning Approaches

Verfasserin: Sara S. Wick

Matrikel-Nr: 10-737-666

Referentin: Prof. Dr. Marianne Hundt

September 28, 2016

Acknowledgement

I owe my sincere gratitude to many people for helping me finish this Master's thesis. First, I would like to thank Prof. Dr. Marianne Hundt for letting me take on a crazy project and trusting in me that I am able to handle it. I am also forever grateful for Dr. Annette Rios' technical support and valuable input for the machine learning part. Moreover, I would like to thank Dr. Simon Clematide for taking time to speed up my endless cycle of evaluations. Thank you to Ally Chandler for proofreading and lending your language expertise.

I would also like to thank my family and friends for the many pick-me-ups, ears lent and cheesecakes. Lastly, thank you Laurent, for being who you are.

Contents

List of Figures

List of Tables

1 Introduction

Quirk et al. (1985) claim that "English is generally acknowledged to be the world's most important language" (3). While this claim can be challenged for a number of reasons, it is true that English has become the lingua franca in many respects. Therefore, it is no surprise that the majority of English-language writing is no longer produced by native speakers; some claim that non-native speakers author up to 74% of all English texts (Gamon et al., 2008, 1). While some non-native speakers can produce English on par with a native speaker, others struggle with various aspects of vocabulary or syntax, especially if basic structures differ from their mother tongue. Academic publication, for example, requires a certain competency in the target language and non-native speakers have been rejected due to their unclear expression of ideas. The ideas can be ingenious but if the language capabilities are not on a level which does justice to the complexity of the idea, there is little an editor can do (Bhaskar et al., 2011, 250). Apart from being able to convey one's point, correct language is important for a number of natural languages processing tasks. For example, simple tasks such as part-of-speech tagging or parsing texts depend on mostly correct language, to perform to their full potential. More complex tasks, like automatic summarization or machine translation, also depend on the input language to be as accurate as possible (Han et al., 2004, 1625). Commonly used text editors, such as MS Office Word, correct typographical errors and basic punctuation, however, it offers no inputs towards the correct use of preposition or articles (Heidorn, 2000). One of the more advanced grammar checkers available is *Grammarly*. It checks text using over 250 complex grammar rules and can detect errors in subject-verb agreement, modifier placement, article use and more. *Grammarly* can be used for free as a basic plug-in on various browsers and text editors to instantly review a text. For a more advanced grammar check, one needs to purchase the full-service version (Grammarly, Inc., 2016). While *Grammarly* is a comprehensive grammar correction tool, this thesis will concentrate on one issue, namely the correction of article usage. Furthermore, the text genre will be restricted to academic texts, as English is the language of science. As has been previously mentioned, grammatically incorrect texts hinder the understandability of the subject matter significantly, and improvements in the article usage increase coherence remarkably.

1.1 Motivation

The inspiration behind this thesis came from my work on the SPARCLING[1] project. I helped extract data from a large parallel corpus, where the German noun was accompanied by an article and the English counterpart was not. It was interesting to see how articles are used differently from language to language even though the context is every similar. Furthermore, it is interesting how familiarity with one language can be used to deduce information about the aligned second language. In addition to my work with articles in this research project, articles can be difficult to use correctly. This became apparent in my own writing, as well as fellow students' texts or research papers. It is particularly noticeable when the mother tongue of the writer does not have articles, or relies on a completely different system. Any linguist knows that the definite article *the* is the most frequently used word in the English language. Moreover, there are only three possibilities for articles definite, indefinite, or no article. Given this, perhaps their usage should be 'easy'. Nonetheless, there are hardly any grammar correction tools which successfully correct article usage.

Apart from article usage as a problem to be solved, I have always found it interesting how rule-based systems usually perform far worse than statistical methods in computational linguistics. One technique relies on a huge volume of human, linguistic expert knowledge, while the other is informed by a sheer mass of data. This juxtaposition of knowledge against immense data sets has fascinated me since the beginning of my Master's studies. Therefore, I wanted to engineer two systems, one rule-based and one based on machine learning techniques. Both systems, developed using similar time scales and effort, will be used to test the same task and their performance analyzed. Moreover, I have not had a chance to understand machine learning and its application in its entirety during my studies. Therefore, I wanted to devote some time to learning and understanding machine learning techniques better.

1.2 Aim of Project

The aim of this Master's thesis is, simply put, to engineer two systems which automatically correct the article use in the input text. The first system will attempt to correct the articles according to a specific set of linguistic rules. In this rule-based approach, there is linguistic knowledge involved; one needs to write a set of grammar rules which then is applied to text. The second system will take a different approach and will correct input texts using machine learning techniques. Here the computer

[1] large-scale parallel corpora to study linguistic variation

recreates linguistic knowledge it gained from studying practice texts. Apart from building two functioning systems, a further aim is to identify strengths and weaknesses for each system, in order to obtain insight for possible future combinations of the two systems. Moreover, the influence of the training material on the performance of the algorithms will be investigated, especially regarding differences in the time of publication.

1.3 Outline

The remainder of this thesis is organized as follows, a general introduction to English articles (2.1) is given, followed by a categorization of nouns (2.2) and several observations about the usage of articles (2.3). In Chapter 3 the focus lies on the resources used for the automatic article correction, namely the Part-of-Speech-Tagger (3.1), the Parser (3.2) and lastly, the machine learning workbench (3.3). In Chapter 4 the rule-based approach is described, after which the illustration of some previous works (4.1), the development of the correction rules (4.3), as well as their evaluation (4.4), will be explored. The second system is discussed in Chapter 5. As machine learning is a complicated topic, the concept of machine learning is introduced (5.1) as well as the algorithms which are used (5.2). The actual machine learning system is explained through pre-processing (5.5), feature extraction and selection (5.6), and model training (5.6). As with the rule-based approach, the system is evaluated (5.8) and conclude. Both systems are discussed in Chapter 6, to elicit comparative strengths and weaknesses. Chapter 7 concludes the findings of this project and gives an outlook into possible future research.

2 Articles in English

The English language is one of a few languages, which uses distinct words for the definite as well as the indefinite article. In addition, English uses an indefinite article which is different from the word for the numeral *one*. *The World Atlas of Language Structures Online* (WALS) lists only 55 languages that share the same structures as English in regards to article use. This is relatively small, as there are 2679 languages documented by WALS (Dryer and Haspelmath, 2013). In comparison, there are 198 languages that have neither indefinite nor definite articles. Some linguists claim that there are over 7000 active languages in the world (Lewis and Fenning, 2016). Consequently, it can be stated that English uses articles in a very rare manner. Given the diversity of article usage across languages Christophersen's claim that "to give a definition of the article such as would be valid for all languages is a difficult task" (Christophersen, 1939, 22) is very true. It will prove to be a challenge doing so for solely the English language.

In this chapter the articles for Standard English will be defined and contrasted to determiners, certain nouns will be categorized according to their article use, and the usage and omission of articles, and their related semantics, will be illustrated.

2.1 The Articles

Word classes have a long tradition in linguistics dating back to ancient Greece. While classes like `noun`, `verbs` and `prepositions` continue uncontested in present-day grammar, `articles` are said to "not deserve a place" in the word classes (Hudson, 2010, 251). The scholarly discussion around what articles are and whether they 'deserve' a category of their own is also mirrored in the definition of *article* given in the *Oxford English Dictionary*. According to the OED, an article is: "Each of the members of a small set of words (in English the, a, or an, traditionally regarded as adjectives, now also classed as determiners) that give definiteness or indefiniteness and specificness or genericness to the application of a noun." (Oxford English Dictionary, 2003b). The comment in the parenthesis echoes the debate around articles.

For example, words like *any*, *this*, or *which* share many similarities with *the* and *a*, however, they are traditionally considered to be adjectives. In more recent time, these words have been grouped together in the new category `determiners` (Hudson, 2010, 252). Articles match well with the definition of a determiner; therefore, some linguists argue that the category *article* has become redundant. Nevertheless, articles have a special role in syntax and are interesting to look at separately from other determiners listed above.

(2.1) I love this house.

(2.2) I love this.

(2.3) *I saw the ~~movie~~.

(2.4) *I saw a ~~movie~~.

In sentence (2.1) and (2.2) the determiner *this* is once used in an adjectival position and once in a pronominal position. Both sentences are grammatically correct. One can see that determiners do not need to be followed by a noun. Sentence (2.3) and (2.4), are grammatically incorrect because articles need to be followed by nouns. This ability to stand without an accompanying noun is one of the main differences between determiners and articles, and gives ample justification for differentiating the two word categories in this thesis (Hudson, 2010, 253).

Articles in English can have three forms, *the*, *a/an* and the zero-form, which will henceforth be represented as ∅. A difference in pronunciation of *the* as voiced ð or as voicelss θ is not of importance for this thesis as it is concerned with written texts, therefore it was not taken into consideration.

	singular	**plural**
definite	the cake	the cakes
indefinite	a cake	
∅	cake	cakes

Table 1: Five Forms Articles can take according to Christophersen (1939, 24)

In Table 1 the five possible noun forms in English are depicted. There is a division between singular and plural forms, as well as ∅, indefinite, and definite forms. While the matrix technically allows for six possibilities, there is no indefinite plural form. In writing, the definite article is far more frequent than both forms of the indefinite article combined. In academic writing, which has the highest use of articles in general, the definite article has a frequency per million words of around 55‘000, while the indefinite form *a* occurs around 15‘000 times in a million words, and the indefinite form *an* is used roughly 5‘000 times per million words. Conversely, in

conversation the frequency of definite and indefinite articles are more similar (Biber et al., 1999, 267).

An issue, which must be touched upon is that there is little agreement on how to label the 'absent' article, which so far has been referred to as the ∅-form. Berezowski (2009) calls the ∅-form "the oddball of the English article system" (1). Sweet (1898) discusses the omission and *absence* of articles. The term 'bare word' would have been the preferred terminology of Jespersen (1949, 403). For Christophersen (1939) the ∅-form is not an article unto itself, as it is merely a form of the noun when neither a definite nor an indefinite article is allowed (23-24). Berezowski claims that the use of the term zero-article is an easy solution to a very complex problem. It gives the illusion that one of the articles "is never given any overt phonological representation" and therefore there actually is an entirely uniform pattern "and an article truly forms a part of each and every nominal" (2009, 11). This thesis has no intention to claim that any of the abovementioned arguments are valid or not. The choice to use the term ∅-form was made simply because the majority of texts consulted referred to the concept as such, and that the form is represented as the numeral *0* in the scripts written for the project.

Before delving into the usage of the aforementioned articles, different noun categories must be introduced in order to gain some understanding of the various influences on article usage.

2.2 Categories of Nouns

Nouns can be organized into two main categories `Proper Nouns` and `Common Nouns`. Proper nouns are usually names of specific people, geographical places, days, or months and so forth (Quirk et al., 1985, 288). Biber et al. (1999) state that typical proper nouns "are arbitrary designations which have no lexical meaning" (245). They usually appear without articles and do not change in number. In most cases they are marked orthographically by the first letter being capitalized (Biber et al., 1999, 245). Common nouns are the class of nouns which refer to entities that are non-arbitrary. The line between proper nouns and common nouns is not as unambiguous as it might seem. Nevertheless, the definiteness of a noun is a good indicator, as proper nouns tend to be accompanied by the ∅-form, while common nouns use both definite and indefinite articles, as well as the ∅-form.

Common nouns can be further divided into `countable` and `uncountable` nouns.[1]

[1]Sometimes they are also referred to as count vs non-count or thing-nouns vs mass-nouns.

Countable nouns refer, as the name suggests, to items which can be counted, like *house*, *cow* or *smile*. These nouns have both a singular and plural form, and they can be used with definite and indefinite articles (Biber et al., 1999, 241). Uncountable nouns denote "an undifferentiated mass or continuum" (Quirk et al., 1985, 246) like *music*, *milk* or *money*. Because they do not vary in number, there is usually no distinction between singular and plural. Uncountable nouns do not take the indefinite article; however, they appear both with the definite article and the ∅-form. The difference between countable and uncountable nouns is far from clear cut.

(2.5) Mary: Can I bring you *a water*?

(2.6) John: No, I do not drink *water*.

In sentence (2.5) *water* is a countable noun, as Mary is referring to a glass or bottle of water she is offering John. In sentence (2.6) John is referring to water in general, e.g. the uncountable amount of drinkable water there is. The difference between countable and uncountable nouns, like in the example sentences, is often purely based on semantics and context, and therefore very hard to deal with for computers.

One last distinction can be made between `abstract` and `concrete` nouns. This division is made on a semantic level rather than a syntactic one (Quirk et al., 1985, 247). Concrete nouns are nouns which are "accessible to the senses, observable, measurable, etc", for example *cow* or *gold*. Abstract nouns on the other hand, are "nonobservable and nonmeasurable" (Quirk et al., 1985, 247), for example *hope* or *homework*. In Figure 1 all categories mentioned are represented with some examples.

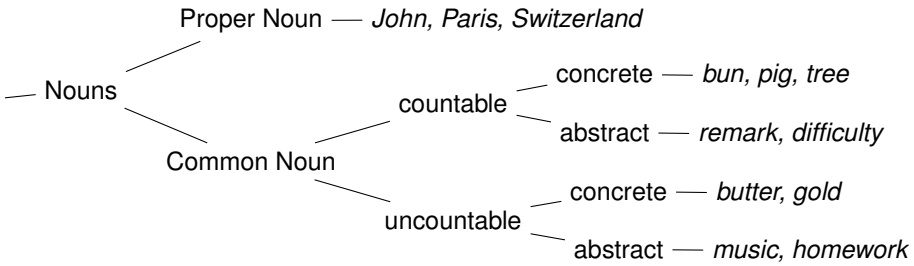

Figure 1: Most common Noun Categories (adapted from (Quirk et al., 1985, 247))

These different categories of nouns have implications for the article use. As Christophersen claimed, that there are five forms which nouns/articles can take. It is important to note, that not all nouns make use of the five categories.

In Table 2, further examples of the categories which have so far been described

	countable/uncountable	countable	uncountable	proper nouns	uniques
singular	cake		butter	Mary	
	a cake	a book			
	the cake	the book	the butter		the equator
plural	cakes	books			
	the cakes	the books			

Table 2: Five Forms Nouns and Articles can take (adapted from (Christophersen, 1939, 24))

can be found. Proper nouns rarely use an article. There are different categories of countable and uncountable nouns, some of which can be paired with all articles, some of which cannot. Moreover, there are the so-called 'uniques', which Christophersen places into a separate category. Uniques are nouns which usually only contain one member. These uniques do not require any context or situational explanation, as there is only one of them, and all communicators are aware of which one one is speaking. Thus, they occur with the definite article (Christophersen, 1939, 24-31).

2.3 Use and Omission of Articles

So far the articles have been introduced and different noun categories have been proposed. However, the actual usage of articles has not yet been discussed. In the following two sections, the basics of article usage and omission are presented. However, there are very few universal rules which are binding in every context.

2.3.1 Use of Definite and Indefinite Articles

The **definite article** *the* signals that the noun or noun phrase is "referring to something which can be identified uniquely in the contextual or general knowledge shared by speaker and hearer" (Quirk et al., 1985, 265). The concept or thing has already been introduced in the conversation, and therefore there is no ambiguity concerning what is being referred to. Christophersen (1939) calls this "a proper basis of understanding" between the speaker and the hearer. This can be achieved in several different ways, for example, through *situational reference*. Situational reference is given when the speaker and hearer are talking about *the table*, when

there is only one table in their immediate surroundings or if one asks "Have you visited *the museum*?" while standing in front of it (cf Quirk et al. (1985), 266; Christophersen (1939), 30; Biber et al. (1999), 264). One can also refer to this as *general knowledge*, if the speaker and hearer have a common ground, such as the same cultural background or they live in the same country. Then both will know which government is referred to by "the government". The *uniques*, which have been introduced in section (2.2) also fall into the category of general knowledge. Usually, it is very clear which sun is meant when one talks about *the sun*. Most often, however, the definite article is used to create an *anaphoric reference*. This is the case when the necessary information is provided earlier in the discourse.

(2.7) My friend bought *a house* last month.

(2.8) Now he needs new furniture for *the house*.

The concept of *a house* was introduced in (2.7), and therefore, the hearer knows which house is meant in (2.8) and no clarification is needed. The concept of *house* "can be treated as 'contextually known'" and consequently it is referred to by the definite article thenceforward (cf Quirk et al. (1985), 265; Christophersen (1939), 31). Lastly, there is *cataphoric reference*. Unlike the previous cases, the context follows the noun rather than precedes it. For example in the phrase "*the tree*, which was cut down", the tree is put into context by the relative clause that follows it (cf Quirk et al. (1985), 268; Biber et al. (1999), 264). All four types of reference specify that the referent of the noun is presumably known to all participants of the discourse, discourse, otherwise the speaker risks misunderstandings between communicants (Biber et al., 1999, 263).

The **indefinite article** *a* or *an* is historically derived from the numeral *one* and therefore often narrows the reference of the noun it accompanies. The other usage of the indefinite article is the introduction of a specific entity, which subsequently is then referred to by the definite article, as was just seen in (2.7) and (2.8). Quirk et al. (1985) argues that in some cases *one* can be substituted "as a slightly emphatic equivalent of *a*" (273). A somewhat different view is put forward by Christophersen (1939). He claims that there are three distinct uses of the indefinite article: (i) introductory use, where the focus is on one particular thing out of many, (ii) characterization, where the focus is on the "generic characters of a single individual" and (iii) the generalization, where the referent is several things of the same class (33). In (ii) the term *generic* is used, which is the opposite to the term *specific*. All of the examples used in this thesis thus far have been specific examples.

(2.9) An apple and two bananas are left.

(2.10) Bananas are my favorite fruit.

In sentence (2.9) we talk about specific specimens of bananas, namely, the ones that were left over after making a fruit salad. In sentence (2.10), on the other hand, not one particular banana is meant but rather the group of fruit called 'banana' is referenced. While number and definiteness are important for the specific use of articles, it is not as crucial for the generic use, as "generic reference is used to denote the class or species generally" and not one or more distinct one(s) (Quirk et al., 1985, 265). Therefore, it is not *the bananas* in (2.10).

To conclude, in Standard English one can use definite or indefinite articles, both generically or specifically for singular common nouns. Moreover, in plural common nouns there are no indefinite articles; however, one can still use the plural definite article as both a generic or specific marker. One needs to keep in mind that usually, uncountable nouns do not change in number, consequently, they rarely use plural articles.

2.3.2 Omission of Articles

It has already been mentioned, that the omission of articles is a hotly debated topic. In the following, an account on where articles can or should be omitted is given. Sweet (1898) states that "the absence of articles is in most cases a tradition of time ... when there were no articles at all" (64-65). Therefore, certain rules still govern it. Other scholars, however, claim that the ∅-form is assumed to be simultaneously indefinite and definite, which makes it nearly impossible to theorize on without getting lost in contradictions (Berezowski, 2009, 2). More generally speaking, the ∅-form can be used with proper nouns, singular, uncountable nouns, and plural countable nouns. These three cases will be elaborated on in the remainder of this section.

Sweet (1898) states that proper names do not take an article, like *John* or *Mary*, which we have seen in Table 2 (63). Furthermore, proper nouns of institutions often appear with the ∅-form, although almost always in combination with a prepositional phrase.

(2.11) They got married in ∅ church.

(2.12) *The* church was charming.

The same goes for meals, means of transportation, or times of the day. As one can see in (2.11) and (2.12), the same noun can appear with and without the ∅-form.

In the prepositional phrase, the articles are not needed, while in (2.12) a specific, already introduced church is referred to; therefore the article is obligatory. (cf Biber et al. (1999), 261-263; Berezowski (2009), 19-20). Moreover, the ∅-form is used in vocative phrases such as *That is alright, mate!*. Here countable nouns are used as forms of address, and consequently, do not need an article (Biber et al., 1999, 263).

Several syntactic constructions make articles redundant. For example, nouns which are part of a genitive construction as in *Peter's ∅ house*, appear with the ∅-form. Sweet argues that in these cases the nouns are already defined by the preceding genitive form (Sweet, 1898, 64). In parallel structures, such as *arm in arm* or *from country to country*, articles are not allowed in front of either noun (Quirk et al., 1985, 280). Furthermore, as was already mentioned in connection to proper nouns, nouns which are part of a prepositional phrase most often do not take an article. Articles also are omitted with temporal expression, such as *we met at ∅ noon*, means of transportation, *he traveled by ∅ plane* and institutions as in example (2.11) (Berezowski, 2009, 20).

For plural nouns and uncountable (singular) nouns there are no articles when the phrase refers to an "indefinite number or amount (often equivalent to *some*)" (Biber et al., 1999, 261, original emphasis). Quirk et al. (1985) mentioned, however, that *some* would not be an acceptable alternative to the ∅-form in all cases. For example, sentence (2.13) changes its meaning if *some* is inserted in front of *ducks*. John would then love to chase one breed of ducks, but not all kinds of ducks. In (2.14), the change in meaning is not as drastic as in (2.13), therefore, *some* could be considered as a valid alternative to the ∅-form.

(2.13) John loves chasing ducks.

(2.14) We had wine with dinner.

(2.15) He loves music.

The last example shows the use of the ∅-form with an uncountable noun. As was seen before, the line between countable and uncountable is very dependent on the context, and consequently so is the omission of the definite/indefinite articles or the use of the ∅-form.

As was shown in this chapter, English has three kinds of articles: the definite article *the*, the indefinite articles *a* and *an*, and the absence of an article, the ∅-form. It has also been shown that article usage is dependent on various factors, including the larger context of the word, the noun the article accompanies or its position in the phrase. In the next chapter, the resources used to engineer the two article correction

tools will be introduced. Afterwards, the development and performance of the two article correction systems will be presented.

3 Resources

In this chapter, the resources used for this study will be outlined. The resources must be explained as their quality holds a significant influence over the quality of the result from both article correction systems later on. The detection of articles and nouns is done automatically, using the linguistic information which the Part-of-Speech Tagger and the Parser provide. Apart from the *TreeTagger* and the *MaltParser*, the *Lightside* workbench will also be presented. *Lightside* was used as a workbench for the machine learning section of the thesis.

3.1 Part of Speech Tagger

The first step in many of linguistic annotation pipelines is Part-of-Speech (POS) tagging. A POS-Tagger automatically assigns each token a POS-tag, some taggers additionally assign lemmas and morphological information (Voutilainen, 2003, 2). Well-known examples for POS-tags are *noun*, *verb* or *adjective*. The number of labels which are assigned depends on the size of the respective tag set. Apart from the size of the tag set, the methods used in the assignment process differ as well. Below is an explanation of the *TreeTagger*. In Table 3 the top row consists of the sentence to be tagged. The middle row shows the POS-tags, and in the bottom row, one can find the lemmas to each token.

This	is	a	sample	sentence	.
DT	VBZ	DT	NN	NN	SENT
this	be	a	sample	sentence	.

Table 3: Sample Output from the Treetagger

The *TreeTagger* was developed by Helmut Schmid in 1995. Schmid's aim was to circumvent the sparse data problem using decision trees (Schmid, 1995, 1). First, each token is assigned a probability for all possible Part-of-Speech tags. These probabilities have been learned from the Penn-Treebank, which consists of over 4.5

million words of American English, mostly garnered from newspapers (Marcus et al., 1993, 1). Once each word has a probability, a decision tree is built recursively. The decision tree is then used to make a choice between the different possibilities, given the two preceding POS-tags. For example, given the two proceeding POS-tags are **DT** and **JJ**, the POS-tag for *store*, is more likely to be **NN**, a noun, than **VB** the verb base form, as in *a small store*. The small context which is needed to disambiguate the tokens enables the TreeTagger to avoid the sparse data problem, which other POS-Taggers face (e.g. Cutting et al. (1992), Kempe (1993)). Using the same principle, but expanding to tetragrams from trigrams, the TreeTagger performs with an accuracy of 96.36 % (Schmid, 1995, 16).

POS-Tag	**Description**	**Example**
NN	noun, singular or mass	*tree, house*
NNS	noun, plural	*trees, houses*
NNP	proper noun, singular	*Switzerland, Kreisler*
NNPS	proper noun, plural	*Americans, Volvos*
DT	determiner	*the, a, these, that, some*

Table 4: Noun and Determiner Tags for the TreeTagger

The tag set of the *TreeTagger* consists of 36 different tags. However, only five of them are pertinent to this project, namely all noun tags and the determiner tag. In Table 4, the relevant tags are listed with an explanation as well as an example. The noun tags are straight forward. Singular and plural forms are distinguished as well as common nouns and proper names. This results in four noun tags, which are quite distinguishable. The noun tags, therefore, do not pose a significant challenge for algorithms. The determiner category is fuzzier as the given examples illustrate. In Chapter 2 it was clearly outlined what the differences between articles and determiners are. Further, it was explained why only the 'classic' articles are considered for this project. Nevertheless, the fact that the individual articles are not isolated with a special POS-tag has a couple of implications. Firstly, all the unwanted determiners (i.e. *this*, *some* , ...) needed to be filtered out for the correction process. Secondly, the determiners had to be considered as *correct* during the correction process, even if the program would consider them as *wrong* because they are not articles. These issues will be addressed in greater details later in Chapter 4 and 5, respectively.

3.2 Parser

A language parser assigns a syntactic analysis to a string of tokens based on a given grammar (Mitkov and Carroll, 2012, 1). There are many different parsing methods, as well as the depth of analysis available. There are rule-based or statistical parsers and shallow versus deep methods of parsing. For this project the *MaltParser* for English (Nivre and Scholz, 2004) was used. The *MaltParser* is a dependency parser. It has been developed for English by Joakim Nivre and Mario Scholz, based on an algorithm originally developed for Swedish (Nivre, 2003). Dependency parsers assign dependencies between the headword and its dependee(s). Each link between tokens is labeled with a grammatical function of the dependee(s) with respect to the headword of the phrase (Mitkov and Carroll, 2012, 3).

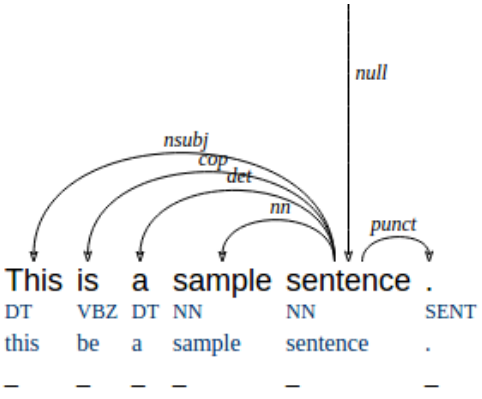

Figure 2: Visualization of a parsed sample sentence

The relations are taken from the Penn Treebank; however, the Penn Treebank does not use dependency labels as it is parsed on a constituent basis. Nivre and Scholz converted the constituents to dependency labels using the Penn TreeBank II Annotation Scheme by Bies et al. (1995). As can be seen in Figure 2 in the noun phrase *a sample sentence*, *sentence* is the head noun, *sample* is dependent via a noun-relation, and *a* is the corresponding article to the compound *sample sentence*[1]. The parser reaches an overall accuracy of 86%; while this is not the highest possible accuracy for parsing English texts, given this project's focus on noun phrases it is sufficient (Nivre and Scholz, 2004, 5). The actual algorithm is similar to the parser engineered by Yamada and Matsumoto (2003); it uses a "deterministic parsing algorithm in combination with a classifier induced from a treebank" as well (Nivre and Scholz, 2004, 1). While Yamada and Matsumoto (2003) only use a bottom-up approach, the combination of a simultaneous top-down and bottom-up approach

[1]It also becomes clear that the parser is not perfect, as the rest of the parse is not correct.

allows the algorithm for the *Maltparser* to be very efficient, as the running time grows linearly with the size of the input (Nivre and Scholz, 2004, 1).

The most important dependency label for this project is the **det** dependency. This denotes the dependency between a head noun and its determiner. As mentioned before, this dependency does include more than just the classic articles. Also of interest are the **nn** and various **mod** dependencies. **nn** is the noun compound modifier, referring to any noun which serves to modify the head noun. There are several different other versions of modifying elements; their labels have been collected in Figure 5.

Dependency Label	Description	Example
amod	adjectival modifier of the head noun	John likes **yellow** *houses*.
advmod	adverbial modifier of a word	**less** *often*
cop	copula	Bill *is* **big**.
nn	noun compound modifier	The **oil** *prices* have plummeted.
num	numeric modifier of a noun	Mary has **three** *childern*.
poss	possession modifier of a noun	**their** *offices*
nsubj	nominal subject to a noun	The *baby* is **cute**.

Table 5: Selection of Modifying Dependency Labels

The dependee is represented in boldface while the modified token is represented in italics. These seven modifiers appear in the data[2] as modifiers to the head noun of the phrase. **cop** and **poss** seem to make little sense, and are most probably the result of parsing errors. These labels have nevertheless been included in Table 5 in order to provide the most comprehensive and accurate account of the data.

3.3 Lightside

The third and last resource which was used is the Machine Learning Researcher's Workbench *LightSide* (Mayfield and Rosé, 2013). This workbench is a tool specifically developed to apply machine learning techniques to text based on WEKA. It is therefore equipped with certain features specifically geared towards natural language processing, unlike the Waikato Environment for Knowledge Analysis (WEKA). WEKA is aimed at academics and professionals as a "comprehensive collection of machine learning algorithms and data preprocessing tools" (Hall et al., 2009, 1). It is an

[2]The data is illustrated in detail in subsequent sections.

all-purpose workbench, which can be used in many different fields of resource or business. Through *LightSide*'s interface, the user can choose different texts to work with, then extract features, train algorithms on the texts and finally evaluate the output. Each of these steps will be elaborated on below with respect to *Lightside*. The concept of machine learning will be explained in greater detail in Chapter 5.

Extract Features Features are the information which the algorithm will later base its learning on. *Lightside* provides some ready to use features for the user in the *Basic Feature*-column. One can choose to extract unigrams, bigrams, POS-bigram, or stemmed N-grams. All these features are automatically extracted by *Lightside*, and it is not entirely clear how it is accomplished. Additionally, the user can provide further features, in the provided csv text file[3], which is loaded into the workbench.

Restructure Features Once all the features have been extracted, *Lightside* provides the user with the opportunity to exclude single instances of features from the training process. This can be very helpful once some initial results have been obtained, to eliminate 'confusing' features.

Train Models In this step, the actual algorithm is trained on the data. One can choose between different algorithms and evaluation methods. After the training process has been completed a confusion matrix is provided along with the accuracy and the respective kappa value of said model.

Evaluate Results A very helpful feature in *LightSide* is the option to explore the results for each trained model. One can evaluate the influence of individual features on the results. Moreover, one can determine 'confusing' features, which produce a lot of false positives and/or false negatives, and then remove them while restructuring the features.

The aforementioned importance of the POS-Tagger and Tagger applies in particular to the features which are extracted by the researcher and then fed into *Lightside* via the csv-file. One can pass on information about the relation of tokens, as well as characteristics about individual words, such as POS-tags. Furthermore, since instance extraction for the algorithms are based on the noun tags, if a token is mistagged as a noun it will produce noisy data. Likewise, if a noun is not tagged as a noun it will not appear in the data.

[3]comma separated values

4 Rule-Based Approach

In the beginning of automatic language processing it was believed that "Human language can fundamentally be explained through the interaction of different generally applicable rules, and these rules can be explicitly formulated" (Foth, 2007, 5). With the rise of programmable computers, the theory was extended to being able to feed these rules to a computer and the computer will be able to understand human language. However, these high expectations could not be met. The attempt to do rule-based machine translation, for example, ran into a rather unexpected problem. It was not that the right translation could not be produced, but rather that too many wrong possibilities were produced as well. Therefore, a human had to ultimately decide which of the proposed sentence was the right one (Foth, 2007, 5-6). This, in turn, did not result in the desired elimination or reduction of manual work in the translation process.

In practice, rule-based approaches mean that a human formulates rules according to which a process is carried out. In the case of machine translation, each token gets assigned one or more translation possibilities. One of these possibilities is then chosen, for example, on the basis of the preceding token's POS-tag. If a token or syntactic construction is not covered by the rules, a predefined default will be implemented. If no default exists, the system breaks down. As such, rule-based systems are not well equipped to analyze unexpected data.

In this section some examples of rule-based systems will be introduced, then necessary pre-processing will be explained. An explanation of the development of the rule-based system to correct articles will follow, and lastly the final system will be evaluated before drawing some preliminary conclusions.

4.1 Automatic Language Correction using Rules

This section will explore two different kinds of previous works. Firstly, two studies which formulated theoretical rules about article usage in English will be introduced, and secondly an overview will be provided of papers which applied rule-based sys-

tems to correct language use.

Yotsukura (1970) aims to "compile a practical guide of formulae showing where to use (and not to use) appropriate articles" (9). She selected the 103 most frequent nouns from nine text books used at local American high schools. These nouns occurred a total of 8936 times in her corpus. She then extracted seven different types of noun phrases which covered all instances in the corpus, for example, *the + Ns* as in *the cats* (Yotsukura, 1970, 45-49). Yotsukara considers *the*, *a/an*, ∅ and *some* as articles for her study, further she categorized her nouns into *countable*, *uncountable* as well as *concrete* and *abstract* nouns (Yotsukura, 1970, 54). After having rigorously categorized all nouns, she proceeded with formulating rules for each category. Each noun has three dimensions: countable vs. uncountable, concrete vs. abstract and definite vs. indefinite, consequently, the rules are formulated using these dimensions. This results in rules like (4.1) (Yotsukura, 1970, 78).

(4.1) $if \quad DN_{1a} \quad then, \quad D = the/a \quad the\ group,\ a\ group$

If the noun is of the category *1a*[1], then it will take either a definite or an indefinite article. The study produces 38 formulae. 17 of the formulae leave only one option, making it very clear which article should be used. However, the remainder indicate up to four possibilities that may be correct. It was beyond the scope of Yotsukara's study to give definite suggestions when there is more than one correct option, as "either divided usage or contextual elements" need to be taken into consideration (Yotsukura, 1970, 106). Yotsukura (1970) claims that one could move on to an unlimited corpus with the same methods she has illustrated in her study. However, this would be immensely time-consuming as there is an enormous amount of manual labor involved, and such a task seems impractical.

In the second study, Kałuża (1981) defines "a few very simple rules governing the whole usage [of articles]" (7). His rules are divided into *specifying uses* and *generalizing uses*, as well as *personal proper names* and *non-personal proper names*. A further important distinction is made between countable and uncountable nouns. Kałuża (1981) then lists many very specific rules on how to correctly use articles.

(4.2) When we have in mind a specific entity of a class paraphrasable by "a certain" or "a particular" not yet expressed or implied, we commonly i n t r o d u c e it by means of *a* (Kałuża, 1981, 23, original emphasis)

The rule cited in (4.2) is one of five rules governing specifying uses of countable nouns with the indefinite article. In contrast to Yotsukura (1970), Kałuża uses prose to

[1]This category includes singular countable nouns.

formulate rules, therefore making it very hard to translate into computer-readable rules. A simple algorithm cannot accurately judge whether one can interchangeably use *a boy* or *a certain boy*. Kałuża concludes that one needs to take into consideration three dimensions of nouns, to determine the correct article, namely, *phrasal vs non-phrasal*, *countable vs. uncountable* and lastly *specific vs. generic* (1981, 83). He further states that if one is not naturally able to categorize nouns, one should resort to phraseological dictionaries. This seems counter intuitive to his previous claim that his rules were simple.

The two studies seen so far do not actually implement their rules. In light of this, some papers which correct language using a rule-based system will be presented. Most studies, take a hybrid approach towards automated language correction, meaning that they use both rule-based and statistical methods to correct text. Bhaskar et al. (2011) use conventional grammar tools and spell checkers to detect an array of errors, for example, "wrong form of determiner", "verb agreement error" or "missing preposition" (251). After singling out mistakes, they use a statistical tool to determine the correct version. In a final step, they merge the rule-based error detection with the correct version provided by the statistical model, which leads to a corrected document (Bhaskar et al., 2011, 252). While their system dealt well with some errors, it had difficulties with many syntactic and semantic errors, particularly the indefinite article.

A different hybrid approach to grammar correction was taken by Kunchukuttan et al. (2013). They focus on three main errors, *noun-number*, *determiner* and *subject-verb agreement*. While they solve the first two with classifiers, *verb-subject agreement* is approached using a rule-based system (Kunchukuttan et al., 2013, 82). The system has two stages; firstly, the subject of a given verb is identified. Secondly, conditional rules are used to correct the given verb if deemed incorrect. The conditional rules obtain linguistic information through POS-tags and lemmas. This set up was quite successful in correcting the subject-verb agreement. However, the lackluster performance of the correction tool towards the other two error types significantly lowered the performance of the approach overall. Errors in noun-number, for example, will have consequences for subject-verb agreement. Therefore, the authors report an F-1 measure for the subject-verb agreement of 28.45 for the complete system. However, if they account for the errors made due to lack of correction for noun-number errors, the F-1 measure increases drastically to 66.12 (Kunchukuttan et al., 2013, 84-86). This demonstrates that often the correction of one error type is dependent on the successful correction of another error type.

Finally, a different approach to rule-based language correction is demonstrated by analyzing Behera and Bhattacharyya (2013). They "consider grammar correction as a translation problem - translation from an incorrect sentence to a correct sentence" (Behera and Bhattacharyya, 2013, 937). The system learns synchronous context-free grammar rules from a parallel corpus with aligned correct and incorrect sentences. These rules are then used to form syntax trees for the wrong sentences, matching to a correct syntax tree and 'translating' the wrong syntax tree into the correct syntax tree. With this approach Behera and Bhattacharyya (2013) are able to correct article choice, preposition, unknown verb, word insertion as well as reordering errors. Because they approach the correction as a translation issue, they measure the improvement in the BLEU score; the baseline had a BLEU score of 0.7551 and with a training set of 3000 sentences they were able to increase the score to 0.7744 (Behera and Bhattacharyya, 2013, 940).

These different approaches to automatic language correction using rules show the variation of approaches as well as some of the main difficulties. In subsequent sections, all necessary steps in developing the rule-based article correction system for this project will be introduced and elaborated on.

4.2 Pre-Processing

Pre-processing the data is rather straightforward. The raw text is tagged using the standard *TreeTagger* for English application. Then the tagger output has to be transformed into a conll-format. This is done because the *Maltparser* needs a conll-file as input. Once the text is in the right format, it can be fed to the parser, which then parses the text based on a pre-trained model. This model has been trained on newspaper texts and is freely available on the *Maltparser* website. A parsed sample sentence can be seen in Table 6; it is taken from the academic section of the Brown Corpus.

The number in the seventh column indicates on which other token the current token is dependent, and column eight specifies this relation. For example, the first token *radio* is dependent on token number two *observations*. As *radio observations* is a compound noun, the dependency between the two is labeled as *nn*, which stands for a noun modifier relation as explained in Table 5.

The pre-processing of the data is only needed once the rules are applied to text. The rule-based system is solely based on linguistic knowledge and therefore is theoretically developed independently from the text.

1	Radio	radio	NN	NN	_	2	nn	_	_
2	observations	observation	NNS	NNS	_	0	null	_	_
3	of	of	IN	IN	_	2	prep	_	_
4	Venus	Venus	NP	NP	_	3	pobj	_	_
5	and	and	CC	CC	_	2	cc	_	_
6	Jupiter	Jupiter	NP	NP	_	9	nsubj	_	_
7	have	have	VBP	VBP	_	9	aux	_	_
8	already	already	RB	RB	_	9	advmod	_	_
9	supplied	supply	VBN	VBN	_	2	conj	_	_
10	unexpected	unexpected	JJ	JJ	_	12	amod	_	_
11	experimental	experimental	JJ	JJ	_	12	amod	_	_
12	data	datum	NNS	NNS	_	9	dobj	_	_
13	on	on	IN	IN	_	9	prep	_	_
14	the	the	DT	DT	_	16	det	_	_
15	physical	physical	JJ	JJ	_	16	amod	_	_
16	conditions	condition	NNS	NNS	_	13	pobj	_	_
17	of	of	IN	IN	_	16	prep	_	_
18	these	these	DT	DT	_	19	det	_	_
19	planets	planet	NNS	NNS	_	17	pobj	_	_
20	.	.	SENT	SENT	_	9	punct	_	_
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Table 6: Sample Output Sentence in the .conll-format

4.3 Development of Rules

As explained earlier, it is difficult to describe language using only a set of rules while simultaneously accounting for all exceptions and eventualities. To facilitate this process, rules were developed in stages, starting with the most simple and moving towards more complex rules. There have been a few attempts at formulating rules for article use, especially for language learners (cf Berry (2013), Murphy (2004), Siepmann (2008)). All of these were written with a human being in mind and not meant to be implemented as computer readable grammar rules.

It is important to remember that the performance of all these rules is dependent on the quality of pre-processing. For this thesis, correct POS-Tagging was paramount; the correct token must be assigned with the tag **DT** or one of the noun tags. Furthermore, the parser needs to correctly assign the relations, for the rules to analyze the right tokens. If the parser produces many mistakes, then the analysis will be

flawed as well. Thus, the rules are only as good as the pre-processing of the data they analyze.

4.3.1 First Set of Rules

The first set of rules that is implemented is illustrated in Figure 3. The rules are a combination of Yotsukura (1970) and Kałuża's (1981) writing. All tokens which have been tagged as a noun, are filtered according to certain criteria. First, all proper nouns are filtered out, and then each remaining noun is checked against the list of uncountable nouns. Following this second separation, the algorithm checks the POS-tags to see whether the countable nouns are plural or singular and makes the respective article suggestions. For uncountable nouns, there is a further step involved as the algorithm checks whether the token is modified and/or followed by *of*, as in *the quality of translation*. With this set of rules all nouns are processed. It should be noted, however, that only the 'core' articles *the*, *a*, *an* and ∅ are considered as correct options. This, of course, does not reflect reality. Determiners such as *some*, *that*, or *any* are valuable alternatives depending on the linguistic situations. Therefore, all of these determiners[2] were considered to be correct. These rules further oversimplify as no difference is made between definite or indefinite article use. This results in a lot of correct articles, where the system suggests two options, even though for a human the immediate context might suggest one of them to be more eloquent.

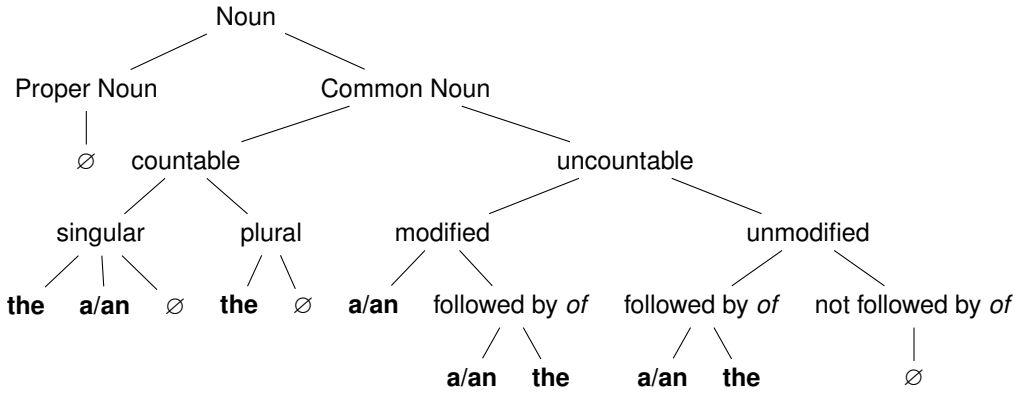

Figure 3: First Set of Rules

In other words, the rules of correction are very lenient; all possibilities are considered correct, even if one option is preferred over another. Furthermore, all nouns are treated the same, regardless of whether they are the head of the noun phrase or not.

[2]Some, this, that, any, those, these

This, by all means, does not adequately describe the rules of article use. Therefore, this was incorporated into the next set of rules.

4.3.2 Second Set of Rules

As an initial step, the second set of rules eliminates modifying nouns from the correction process. This step is inserted into the decision tree, before checking if the target noun is a proper or common noun as illustrated in Figure 4.

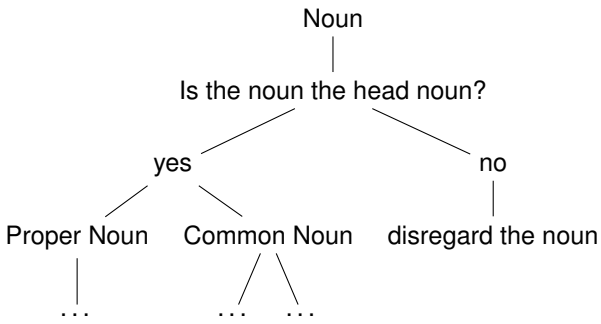

Figure 4: Second Set of Rules

With this additional step, it is possible to ensure that the modifying noun does not interfere with the head noun's article correction. As suggested in Figure 4 the correction process continues onward as described in Figure 3.

(4.3) *Radio* **observations** of Venus and Jupiter

(4.4) the measured *antenna temperature* **change**

The examples (4.3) and (4.4) are taken from the Brown academic subcorpus[3]. The noun(s) in italics are modifying nouns dependent on the nouns in bold via a noun compound modifier-dependency. As such, they are no longer checked for their article use.

The choice between definite and indefinite articles is usually made with semantic information. Has the concept been introduced before or is a class in general referred to and not a single entity? These questions are difficult for the computer to answer; therefore an attempt at providing this information was made. For each noun, it was checked whether the noun has appeared in the previous five sentences. The noun was lowercased; however, no real co-reference resolution was made. Consequently, if a noun is later on referred to by a pronoun or a different name it was not considered

[3]The Corpus will be introduced in more detail in Section 5.4

as 'already seen'. This check was implemented for all noun categories and all special cases (for example followed by *of*). There are several difficulties which were not anticipated, for instance, uniques like *world* or *universe*, (almost) always take the definite article. On the other hand, in examples such as

(4.5) the moon and planets

(4.6) the radio emission of a planet

the article usage for *planets* and *planet* were both labeled as wrong. In (4.5) *planets* has not been seen before, therefore the rules called for an indefinite article, which makes little sense as it is plural, plus it is within the scope of the definite article assigned to *moon*. In (4.6) the word *planet* was utilized in the preceding sentence, thus the rules suggested the definite article *the*, as the concept was already introduced. As a proficient speaker of English, it is quite clear that neither rule makes sense. In an effort to improve the performance of this particular rule the window was increased to 10 sentences and even 15 sentences, and it was ensured that plurals were not suggested to use indefinite articles. Nonetheless, the rule was ultimately removed altogether, as it did not produce enough accurate suggestions.

For the third round of rules, the first additional step was kept, but the attempt at adding semantic information was deleted.

4.3.3 Third Set of Rules

The main difficulty so far was that the rules were too lenient, meaning that there were too many cases where more than one option was correct. Yotsukura (1970) formulated 17 rules that lead to a single result. In the third set of rules, these 17 rules are implemented as far as possible and integrated into the basic structure from Figure 3. In Table 7 all twelve new rules can be found, although only eight resulted directly from Yotsukura (1970). The remaining four seemed sensible after analyzing some results from previous sets of rules.

The first rule should be read as 'if the token immediately preceding the noun has the POS-tag `POS`, then this noun is accompanied by the ∅-form'. If items are divided by a slash as in line two, this signifies that any of these options are possible. Thus, another valid example for the second rule could be *by car and plane* as in "They traveled by car and plane". The majority of rules lead to the ∅-form, as it is the least ambiguous. Many of the rules have different variables; this is done in order to keep the rules as modular as possible. Only the last two rules are highly specific, meaning they are hard coded for one particular expression and not for a construction

Noun Category	Rule	Example
all Nouns	POS noun → ∅	one's ∅ hands
	CC/IN/TO DT noun CC/IN/TO DT noun → ∅	from country to country
Countable Plurals	half DT noun → *the*	half the time
	DT certain/such noun → ∅	∅ certain horses
	DT same modifier noun as → *the*	the same color as
	both/(n)either/many/one/several/all/most/same of DT noun → *the*	most of the people
Singular Countable	such DT noun → *a*	such a house
	DT certain noun → *a*	a certain tree
	DT adj-est noun → *the*	the slowest car
	DT noun such as → *a*	a house such as
	in DT order to → ∅	in order to
	to DT date → ∅	to date

Table 7: Constructions added in the Third Set of Rules

where different tokens can take a certain position. Depending on the data set, up to 25% of all article corrections are narrowed to one option with these additional rules. Moreover, there is no case left where all three article options are considered correct. Obviously, the number of instances where there is only one option, depends immensely on the text. If the author of the text does not use constructions that lead to singular outcomes, then the rules will have to suggest two options. Nonetheless, the rules added in this set of rules resulted in corrections which were good enough for the scope of this project. Furthermore, the effort necessary to further improve the rules would have been beyond the scope of this project, as a second correction system is developed as well.

In conclusion, the development of the rule-based article correction system was done in three steps, moving from very general to more specific rules by implementing additional rules from publications like Yotsukura (1970). Several other rules emerged during the process of rule writing. This was the case in fixed expressions like *in order to*, which were not accurately judged by the rules in different practice texts. Even though some rules are inspired by real-life texts, it is important to underline again, that the rules were not deduced from training texts, unlike in the machine learning

approach outlined in Chapter 5. To conclude the development of rules, many more rules could be formulated with sufficient time; however, deeper investigations into rule writing would have been beyond the time scope of this project. Nevertheless, the evaluation will show that a strong foundation has been achieved while keeping the rules as modular as possible.

4.4 Evaluation

Before presenting the results from the evaluation, the texts used to evaluate both systems are presented. The actual results are then illustrated with examples to pinpoint difficulties, as well as ways to further improve this approach.

4.4.1 Texts for Evaluation

For the evaluation of both systems, the same texts were used in order to be able to compare the performance of the systems directly. While the text type and time of publication are not as important for the rule-based approach, it will become clear in Chapter 5 that it is vital for the machine learning approach. Therefore, the evaluation texts will be described here in more detail than currently needed, as it is critical to a proper understanding of the second evaluation.

To mimic the training data used in the machine learning system, two sets of texts were compiled for the evaluation. For the first set 5 texts snippets published in the year 1961 were extracted from the Corpus of Historical American English (COHA) (Davis, 2010). There is no *academic* or *science* genre, therefore, the *non-fiction* category was chosen, which contains mostly scientific publications. In order to find 'random' text passages the corpus was queried for *and it is*, which is a common, non genre specific trigram. From the search results, five text snippets from different scientific fields were selected. The same procedure was done for the second data set with academic texts published in 2006. For this set the data was collected from the *science* genre in the Corpus of Contemporary English (COCA) Davis (2008). This resulted in two small data sets; an overview is given in Table 8.[4]

As was done during the development of the systems, the native speaker texts are assumed to be correct. Thereby, the articles of the texts needed to be altered, to give the systems correctable input. Thus, all nouns were randomly assigned new articles, keeping in mind the ratio of indefinite, definite and ∅ articles during the

[4]All text snippets can be found in the appendix.

	number of words	number of nouns
1961	674	154
2006	731	169

Table 8: Overview of the Evaluation Texts

given time period. To ensure that the texts remained authentic, determiners such as *that*, *this*, and *some* were left untouched. These falsified texts were then used to evaluate both article correction systems which were developed for this thesis.

4.4.2 Results of Evaluation

The two jumbled data sets are both re-tagged and re-parsed before running them through the rule-based correction script. This leads to minor differences in the numbers of nouns as not all tokens were tagged the same way, once the articles were jumbled. Shuffling articles results in the creation of different trigrams, which can in turn influence the tagging process as described in section 3.1. In Table 9, one can see how the system performed on the two data sets. The number of corrected nouns is smaller than the total number of nouns, as the correction algorithm does not deal with determiners. Therefore, all nouns which were preceded by *this* or *that* and so forth are not listed in this table. An interesting case which resulted from this procedure is listed in (4.7).

(4.7) that the conditions – *original*

that conditions – *corrected*

In the jumbled version the definite article was deleted in front of *conditions*, which lead the tagger and parser to conclude that the determiner *that* must be linked to the noun *conditions*. Consequently, the rules viewed the determiner as correct, because it was specified as such.

As was expected, the vast majority of instances are cases where the algorithm proposed two possibilities. One result is the article used in the original text, while the second option is a different option still considered correct by the algorithm. In most cases, the alternatives are not grammatically wrong, but rather stylistically less desirable or strange.

(4.8) a self-induced injury or a false history in order to mislead a physician into

	1961	2006
one correct option	15 (10.6%)	17 (11.2%)
one or more wrong options	25 (17.7%)	25 (16.4%)
2 options, one correct	101 (71.7%)	110 (72.4%)
Total nouns corrected	141	152

Table 9: Overview of Evaluation for the Rule-Based System

making an erroneous diagnosis and administering some type of treatment – *original*

the [an/the] self-induced injury or false [the/∅] history in *the* [∅] order to mislead [the/∅] *physician* into making the [a/the] erroneous diagnosis and administering some type of [the/∅] treatment – *corrected*

In example (4.8), the two wrong articles (in italics) were correctly recognized by the system. For the remaining four articles, the articles in the falsified text are not grammatically wrong, but as aforementioned not entirely correct either given the context. For all four, the system proposes the correct alternative as well. Furthermore, it needs to be noted that the majority of the two options instances are "the/∅". For the 1961 texts, in 93 out of 101 instances, the system suggests using either the definite or ∅-form. For the text snippets from 2006, the ratio is a little less clear, in 67 out of 110 cases the rules propose using either *the* or no article at all.

Not surprisingly, all cases of the accurate correction of articles are instances where a ∅-form was needed. In order for an instance to be considered correct, it needed to produce only one result, which is the same as the article in the original text. As the majority of single-result cases lead to ∅-forms, Table 10 consistent with initial understandings.

original	jumbled	correction	# of instances 1961/2006
∅	∅	∅	10/10
∅	a	∅	2/1
∅	the	∅	3/6

Table 10: All combinations of accurately corrected instances

Table 10 shows that there is little variation in the accurately corrected instances. For both data sets only ∅-forms were accurately detected as either already correct or as wrong and then the appropriate suggestion was made. A more complex picture presents itself when one analyzes the erroneously corrected instances.

original	jumbled	correction	# of instances 1961/2006
the	∅	∅	7/2
the	the	∅	2/0
a	the	the/∅	0/1
a	∅	∅	0/17
a	∅	the/∅	4/0
a	a	an/the	1/0
a	a	the/∅	1/0
an	∅	the/∅	3/0
an	the	a/the	1/0
∅	a	an/the	1/1
∅	the	an/the	2/0
∅	the	a/the	2/1

Table 11: All combinations of wrong instances

Table 11 lists all the instances in which the suggested options are wrong. It is interesting to see that unlike in the correct instances, there is not much overlap between the two data sets. For example, there are no *an*-forms in the 2006 data set, therefore, it cannot be corrected at all. In the two instances where the 'wrong' indefinite article, *a* for *an* or *an* for *a*, was predicted, the noun starts with a vowel but is modified by a token starting with a consonant (or vice-versa).

(4.9) a correcting constitutional amendment – *original*

a [an/the] correcting constitutional amendment – *corrected*

In example (4.9) the rules marked the indefinite article *a* as wrong, and suggested either *an* or *the* to be correct, as the noun begins with a vowel. The rules do not consider the modifying parts of this noun phrase, which make the article *a* correct. In the future, it would make sense to prevent such mistakes by checking the token that directly follows the article and not just the target noun.

(4.10) condemn an entire group of animals – *original*

condemn *the* [a/the] entire group of *a* [the/∅] animals – *corrected*

Example (4.10) illustrates the problem of the modifier, as well as the small semantic differences between definite and indefinite, or definite articles and ∅ forms, which are nearly impossible to grasp for the rule-based correction system. The phrase *the entire group of the animals* sounds strange and not well formulated; nonetheless it is not wrong in a grammatical sense. This feel for correct or incorrect is hard for non-native speakers to learn, and even harder to teach to a machine using rules.

One last difficulty that should be highlighted is the fact that the tagging and parsing of incorrect texts is hard. The difference has already been demonstrated by the different numbers of nouns detected by the two tools. The implications of this can be seen in example (4.11). The rule for proper nouns is that they always have the ∅-form.

(4.11) The decision of the Supreme Court of the United States – *original*

> the [a/the] decision of *the* [∅] Supreme [∅] Court of [∅] United [the/∅] States – *corrected*

This leads to the deletion of the definite article preceding *Supreme Court*, as well as the failure to insert an article in front of *United*. Apart from this problem created by an insufficiently comprehensive rule, the parser has not recognized that *United States* is one proper noun. Therefore, if a definite article had been inserted before *States*, the rules would have considered it to be correct. Moreover, the proper noun *United States* appeared four times in total, resulting in four wrong article corrections with *United*, and four partially correct ones for *States*. This issue and its possible solutions will be discussed further in the conclusion of the current chapter and in Chapter 6, where both correction systems are contrasted and discussed.

To conclude the evaluation, it can be stated that the rules make few mistakes. However, they also get few instances correct in their entirety. The evaluation has also shown the difficulty of analyzing language through the rigid constraints of immutable grammar rules. Nonetheless, the results suggest that such analysis can be done, and with a few improvements the ratio of single option rules may be increased in the future.

4.5 Rule-based Approach Conclusions

The first system engineered to correct articles used prescribed rules to determine the correct article. These rules were deduced from literature and arose during the process of rule writing. This lead to a solid foundation for rule-based article correction. The

development process was done in three steps, moving from simple, generic rules to more complex and specific rules. The rules categorize nouns and then apply different logics to them as needed. The rules were kept as modular as possible, in order to keep the system flexible and limit the number of rules that required hard coding. This process lead to a system that corrects articles with some success. Using this formulation, only 17% of articles are corrected completely erroneously. Conversely, the system fails in guaranteeing the best outcome, as only 10% of all corrected cases are entirely accurate. In the majority of cases, the system proposes two options for the article usage, at least one of which is correct. As was mentioned at the beginning of this chapter, the problem is not necessarily that the correct option is not produced, but rather that too many wrong options or suboptimal options are produced.

The rules which lead to a single article result all produced more correct outcomes than wrong ones. Therefore, it would make sense to invest more time in finding such rules. It may prove to be fruitful to do this from a Construction Grammar point of view, as one needs sentence constructions with fixed article usage, yet other parts of the phrase should be interchangeable. One of the biggest sources of errors is proper nouns. It would be interesting to see if *Named Entity Recognition* (NER) would improve the differentiation between definite articles and ∅-forms. For example, if the compound *United States* were tagged as `country` one could make lists with all the countries that usually take a definite article, like *the United States*, while most other countries use the ∅-form like *Germany* or *Canada*. It would be especially helpful in identifying institutions like *the UN*, for example. NER would add a certain amount of semantic information about the proper nouns. Another way to add semantic information would be to do real co-reference resolution. This was attempted on a smaller scale in the second set of rules, however it was abandoned, after causing too much confusion in the correction process. Co-reference resolution, done properly, could provide crucial information to help make smarter choices between definite and indefinite articles based on the subject's novelty in the broader context.

In conclusion, it can be stated that the rule-based approach performs fairly well and within the expected realm of correctness. Valuable insights into article usage have been gained, some of which might be helpful for extracting features for the machine learning approach in the next chapter. Furthermore, several options for improving the rules have been identified for future research. The rule-based system will be compared to the machine learning system in Chapter 6, and the possibility of a combination of both systems will also be explored .

5 Machine Learning Approach

We are living in the age of 'big data', meaning that the sheer volume of available data can appear to be quite overwhelming to process for analysis. The storage capabilities of our devices are greater than ever before, yet one could also "testify to the growing gap between the *generation* of data and our *understanding* of it"(Witten et al., 2011, original emphasis, 4). Data Mining and Machine Learning methods are meant to help researchers, marketing agents, and producers to better understand the massive amount of data available. Witten et al. state that "intelligently analyzed data is a valuable resource" (2011, 4). In order to analyze intelligently, one needs to search for hidden patterns in the data. This is exactly the purpose of *data mining*; it is the process of automatically detecting patterns in large quantities of data (Witten et al., 2011, 5).

In the following chapter, the concept of machine learning is briefly introduced. Then the algorithms used in this project are explained, followed by an overview of previous works on machine learning and automatic language correction. In the remaining section, the machine learning approach used for this thesis is elaborated on, and the final system is evaluated.

5.1 Concept of Machine Learning

Generally speaking, one can say that humans learn by experience. Witten et al. (2011) argue that in machine learning the *learning* is more tied to *performance* than *knowledge* (7). *The Oxford Handbook of Computational Linguistics* defines machine learning as the "study of computational systems that improve performance on some task with experience" (Mitkov and Mooney, 2012, 2). There are four different basic types of learning in data mining: *classification*, *association*, *clustering* and *numeric prediction* (Witten et al., 2011, 40). For this project *classification* is the only type which is of importance. In *classification* tasks, the machine is taught how to classify different instances into categories. This can be done either *supervised* or *unsupervised*. In the case of *supervised* learning the algorithm is given a set of labels

to learn and chose from, with *unsupervised* learning the algorithm is expected to derive categories from the data.

This project is a supervised classification task. A similar, simplified task will be presented to explain supervised classification. The example task is taken from Witten et al. (2011) and was slightly adapted to fit this thesis. The question is whether or not one should play a specific game outside given the circumstances. The circumstances are determined to be the *outlook*, *temperature*, *humidity* and *wind*, these circumstances are called `features` in machine learning.

outlook	**temperature**	**humidity**	**wind**	**play**
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	**??**

Table 12: Data Set about the Weather

Each row in Table 12 represents an `instance`. In this case, one instance stands for one day, where the game was played or not, and the weather conditions on that particular day. The task then becomes predicting the outcome of the last row based on all the previous experiences. The algorithm takes into consideration how many times the game was played under similar conditions and then makes a prediction. If the predictions improve with repeated exposure, or training, one says that the machine has learned. The example is a binary decision, so either the players play or they do not, and therefore statistically the algorithm will be right about 50% of all cases. Consequently, the algorithm needs to be above the 50% accuracy rate, otherwise it is very obvious that a fundamental flaw exists. Formulated more abstractly machine learning consists of the following steps:

Firstly, the data needs to be formatted in a way that the machine learning tool

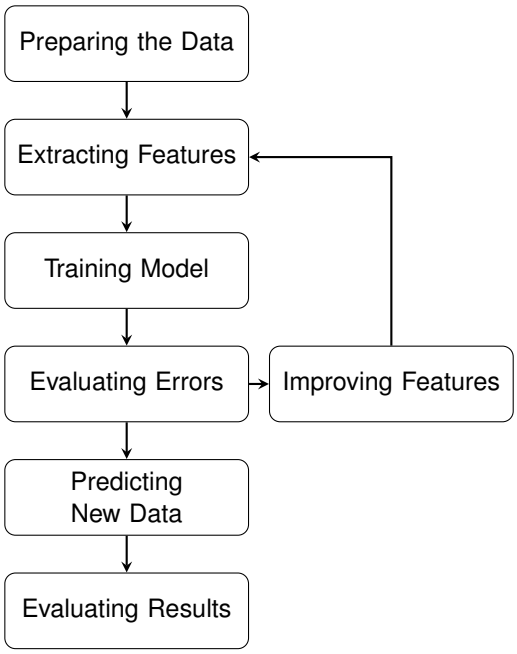

Figure 5: Maschine Learning Process

can process; it additionally might need to be linguistically annotated. In the next step, features are extracted. These features are assumed to help the algorithm make correct predictions. Following the feature extraction, the actual models are trained using different algorithms. Afterwards, the results are analyzed to improve feature extraction and eliminate features which lead to bad predictions. Then the process begins anew. Once a satisfying result has been reached, the trained model is used to predict labels on new, unseen data. It is important that the final prediction on new input is carried out on data which has not been involved in any part of the training or feature extraction. Only entirely new data provides a real, unbiased challenge for the algorithm, and consequently yields the 'true' performance of the trained model. In the final step, results from the prediction are again evaluated for future research.

5.2 Algorithms

An algorithm is defined as "a procedure or set of rules used in calculation and problem-solving" and as "a precisely defined set of mathematical or logical operations for the performance of a particular task" (Oxford English Dictionary, 2003a). In our case, the specific task is to decide what kind of article should accompany a given noun. The results of the machine learning-approach depend on the algorithm

as much as the preparation of the data and feature selection. Therefore, the three algorithms used for this thesis are briefly presented below.

5.2.1 Naive Bayes

Naïve Bayes is based on the Bayes' theorem, proposed by Thomas Bayes in 1763 (Bayes and Price, 1763). The theorem has been seen as a cornerstone of probabilistic theory since its publication. It assumes that all events are independent of each other, and therefore one can multiply the probabilities of single events. The simple formula can be seen in (5.1).

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{5.1}$$

What is the probability of H happening given E has happened? This is calculated by multiplying the probability of E given H and the probability of H divided by the probability of E. Naïve Bayes is proof that "simple ideas often work very well" (Witten et al., 2011, 86), as this simple algorithm rivals or even outperforms many more advanced or sophisticated classifiers (Witten et al., 2011, 99).

5.2.2 Support Vector Machine

Support Vector Machine (SVM) is a more complex algorithm than Naïve Bayes. It can be explained most clearly by using a simplified example. The basic idea is to measure similarity between concepts using vectors. Widdows formulates the mathematical thinking behind SVM as follows:

> If the two points are close together, then the angle in between them is small, and we might say that they are fairly similar to one another: if they are *exactly* the same point, then we shall say that their similarity is equal to 1. On the other hand, suppose the points a and b are at right angles to one another [...] then we might be tempted to say that they have nothing in common at all ... (Widdows, 2004, 105)

Figure 6 illustrates this quote nicely. a and b are two entities which are compared to each other, given their dimensions, they point in different directions. The angle, Widdows mentions is θ, and the smaller it is the closer related are the two entities.

The vectors hold information which describes our instances. In Figure 6, there are only two information pieces per entity, and this obviously is not enough to represent

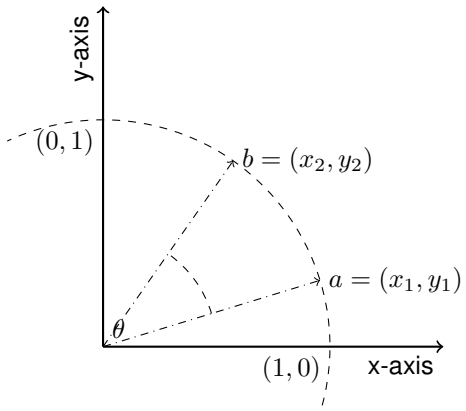

Figure 6: Cosine Similarity adapted from (Widdows, 2004, 105)

'reality', but it is easier to visualize a two dimensional example than a five to 100-dimension real world problem. To return to the weather data, the vectors for the first two days could look like this:

$$\begin{pmatrix} sunny \\ hot \\ high \\ no-wind \end{pmatrix} \begin{pmatrix} sunny \\ hot \\ high \\ wind \end{pmatrix}$$

The order of the values needs to remain the same for all vectors, namely first *outlook*, followed by *temperature*, *humidity* and *wind*. The machine obviously needs numerical values to process the vectors, therefore, a value has been assigned to each weather condition; *1* for *sunny*, *2* for *rainy*, and *1* for *hot*, *2* for *mild* etc. This encoding translates into the following vectors:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Now one can see that the vectors are very similar, as they only differ in the last position, therefore, they can be considered to be similar and will consequently yield the same answer to the question 'should the game be played?'. If a new vector is seen, it will be compared to the known vectors and based on the similarity to either *play* or *not play* the day will be classified accordingly.

The mathematical equations which are needed to successfully calculate the similarity between any two given vectors are the following:

$$||a|| = \sqrt{\sum(a_i^2)} = \sqrt{a \cdot a} \tag{5.2}$$

$$(x, x') := \sum_{i=1}^{N}[x_i][x_i'] \tag{5.3}$$

(5.2) calculates the norm of a vector, and with this one can normalize all vectors to a length 1, which is called the *unit vector*. This is usually done to circumvent penalties for extremely frequent or infrequent contexts. The *dot product* or *scalar product* in (5.3) computes the angle between two vectors, which results in a similarity measurement as illustrated before (Widdows, 2004, 152-157), (Scholkopf and Smola, 2001, 1-3).

For this project, the vectors describe each noun in the data. This means that the vectors have over ten coordinates. The algorithm learns in which direction vectors with a definite article point and then, based on the similarity to a category labels new data.

5.2.3 Logistic Regression

The third algorithm which was chosen is called *Logistic Regression*. Logistic Regression was first proposed by David Cox (1958). He assumed that if you have a binary class of either *0* or *1*, the probability of an instance being either one, depends on the values of independent variables (Cox, 1958, 215). Formulated differently, the aim is to model a conditional probability $Pr(Y = 1|X = x)$ as a function of x, the unknown parameters will be estimated using maximum likelihood (Shalizi, 2013, 224). This can be achieved with a logistic regression model.

$$log\frac{p(x)}{1 - p(x)} = \beta_0 + x \cdot \beta \tag{5.4}$$

If one solves (5.4) for p, the formula then translates to:

$$p(x; b, w) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}} = \frac{1}{1 + e^{-(\beta_0 + x \cdot \beta)}} \tag{5.5}$$

This results in a linear classifier. However, logistic regression is more than just a classifier, because it states that "the class probabilities depend on distance from the boundary in a particular way", this way it makes "stronger, [and] more detailed

predictions" than other algorithms (Shalizi, 2013, 225). Logistic regression, similar to Naïve Bayes, performs very well given its simplicity. Additionally, there is a long tradition of applying it to text data (Shalizi, 2013, 227). Therefore, it was included as a third algorithm for this thesis. Our classification is not a binary one, and therefore some minor modifications need to be done to the equation (5.5). The modified version can be found in the appendix.

5.3 Automatic Language Correction using Machine Learning

Machine Learning has been used on an array of grammar correction tasks, though they have largely focused on determiners and preposition errors. Before introducing a selection of past studies, two different approaches to language correction using machine learning will be presented. All of the presented studies depend on training data, as well as good linguistic annotation. Sakaguchi et al. (2012) propose a system to correct spelling errors jointly with POS-tagging mistakes. This is done because many English as a Second Language (ESL) studies depend on correct POS-tagging and parsing of the data. However, if the data contains many spelling errors, the POS-tagging will be riddled with mistakes, and consequently, the parsing will not work properly either (Sakaguchi et al., 2012, 2358). Therefore, the team developed a machine learning system which first corrects seven different types of spelling errors[1] and then tags the ESL text. Using the Cambridge Learners Corpus First Certificate in English (CLC FCE) data set, they were able to see a 2.1% increase in their F-value compared to the baseline, which is statistically significant (Sakaguchi et al., 2012, 2366). The classifier performed even better on the Konan-JIEM learner corpus, which consists of essays written by Japanese ESL students. There, the increase in performance measured 3.8% (Sakaguchi et al., 2012, 2361-66). The most important insight from this study is that this approach results in better POS-tagging than the pipeline approach, where spelling mistakes are corrected prior to the POS-tagging (Sakaguchi et al., 2012, 2370). Tajiri et al. (2012) focus on another area that often causes difficulty for ESL learners, namely tense and aspect. This type of correction is very difficult, as it relies heavily on global context (198). They defined 14 local features, meaning that the features relate directly to the verb phrase which needs to be corrected, such as *auxiliary verb* or *word to the left*. They also defined global fea-

[1] typographical (**gramar/grammar*), homophone (*see/sea*), confusion (**form/from*, split (**never the less/ nevertheless*, merge (**baseballbat/baseball bat*), inflection (**swim/swims*) and derivation (**well/good*) errors (Sakaguchi et al., 2012, 2358)

tures. These features include information from verb phrases preceding or following the current target verb phrase (Tajiri et al., 2012, 199-200). With this approach, the system detects 61 out of 211 instances where simple present tense was used instead of simple past, and corrects 52 instances. The second most common error, using simple past instead of simple present, is only detected in nine instances out of 94 by the system. This significant drop in performance can be easily explained by the fact that the second error is harder to detect for the computer. Thus they intend to intensify their research to detect less obvious errors as well (Tajiri et al., 2012, 201).

Dahlmeier and Ng (2011) introduce the *NUS Corpus of Learner English* (NUCLE). This corpus consists of around 1400 essays written by non-native, university students on a variety of topics. This results in over one million tokens fully annotated with error tag corrections (Dahlmeier and Ng, 2011, 918). The NUCLE has been used by many studies as a training or evaluation data set. Dahlmeier and Ng (2011) are able to show that grammar correction tools perform better when trained on data from NUCLE rather than tools which were trained on correct data. They corrected article and preposition usage and concluded that especially for articles the main difficulty lies in the fact that often more than one choice is correct (920-21). One of the studies which makes use of NUCLE is carried out by Xiang et al. (2013). They claim that up to 12% of all errors recorded in NUCLE are errors in article usage. To reduce complexity during the training process without losing performance a so-called genetic algorithm was used to lower the number of feature dimensions. In other words, the features leading to the clearest predictions were determined, and confusing features were eliminated. After some confidence tuning, the Maximum Entropy classifier outperforms the system by Dahlmeier and Ng (2011) by 2.2% in the F-value (Xiang et al., 2013, 1067-1069). The same type of classifier, Maximum Entropy, is also used by Han et al. (2004). They trained their classifier on about eight million noun phrases stemming from the *Meta Metrics Corpus*. An interesting insight from their feature evaluation is that the more a head noun appeared, the better the classifier predicted the correct article. The classifier was then tested on *Test of English as a Foreign Language* (TOEFL) essays written by native speakers of Russian, Japanese, and Chinese. These languages were intentionally chosen as they do not have articles. The classifier reached an average accuracy of an impressive 83.00% (Han et al., 2004, 1626-1627). A large portion of the mistakes originated in the fact that often "the highest probability outcome was only marginally greater than the second highest", moreover both options were often grammatically correct. Nevertheless, based on context a human annotator clearly favors one over the other, for example "because an entity was being mentioned the first time" (Han et al., 2004, 1268).

Finally, two slightly different approaches to article correction will be illustrated. The first reiterates the point that "writing errors often present a semantic dimension that renders it difficult if not impossible to provide a single correct suggestion" (Gamon et al., 2008, 1). To circumvent this problem, Gamon et al. (2008) develop a three step system. First, they train two classifiers for the *suggestion provider*, for both types of errors (articles and prepositions). The first classifier determines whether or not an article/preposition is needed, and the second makes a choice as to which article/preposition should be used, given that one is required. Second, a *language model* is learned, to rate the original user input against the suggestion made by the suggestion provider. And third, they construct an *example provider*, which provides the user with similar examples to the sentence the user has written (Gamon et al., 2008, 2-4). This system performs with an accuracy of 86.07% for the determiner correction, where the most common correction is the insertion of a missing article (Gamon et al., 2008, 5-7). The final paper which will be presented approaches the problem of sparse incorrect data from which the machine can learn by introducing errors into their data. Rozovskaya and Roth (2010) use different methods to introduce the erroneous articles ranging from randomly replacing articles to using error patterns from ESL writings to mimic the writing of an ESL learner (158-159). Classifiers were trained on the different modified data sets and then applied to three sets of data, namely ESL writing by Chinese, Russian and Czech speakers. For each native language, a different method of introducing errors worked best, but the error reduction lay between 8% and 16% (Rozovskaya and Roth, 2010, 160). These results suggest that classifiers which are trained on data containing errors perform better than classifiers which are trained on clean or correct data (Rozovskaya and Roth, 2010, 160).This is corroborated by Dahlmeier and Ng (2011).

The field of automatic language correction is very broad, and there are a lot of new insights constantly emerging. Much can be achieved with the proper set of tools and training data as was demonstrated by the by the cited studies. In the following sections, the data used for this project will be presented.

5.4 Data

The quality and type of the data is much more important for the machine learning approach considering the data is the basis for the learning. Consequently, the data will be presented in greater detail.

The corpora chosen for this project are all part of the *Brown Corpus Family*. *The Brown Corpus* was the first large text collection which was computer-readable and

made available with linguistic annotation. The original corpus was compiled during the 1960s and contained over 1 million words of American English printed during the calendar year 1961. The corpus is split into 15 genres; a complete list can be found in the appendix. This project seeks to correct article usage in *academic* writing, thus the chosen data stems from the academic section of the corpora. This data set will be referred as *Brow*n for the rest of the thesis. Brown academic consists of around 187‘000 tokens (including punctuation) which make up over 6‘800 sentences. The distribution of articles can be seen in Table 13. This corresponds to data cited by Biber et al. (1999), although no claims about ∅-forms are made there.

	Brown		**AmE06**		**BOTH**
definite	13‘367	35.33%	11‘648	26.80%	25‘015
indefinite (*a*/*an*)	4‘811 (3‘975/836)	12.71%	4‘785 (3‘938/847)	11.02 %	9‘596 (7‘913/1‘683)
∅	19‘641	51.92%	27‘006	62.17%	46‘647
total	37‘819		43‘439		81‘258

Table 13: Overview over the Distribution of Article in the Data Sets

The second data set, AmE06 comes from the Brown family as well. It was compiled at Lancaster University in early 2011 to mirror the BE06. It consists of texts written by Americans or people who have lived in America for a substantial amount of time, published between 2004 and 2008. 400 out of the 500 texts were published in 2006 (Potts and Baker, 2012, 301-302). Using only the academic portion of the corpus as training material, roughly 192‘000 tokens (including punctuation) in 6‘800 sentences were identified. The distribution of articles can also be found in Table 13. The third data set listed in Table 13 combines both data sets.

All articles in these data sets are assumed to be used correctly. This is done for a number of reasons. Firstly, a lot of manual annotation on behalf of the researcher can be avoided. Secondly, as was shown in Chapter 2, article use is subjective and highly variable. By assuming the real life data is correct, an attempt at representing the complexities of normal article use by native speakers is made. Thirdly, a natural distribution of articles is achieved. The following procedure was performed on all three data sets, in order to make comparisons possible and see how different inputs affect the machine learning results.

5.5 Pre-Processing

The pre-processing procedure is similar to the rule-based approach elaborated in Chapter 4.2. The raw text was tagged using the *TreeTagger*. The tagged output was transformed into the conll format for the parsing process with the *Maltparser*. In contrast to the rule-based approach, the pre-processing is vital to the performance of the machine learning algorithms, as the algorithms learn what is correct on the basis of the data. An additional step required for machine learning is the conversion of the data into a csv-file.[2] *Lightside* expects this format, as has been mentioned in Chapter 3. The features are extracted from the data and then stored in a file as csv. Afterwards, the pre-processed data is fed into the machine learning workbench.

5.6 Feature Extraction & Selection

The features form the bedrock of the algorithm's learning pattern. Therefore, they greatly influence the results. Features are attempt to define all the factors which affect article choice. Thus, several features describe underlying syntactic structures, while others try to give the algorithm semantic information about the instance. In this section, all features considered in this process will be described and the features ultimately selected will be explained in greater detail. The final features can be found in Table 14, alongside a short description, example, and the number of values the features resulted in.

Similar aspects were taken into consideration as in the rule-based approach. The noun itself influences the article choice, therefore the noun and its POS-tag were selected as a feature. Whether or not a preposition follows the noun, also influences article selection. Furthermore, depending on the preposition, different choices regarding the articles are made. This was taken into consideration with the *preposition* feature. If the noun is not followed by a preposition, the default value 'none' is assigned. The distinction between indefinite *a* and indefinite *an* is addressed with the feature *vowel*. However, this solution does not take into consideration whether the article is immediately followed by the target noun or if a modifier exists.

(5.6) a house → `no`

(5.7) an interesting house → `no`

Consequently, both examples (5.6) and (5.7) have the value `no` for the vowel feature,

[2] comma separated values

even though there is an adjective beginning with a vowel in (5.2). This is something which was ignored. *Lightside* has several built-in features, including both standard and POS-bigrams. However, the context of bigrams proved to be too small. In (5.8) one can see that the bigrams do not adequately represent the syntactic structure of the noun phrase.

(5.8) *The green house on the hill.*
DT JJ NN IN DT NN.
DT-JJ JJ-NN NN-IN IN-DT DT-NN

For the noun phrase *the green house*, the bigrams do not encompass the entire noun phrase, and therefore results in confusing bigrams. This problem is circumvented by using POS-Trigrams. At first, all trigrams are utilized using the built-in feature, which made the training process very slow and did not improve the performance significantly. Therefore, the features *Trigramm-ONE/TWO/THREE* are extracted. These trigrams only include POS-trigrams where a noun tag is in either the first, second or third position. This way, the entire first noun phrase in (5.8) is captured. As can be seen in Table 14, the number of unique values is still rather large, though a lot smaller than if all possible POS trigrams were used.

Another two features relate to the immediate context of the noun, namely *Modifier Relation* and *Modifier Tag*. The relation feature describes the kind of relation the modifier has to the noun, and this gives clues as to the type of noun phrase. For example, if the relation is *nn* it is clear that the noun is part of a compound, as in *baseball bat*. The second feature, the *Modifier Tag*, simply registers the POS of the modifying element. If the noun is not modified, then the default value for both features is 'none'. If there is more than one modifier, the one immediately to the left of the noun is registered in the features. The last feature attempts to provide some semantic information. *Co-Reference Resolution* checks whether the noun has been used in the previous five sentences. However, it does not do real co-reference resolution, as the extraction process will not recognize that the sentences (5.9/10/11) are about the same person. It examines only whether the exact word has been used before.

(5.9) The president is giving a speech.

(5.10) People cheer for him.

(5.11) Obama makes a great speaker.

A further complication is that proper names like *president* almost always take the

[2]The number of values differ between the different data sets.

Feature	Description	Example	Number of Values[2]
Noun	The noun which this instance is concerned with.	*moon*, *emission*	8‘000-10‘000
Noun Tag	The POS-tag of the noun which this instance is concerned with.	NN, NNP	4
Modifier Relation	The dependency relation between the modifying element and the noun.	nn, amod	22-26
Modifier Tag	The modifying elements POS-tag.	JJ, NN	30-33
Preposition	The preposition which follows after the noun.	*of*, *in*	120-130
Co-Reference Resolution	If the noun has appeared in the preceding five sentences.	yes/no	2
Vowel	If the noun starts with a vowel.	yes/no	2
Trigram-ONE	POS-Trigram, where the noun tag is in first position.	NN-PREP-DET	1‘091-1‘192
Trigram-TWO	POS-Trigram, where the noun tag is in second position.	DET-NN-EOL	942-1‘150
Trigram-THREE	POS-Trigram, where the noun tag is in third position.	DET-JJ-NN	1‘316-1‘168
***LightSide* UNIGRAM**	The *LightSide* built-in unigram feature, except *the*, *a* and *an*.	*good*, *is*	unclear

Table 14: List of all Features used in the Machine Learning

definite article, independent of how often they have appeared. Finally, *UNIGRAM* the built-in feature of *LightSide* was used. This feature uses a bag-of-words approach, and the workbench completes the selection process for the user. One modification was made though. Given that the article is what the algorithm is supposed to predict, *the*, *a* and *an* were excluded from the feature list in a restructuring step. If left in the feature list, the algorithm would already ‘see’ in the data what it is supposed to learn.

The eleven described features were then used to train individual models. For the purpose of understandability, the process of feature extraction, feature selection, and training the models is described as a linear process. However, as was shown in Figure 5, the process is far more circular than suggested here.

5.7 Model Training

In total, over 50 different models were trained for this thesis. This is because each feature combination was used to train three different algorithms, three different data sets were used, and promising initial results prompted more attempts. For each set of features, a specific algorithm, data set, and set of categories had to be chosen. An overview of all options can be found in Table 15.

Algorithms	Naïve Bayes, SVM, Logistic Regression
Data Set	Brown, AmE06, combined
Categories	definite, indefinite, zero; definite, indefinite-vowel, indefinite-consonant, zero

Table 15: Different Dimensions to choose from for Model Training

Each trained model is evaluated with a ten fold cross validation by *Lightside*; this returns an accuracy as well as a confusion matrix. The ten fold cross validation is the standard evaluation form for machine learning tasks.

As a baseline, the two features *noun* and *noun tag* were chosen, resulting in the accuracies found in Table 16. It became apparent that the three algorithms are

	Brown	AmE06	Combined
Naïve Bayes	57.4%	64.52%	61.32%
SVM	58.53%	65.48%	62.3%
LogReg	58.62%	66.16%	62.68%

Table 16: Baseline for all Algorithms and Data Sets

very similar, however, there is great variation between the data sets. This finding will be discussed in more depth in the section 5.8. For space purposes not every model trained will be introduced, but the general learning curve will be presented in six training cycles. All models presented are contrasted to the baselines found in Table 16.

5.7.1 Brown – First Training Cycle

For the first of six training cycles, the dimensions described by Table 17 were used in conjunction with the Brown data set. After some preliminary testing, bigrams

proved to result in overly noisy data, as was explained in section 5.6, and therefore they were not included in any training cycle.

	Noun	
	Noun Tag	
Naïve Bayes	Modifier Tag	definite
SVM	Modifier Relation	indefinite
Logistic Regression	Preposition	∅

Table 17: Dimensions for the First Training Cycle

The dimensions are rather simple and attempt to capture about the same level of information as the first round of rule development in section 4.3.1. The one addition is, that it was not only checked whether a noun is modified but also recorded how the noun was modified. The Brown data set was used as training material, as aforementioned, and all three algorithms were applied to it. This resulted in the accuracies found in Table 18. One can see, that all algorithms improved

	Accuracy	**Baseline**	**Improvement**
Naïve Bayes	62.09%	57.4%	+3.69%
SVM	62.61%	58.53%	+4.08%
Logistic Regression	63.24%	58.62%	+3.63%

Table 18: Results for the First Training Cycle

significantly compared to the baseline. While the SVM model has improved the most in absolute terms, the Logistic Regression model has the highest accuracy, even though it improved the least. If one studies the confusion matrix for the SVM[3], it becomes apparent that not all article predictions were equally valid. The

Act\Pred	**definite**	**indefinite**	∅
definite	**7683**	785	4899
indefinite	2210	**979**	1622
∅	4022	601	**15018**

Table 19: Confusion Matrix for SVM, First Training Cycle

Confusion Matrix displays the actual article and the predicted article. Consequently, the diagonal cells in bold represent the correctly predicted instances. The 2210 instances found at the juncture of the first column and the second row represent all

[3]All confusion matrices can be found in the appendix.

the instances where the text originally had an indefinite article, but the algorithm assigned a definite one, as a reading example. Upon closer inspection, it becomes apparent that the algorithm had fewer difficulties deciding between 'no article' and 'some article' then definite or indefinite article. In regards to the indefinite article, more predictions were ultimately incorrect. The decision between indefinite and definite articles also posed a difficult challenge for the rule-based approach, as noted before.

More contextual information is needed to facilitate the definite/indefinite choice. Therefore, POS-Trigrams were added in the second training cycle.

5.7.2 Brown – Second Training Cycle

The POS-Trigrams, which were added as features, in this training cycle contained only trigrams with a head noun, as was explained 5.6. Though *Lightside* has a built-in POS-Trigram feature, there were too many possible combinations for POS-Trigrams, and most of them were unnecessary and created undue noise in the data. Table 20 contains a list of all dimensions for the second training cycle, taken again from the academic portion of the Brown Corpus.

	Noun	
	Noun Tag	
	Modifier Tag	
	Modifier Relation	
	Preposition	
Naïve Bayes	Trigram-ONE	definite
SVM	Trigram-TWO	indefinite
Logistic Regression	Trigram-THREE	∅

Table 20: Dimensions for the Second Training Cycle

The results can be seen in Table 21 and are quite promising. All algorithms have increased the accuracy over 10% compared to the baseline, as well as a significant improvement over the first training cycle.

In order to determine if the POS-Trigrams facilitated the decision between indefinite and definite articles, the confusion matrices for Logistic Regression of the first and second training cycle are presented in Table 22.

The number of correctly predicted indefinite articles has risen significantly, and fewer indefinite articles were labeled as ∅-forms. However, there was a slight increase in

	Accuracy	Baseline	Improvement
Naïve Bayes	67.59%	57.4%	+10.19%
SVM	70.39	58.53%	+11,86%
Logistic Regression	71.23%	58.62%	+12.61

Table 21: Results for the Second Training Cycle

Act\Pred	definite	indefinite	∅
definite	**7866**	518	4983
indefinite	2430	**776**	1605
∅	4007	356	**15278**

Act\Pred	definite	indefinite	∅
definite	**9775**	793	2799
indefinite	2464	**1282**	1065
∅	3337	422	**15882**

Table 22: Confusion Matrices For First (left) and Second (right) Training Cycle, Logistic Regression

the number of indefinite articles mislabeled as definite. A further observation can be made about the definite articles: more definite articles were predicted correctly the first time, and fewer were confused with the ∅-form. It nevertheless appears that the desired effect of better distinguishing between definite and indefinite articles has not taken place; instead, the decision between an article or no article was made clearer. Another attempt at adding the necessary information for making the indefinite/definite decision more straightforward for the algorithms was made in the third training cycle.

5.7.3 Brown – Third Training Cycle

In the preliminary test, the built-in feature UNIGRAM from *LightSide* promised good results and so it was incorporated at this stage. It was not used earlier because it is not entirely clear how *LightSide* extracts these features, as was explained in section 5.6. The articles were removed from the UNIGRAM-featutre for reasons explained in 5.6 as well. As listed in Table 23, the feature *resolution* was added as well.

This was done using a similar methodology as the rule-based approach described in section 4.3.2. The feature was included with the hope that it, in conjunction with the UNIGRAM feature, would provide the necessary semantic information. The third feature added in this cycle is the *vowel*-feature.The shortcomings of this feature, discussed in section 5.6, are relevant to all further analysis.

	Noun	
	Noun Tag	
	Modifier Tag	
	Modifier Relation	
	Preposition	
	Trigram-ONE	
	Trigram-TWO	
	Trigram-THREE	
Naïve Bayes	Resolution	definite
SVM	Vowel	indefinite
Logistic Regression	*LightSide* UNIGRAM	∅

Table 23: Dimensions for the Third Training Cycle

	Accuracy	**Baseline**	**Improvement**
Naïve Bayes	68.25%	57.4%	+14.08%
SVM	72.17%	58.53%	+13.64%
Logistic Regression	73.98%	58.62%	+15.36%

Table 24: Results for the Third Training Cycle

The results shown in Table 24 are promising and have significantly improved compared to both the baseline and the previous training cycle, which is reflected in the confusion matrices. In general, more instances were labeled as indefinite;

Act\Pred	**definite**	**indefinite**	∅
definite	**9775**	793	2799
indefinite	2464	**1282**	1065
∅	3337	422	**15882**

Act\Pred	**definite**	**indefinite**	∅
definite	**9750**	1033	2584
indefinite	1909	**1923**	979
∅	2731	606	**16304**

Table 25: Confusion Matrices for Logistic Regression, Second (left) and Third (right) Training Cycle

consequently, there were more correctly predicted indefinite articles. Nonetheless, there were still 1900 instances where an indefinite article was labeled as definite and around 1000 where the definite article was labeled as indefinite. Furthermore, results showed an increase in the number of instances of actual ∅-forms being labeled as indefinite. One can argue that the errors made while deciding between the definite form and the ∅-form are more numerous, however, in comparison to the correctly

made decision, they do not have too much of an impact.

It is interesting to note that, if one looks at the Confusion Matrices for SVM with the same features, a different picture presents itself. It appears that Support

Act\Pred	definite	indefinite	∅
definite	**9487**	1134	2746
indefinite	2231	**1555**	1025
∅	3373	690	**15578**

Act\Pred	definite	indefinite	∅
definite	**9258**	1406	2703
indefinite	1725	**2100**	986
∅	2854	851	**15936**

Table 26: Confusion Matrices for SVM, Second (left) and Third (right) Training Cycle

Vector Machine has less problems assigning the indefinite article correctly, even in the second training cycle. Compared to the values in Table 25, the SVM algorithm assigns higher numbers of correct indefinite articles overall. However, it is not able to distinguish between definite and ∅-forms as well as the Logistic Regression. To better understand this difference, a deeper examination of the algorithms would be required, which is beyond the scope of this thesis.

5.7.4 Brown – Forth Training Cycle

The impetus behind this fourth cycle was to determine if the algorithms make better decisions when the indefinite category is split into two spelling variants. Thus, *indefinite-consonant* stands for all instances with an *a* as the article, and *indefinite-vowel* for all instances with *an* as the article. The feature *Vowel* was already added during the last cycle, and it will attempt to provide the necessary context for the choice between the two categories.

In Table 27 one can see the additional categories listed, and identical features. The results displayed in Table 28 show great of improvement compared to the base line.[4]

Nevertheless, the results are slightly worse than the ones in the third training cycle, see Table 24. Naïve Bayes, especially, created problems due to the additional category. In all likelihood, the algorithm functions best with binary classifications; with four categories to analyze, the task is far from binary. The confusion matrices for SVM in Table 29 and Logistic Regression in Table 30 respectively show that an additional category does not make the article prediction any easier.

For SVM, the algorithm assigns more indefinite-vowel articles than Logistic Regres-

[4]The models were also applied to the AmE06 data set, the results are listed in the appendix.

	Noun	
	Noun Tag	
	Modifier Tag	
	Modifier Relation	
	Preposition	
	Trigram-ONE	
	Trigram-TWO	
	Trigram-THREE	definite
Naïve Bayes	Resolution	indefinite-consonant
SVM	Vowel	indefinite-vowel
Logistic Regression	*LightSide* UNIGRAM	∅

Table 27: Dimensions for the Forth Training Cycle

	Accuracy	**Baseline**	**Improvement**
Naïve Bayes	67.44%	57.4%	+10.04
SVM	71.42%	58.53%	+12.89%
Logistic Regression	73.3%	58.62%	+14.68%

Table 28: Results for the Forth Training Cycle

sion does. While they both confuse many indefinite cases for definite articles, neither chose the 'wrong' indefinite article a significant number of times.

Overall, it can be said that differentiating between indefinite articles did not facilitate better prediction of indefinite articles, but rather it resulted in more opportunities for the program to make errors. Therefore, the indefinite article was amalgamated once more.

5.7.5 AmE06 – Fifth Training Cycle

So far every model has been trained on the academic portion of the Brown corpus. To compare how well the features perform on other texts, the same features used in the third training cycle were used on the AmE06 academic portion. As was described in the section 5.4, this data set has been compiled out of texts published in 2006; therefore there is a 45-year time difference between the two different data sets. The results, found in Table 31, compare the accuracy baseline with the AmE06 baseline. As such, while the accuracies are the highest so far, the improvement is

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**9326**	1110	230	2701
indefinite-consonant	1563	**1528**	133	855
indefinite-vowel	332	161	**179**	164
∅	2861	705	131	**15944**

Table 29: Confusion Matrix for SVM Forth Training Cycle

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**9886**	768	93	2620
indefinite-consonant	1655	**1395**	60	866
indefinite-vowel	394	146	**103**	193
∅	2782	471	49	**16339**

Table 30: Confusion Matrix for Logistic Regression, Fourth Training Cycle

less significant because the AmE06 baseline is already better than the one for the Brown data set.

	Accuracy	Baseline	Improvement
Naïve Bayes	70.48%	64.52%	+5.69%
SVM	76.74%	65.48%	+11.26%
Logistic Regression	78.19%	66.16%	+12.03%

Table 31: Results for the Fifth Training Cycle

The same general trends can be observed as before: Logistic Regression has the highest accuracy, while Naïve Bayes has the lowest. However, the difference between the algorithms is more pronounced than before. To illustrate the difference between the two algorithms, the confusion matrices are listed in Table 32.

It is interesting to note that Naïve Bayes labels more indefinite articles correctly, though it also mislabels far more definite and ∅-forms as indefinite. Logistic Regression, comparatively, mislabels fewer instances of definite and ∅-forms as indefinite, but misses many of the actual indefinite articles. A further observation can be made about the choice between definite articles and ∅-forms. While Logistic Regression makes about the same number of mistakes in regards to definite instead of ∅-forms

Act\Pred	definite	indefinite	∅
definite	**7945**	1752	1951
indefinite	1756	**2421**	608
∅	4354	1966	**20686**

Act\Pred	definite	indefinite	∅
definite	**7878**	949	2821
indefinite	1525	**2023**	1237
∅	2298	645	**24063**

Table 32: Confusion Matrices for Naïve Bayes (left) and Logistic Regression (right), Fifth Training Cycle

and ∅-forms instead of definite, Naïve Bayes mistakes twice as many of the instances of definite articles for ∅-forms. To determine causation, one would have to analyze many of these instances in depth to determine the main driving factor behind the recommendation. With this new knowledge, the features might be adapted and the accuracies further improved. However, this sort of time consuming investigation would have been beyond the scope of this thesis.

5.7.6 Combined – Sixth Training Cycle

In machine learning it is usually the case that the more training data one has the better the results get. Therefore, the two data sets were combined, and used to create a bigger training set for the features used in the third and fifth training cycle. The accuracies in Table 33 are compared to the baseline drawn from this data set.

	Accuracy	Baseline	Improvement
Naïve Bayes	70.58%	61.32%	+9.26%
SVM	75%	62.3%	+12.7%
Logistic Regression	76.4%	62.68	+13.72%

Table 33: Results for the Sixth Training Cycle

As was the case before, all the algorithms improve compared to the baseline, though the accuracies are slightly lower for SVM and Logistic Regression than with the smaller AmE06 data set. Naïve Bayes has a marginally better accuracy with the combined data set than with solely AmE06.

The confusion matrix in Table 34 for Logistic Regression does not reveal any new observations. The predictions are made with similar ratios of mislabeled instances as before. Roughly the same number of cases are mistaken for the definite and ∅-form, and vice versa. Overall, it does not appear to be the case that more data improves

Act\Pred	definite	indefinite	∅
definite	**17763**	1991	5261
indefinite	3470	**3981**	2145
∅	5058	1249	**40340**

Table 34: Confusion Matrix for Logistic Regression, Sixth Training Cycle

the predictions in this particular case. However, the diversity of the data may be affecting these results. Even though the data stems from the same genre, there is a 45-year difference in the production time, and research has shown that article usage is in decline. Some linguists argue that the use of the definite article has decreased by 50% over the last 100 years (Libermann, 2016, p. 1). Given the significant time difference, there seems to be ample reason to call the data diverse, in the sense that it represents different language systems. This issue will be investigated further in the evaluation section.

In conclusion, the models for the machine learning have been trained in six cycles, using different features and three different training data sets. As a general trend, Logistic Regression proved to be the best performing algorithm and the same features lead to better results on more recent training material. POS-Trigrams, information about the noun itself, and the built-in feature UNIGRAM proved to be most accurate in predicting correct articles. In the next step, a few models will be evaluated on the same texts using the rule-based approach (4.4.1), which have not been included in any step of the training.

5.8 Evaluation

For the evaluation, the initial concept was to choose the most effective models for each data set and algorithm and compare them using the two evaluation texts. However, *LightSide* has a bug which deemed unfixable given the time available; therefore the evaluation process was modified. Nevertheless, the evaluation texts remain identical to the texts used in the rule-based approach. How and why the texts were selected was described in detail in section 4.4.1. As mentioned earlier, small differences between the two data sets become highly relevant when evaluating machine learning performance. The different time span in which the data were compiled led to large variations in performance. During the evaluation, models

trained on either data set will be tested on the evaluation texts from 1961 and on the text snippets from 2006. Apart from the influence of time of publication of the training texts, it will also be examined whether the models trained on the combined data set perform equally well on both evaluation sets or see a marked improvement, as it is familiar with more diverse data.

As previously mentioned, *LightSide* did not allow for the evaluation continue as planned; therefore, WEKA was used to evaluate a smaller version of the data sets using just one algorithm. WEKA does not support the built-in feature UNIGRAM from *LightSide*, however, so the new models had to be trained without it. All three data sets were downsized to 10‘000 instances per data set, as this was deemed sufficient for the evaluation. However, different versions of the data sets were used. The label distribution is not balanced,which can create complications during the evaluation. For example, in the Brown data set, there are 5550 instances with the label *zero*, 3331 instances with the label *definite* and only 1118 instances for the label *indefinite*. This imbalance leads to poor performance in the evaluation of predicting indefinite articles, as the model is not able to find enough distinct patterns. In order to counteract this phenomenon, one can balance the labels by suggesting to the learning algorithm that all labels appear equally often. The models which are trained on balanced data, predict indefinite articles very well, though overall they predict too many indefinite articles. This then leads to more errors in regards to definite articles and ∅-forms. The confusion matrices for all models trained for the evaluation can be found in the appendix. In the following evaluation, not all models are discussed in detail. The general trends and problems with both balanced and imbalanced data sets will instead be shown.

The first data set to be evaluated is the Brown data set. Training a model applying Logistic Regression leads to an accuracy of 74.04% for the regular data set and 64.60% for the balanced data. This difference mostly stems from definite and ∅-forms being predicted as indefinite forms, which is a direct consequence of balancing the data. In Table 35, the accuracies of the four evaluations are listed.

Brown	1961 Texts	2006 Texts
normal	68.62%	68.86%
balanced	67.97%	74.25%

Table 35: Accuracies from the Evaluation of the Models trained on Brown

The accuracies vary widely. The assumption was that models trained on text from the same time period as the evaluation text will perform better. This does not hold true for the Brown portion of the evaluation. In fact the balanced data set reaches

almost a 6% higher accuracy on the 2006 evaluation text. This is most likely due to favorable patterns in the evaluation text. As has been stated before, the models perform better on data which is similar to their training data.

Aside from the the accuracies, the confusion matrices paint an interesting picture in regards to the type of mistakes that were made. In Table 36, it is immediately apparent that no indefinite articles were predicted. This is again a function of the unbalanced distribution of labels. It is more interesting to have a look at the choice between ∅-form and definite article. 19 ∅-forms have been mistaken for definite instances, while only six definite articles have incorrectly been assigned a ∅-form. The two phrases in (5.12) show cases where the algorithm predicted a definite article

Act\Pred	**definite**	**indefinite**	∅
definite	**31**	0	6
indefinite	26	**0**	1
∅	19	0	**84**

Table 36: Confusion Matrix for Brown evaluated on 2006 Texts

instead of a ∅-form. The first phrase, is clearly wrong with a definite article for two reasons. First, *love* is an uncountable noun, and therefore, does not take an article for most of the cases. Second, the expression *such + NOUN + VERB* is a semi-fixed expression, where the noun is never accompanied by an overt article. The second phrase highlights the difficulty of dealing with determiners. The determiners were left in the process of prediction and assigned the label `zero`. This, of course, is not correct. The determiner *this* refers to a concept from the previous sentence which makes the sentence understandable, and the definite article cannot do this, rendering the phrase ungrammatical.

(5.12) Such [the] love manifests itself ...

This [the] is part of ...

(5.13) And *the* [∅] judgments are carried out in decisions ...

One of *the* [∅] more well-known factitious illnesses ...

In (5.13) the mistakes happened the other way around: the model inserted ∅-forms when definite articles would have been correct. The first phrase is grammatically correct with the ∅-form, but the semantics change slightly. The second phrase, comparatively, is grammatically incorrect without the definite article. The phrase is again a semi-fixed expression, where an article is necessary.

When evaluating the machine's decisions between indefinite and definite article, the confusion matrix for the model, trained on the balanced Brown set as seen in Table 37, will be used for reference. Although great improvement was seen in predicting

Act\Pred	**definite**	**indefinite**	∅
definite	**30**	1	6
indefinite	12	**13**	2
∅	8	14	**81**

Table 37: Confusion Matrix for balanced Brown evaluated on 2006 Texts

indefinite articles, the wrong instances are again quite telling. In (5.14), the model assigned a definite article when there should have been an indefinite article. The phrase is not ungrammatical with a definite article, but the context does not require definiteness for the compound *sport team*. The author does not write about a specific sport team, but rather about generic sport teams. The exact opposite is the case in example (5.15). The analogy used is the topic of the entire text snippet, therefore it needs to be definite, although grammatically the indefinite article is correct as well.

(5.14) ... ties to *a* [the] local sport team ...

(5.15) *The* [a] psychological analogy has...

(5.16) ... the study is our geographic [a] setting.

In the last example (5.16), the phrase *our geographical setting* does not need an article, as the possessive pronoun *our* takes the position of the article. Inserting the indefinite article makes the phrase incorrect. The features used during the training do not take into consideration if another Part-of-Speech takes on the function of the article, therefore the models treat all non-articles as non-existent. The same phenomenon was seen in the second phrase in (5.12).

The second data set which will be evaluated in greater detail is the AmE06, containing the training data published in 2006. The procedure during the evaluation was the same as with the Brown data set. The model trained on the regular data set returned an accuracy of 77.77%, while the balanced data resulted in a 60.40% accuracy. For the normal data set, the accuracy is higher for the 2006 evaluation texts, which was expected. However, its performance is noticeably lower than what could be expected, given a training accuracy of over 77%. The balanced data does not conform to initial expectations, as the accuracies are higher than during the training, which is usually not the case. Furthermore, the model performed better on the evaluation texts from 1961 than on the texts from the same time period as the

AmE06	1961 Texts	2006 Texts
normal	69.93%	71.25%
balanced	67.97%	63.47%

Table 38: Accuracies from the Evaluation of the Models trained on AmE06

training. Again, this is likely due to patterns in the evaluation texts fitting more accurately to the patterns in the training data.

Taking a closer look at the confusion matrices, no indefinite articles were predicted with the imbalanced model. Table 39 shows a similar picture to what was seen with the Brown data. The vast majority of definite articles are predicted correctly, and the same can be observed with the ∅-forms. Similarly, the majority of indefinite articles were mistaken for definite articles. Example (5.17) is not necessarily grammatically

Act\Pred	definite	indefinite	∅
definite	**34**	0	3
indefinite	25	**0**	2
∅	18	0	**85**

Table 39: Confusion Matrix for AmE06 evaluated on 2006 Texts

wrong with the ∅-form, however, given the context the definite article makes more sense, as a specific person at a specific point in time is referred to. The same noun phrases as in (5.13) were the other two inaccurately predicted cases. It appears that both models have learned similar patterns that inform these incorrect choices.

(5.17) ... whether *the* [∅] job applicants standing before them ...

(5.18) ... to work in *the* [the] United [∅] States or not.

A very interesting problem can be illustrated with the phrase (5.18). The original article is dependent on the head noun *States*, the model 'sees' the phrase as *United the States*. The model deletes this definite article by inserting a ∅-form. At the same time, the algorithm predicts *United* to have a definite article, which gives the appearance that the phrase is correct, however, the article is accompanying the wrong noun. This example will be discussed in more detail in Chapter 6. The compound appears several times, and both approaches have difficulties in predicting it correctly.

The confusion matrix for the balanced model's evaluation on the 2006 text snippets

show the effects of the balancing very well. The indefinite articles are predicted, though there were a number of so-called false positives. This means that the algorithm predicted *indefinite* even though the correct choice would have been definite or ∅-form. Moreover, the patterns for definite articles and ∅-form seem to have lost some of their distinctness, thus leading to more mistakes in predicting the two other forms. (5.19) is a list of possible reasons for taking drugs like Ritalin. The

Act\Pred	**definite**	**indefinite**	∅
definite	**25**	6	6
indefinite	14	**11**	2
∅	17	16	**70**

Table 40: Confusion Matrix for balanced AmE06 evaluated on 2006 Texts

indefinite article is not grammatically wrong, however, convention forbids articles in such a list. The second phrase, (5.20), is similar in that the indefinite article is not grammatically wrong, though in the context the definite article is required. The *state* has been referenced before, further *the state of being in love* has been contrasted to other states of being.

(5.19) ... (i.e., [a] better concentration, [an] increased alertness) ...

(5.20) ... *the* [a] dynamic state of being in love.

(5.21) ... had drawn the blood from *an* [the] arm vein ...

The same principle holds true for phrase (5.21); it is not important which arm vein the blood was drawn from, but that it did not originate from the neck. This becomes clear when read in context. The context is vital in the decision between definite and indefinite article, something which was demonstrated in the rule-based approach analysis as well, and will be revisited in Chapter 6.

The last data set is the combination of the two data sets. The regular data returned an accuracy of 77.31% and the balanced model performed with an accuracy of 65.45%. In Table 41, the accuracies for the evaluation on the texts form 1961 and 2006 can be found. Many of the confusion matrices and types of mistakes are

Both	**1961 Texts**	**2006 Texts**
normal	71.89%	71.25%
balanced	60.78%	65.86%

Table 41: Accuracies from the Evaluation of the Models trained on Both

very similar to whatbeen discussed with the Brown and AmE06 data sets. For the imbalanced data, the confusion matrix in Table 42 contains the exact same values as was seen in Table 39 for the AmE06 data set.

Act\Pred	**definite**	**indefinite**	∅
definite	**34**	0	6
indefinite	25	**0**	2
∅	18	0	**85**

Table 42: Confusion Matrix for Both evaluated on 2006 Texts

The confusion Matrix for the balanced data set, in Table 43, also resembles other matrices seen so far as well. Once again, indefinite articles are predicted reliably at the expense of the correct predictions of definite articles. While the imbalanced model predicted 34 out of 40 definite articles correctly, the model trained on the balanced data only predicted ten instances correctly. This illustrates once more that the decision between definite and indefinite article is heavily dependent on the context.

Act\Pred	**definite**	**indefinite**	∅
definite	**10**	19	8
indefinite	3	**23**	1
∅	2	24	**77**

Table 43: Confusion Matrix for balanced Both evaluated on 2006 Texts

(5.22) The study had *a* [the] large sample but with *a* [the] low response rate of 32% ...

(5.23) ... because *the* [a] model being proposed ...

The model predicted two definite articles instead of the two indefinite articles in (5.22), which again is not grammatically wrong, however the reader assumes that there must be a smaller sample and a higher response rate, as it was specified which sample was used. The phrase in (5.23) illustrates the inverse problem, as the context demands the definite article, and the model was described in the previous sentence.

In conclusion, even though the evaluation of the machine learning-based article correction did not work as expected, the results are nevertheless promising. The enormous impact of training data on the quality of results was strongly reiterated, although the assumption that great similarities between the training material and the target data will result in greater accuracy has not been confirmed.

5.9 Machine Learning Approach Conclusions

For the second system of automatic language correction, the techniques of machine learning were applied. Three algorithms, Naïve Bayes, SVM, and Logistic Regression were used on three data sets: Brown academic, AmE06 academic, and a combination of the two academic corpora. The process of finding productive features spanned over six training cycles and ultimately resulted in 11 features. These features attempt to describe factors which influence the article usage of a noun. This includes the type of noun and the presence of any modifiers. The accuracies which the algorithms reach are higher on the more recent data as compared to the data published in 1961. The evaluation has shown that the training data as well as the data used to evaluate play a vital role in the performance of the models. This was especially true for indefinite articles, as the imbalanced models failed to predict them entirely due to their low frequency. The problem of the imbalanced data might be solved by providing more contextual information, leading to more distinct patterns guiding the use of indefinite articles.

Similar to the rule-based approach, additional semantic information will most likely improve the results further. Adding features like Named Entity Recognition (NER) information for proper nouns, or information about the locational phrases, will probably increase accuracies. Moreover, it makes sense to train different models on the individual steps in the process of choosing the correct article. Similar to Gamon et al. (2008), one could train a classifier to decide whether an article is necessary or not, and a second step would then choose the correct article if required. Proper nouns have been difficult to deal with; therefore, one option would be to train a model with the sole purpose of correctly handling article usage with proper nouns. Another possible feature would be real co-reference resolution. While the feature *resolution* adds a certain amount of semantic information, co-reference resolution could provide crucial information when deciding between indefinite and definite articles, as was already mentioned when evaluating the rule-based approach. One danger in continually adding features to describe the context of article usage best is over-fitting. If features are made to perfectly describe one set of data, the same features may perform very poorly on new data. Moreover, as was already seen several times, the simplest ideas usually work best.

To conclude, it can be stated that the performance of the machine learning systems has surpassed expectations during the training cycles and shown how complex the influencing factors for performance are in the application to new, unseen data. Additionally, interesting insights have been gained into the next steps of development

for this approach. In the next chapter, both systems will be compared to elicit strengths and weaknesses, as well as possible combinations of the approaches to increase accuracy further.

6 Discussion

So far the two article correction systems have been evaluated separately. In this section, their merits and disadvantages are directly contrasted. Moreover, the type of mistakes which are common to both systems will be elaborated on and suggestions on how to avoid such issues will be presented. As was mentioned in the introduction, the juxtaposition of linguistic knowledge and massive amounts of data is a fascinating one, as the sheer amount of data usually outperforms linguistic knowledge in many natural language processing tasks. This is partially true for the two systems engineered in this thesis, though the results garnered through the rule-based approach are somewhat more subjective. Nevertheless, initial results show the machine learning system is significantly better in providing clear and useful corrections as it always provides a singular answer. However, the systems are closer to each other in performance than expected. In Table 44 the percentage of erroneous corrections are listed. For the machine learning system, the regular, imbalanced data sets were used.

	1961 Texts	**2006 Texts**
Rule-Based	17.7%	16.4%
ML-Brown-LogReg	31.4%	31.1%
ML-AmE06-LogReg	30.1%	28.8%
ML-Both-LogReg	28.1%	28.8%

Table 44: Percentage of Wrong Corrections

Though the rule-based system has relatively low accuracies compared to the machine learning approach, there are cases where the program suggested two options and one of them would be grammatically incorrect. Therefore, in actuality the percentage of erroneous corrections should be higher. Newer texts appear to be easier to correct than the texts from 1961. It was not possible to determine whether this conclusion is generalizable, or whether the specific texts chosen for the evaluation had some influence. Either case is possible, as it is difficult to extrapolate broadly from such small data sets. It can additionally be stated that the types of mistakes made by both programs are very similar. Indefinite articles pose a challenging problem

for both tools, as do proper nouns. Moreover, the choice between definite article and ∅-form is more difficult than expected. These three cases will be discussed in more detail and data for both article correction systems will be used as illustrative examples.

All systems had difficulties in dealing with the phrase *The decision of the Supreme Court of the United States ...* from the fifth text snippet from 1961. In the examples (6.1) to (6.4) the outputs for this phrase are listed. The original article is in italics, while the correction from the systems are in square brackets.

(6.1) *the* [a/the] decision of *the* [∅] Supreme [∅] Court of *the* [∅] United [the/∅] States ... – Rule-Based

(6.2) *the* [the] decision of *the* [∅] Supreme [∅] Court of *the* [∅] United [∅] States ... – ML-Brown

(6.3) *the* [the] decision of *the* [∅] Supreme [∅] Court of *the* [∅] United [the] States ... – ML-AmE06

(6.4) *the* [the] decision of *the* [∅] Supreme [∅] Court of *the* [∅] United [the] States ... – ML-Both

It becomes apparent that all systems prefer the ∅-form with proper nouns. For the rule-based system, of course, this was an explicit instruction. However, the hope was that the machine learning approach would be able to learn the nuanced differences between some proper nouns. As was argued in the conclusions for both rule-based and machine learning systems, Named Entity Recognition (NER) could help improve outputs in cases like this. *Supreme Court* should be recognized as an `organization` while *United States* should be tagged either as an `organization` or a `location`. Provided with this information, the systems should be able to make more informed decisions concerning article usage. An additional obstacle is that the systems require further guidance in dealing with compounds. Although it is registered if a noun is modified, the modifying element is not aware of the fact that it 'belongs' to some other token. Therefore, *United States* is assigned two articles, even though it is one entity. To illustrate this problem it makes sense to have a look at the parse of this phrase in Figure 7.

The article *the* and *United* are dependent on *States*, however the computer does not realize that *United* is thus within the scope of the article and so it assigns an additional one. A further complication is the placement of the article; as in the example phrases above, the article was always placed directly in front of the given noun, which is not necessarily correct. The same problem is illustrated with

15	The	the	DT	DT	_	16	det
16	decision	decision	NN	NN	_	27	ccomp
17	of	of	IN	IN	_	16	prep
18	the	the	DT	DT	_	19	det
19	Supreme	Supreme	NP	NP	_	17	pobj
20	Court	Court	NP	NP	_	19	partmod
21	of	of	IN	IN	_	20	prep
22	**the**	the	DT	DT	_	**24**	det
23	United	United	NP	NP	_	**24**	amod
24	States	States	NPS	NPS	_	21	pobj

Figure 7: Parse of the Phrase *The decision of the Supreme Court of the United States*

example (5.18) in section 5.8. A final major difficulty, which is related to faulty tagging and parsing processes, is the fact that here presumably correct Standard English was processed. In actuality, when correcting ESL writing, the syntax will not be as standard, and therefore pose a much more difficult problem for the natural language processing tools. This issue was already touched upon in Chapter 4, during the evaluation of the rule-based approach. This context must be relevant to future designs and developments for these tools, to ensure the best results possible for ESL students.

The choice between indefinite article and definite article is largely dependent on semantics and extralinguistic context. Consequently, it was expected that this task would prove to be a difficult one for all systems. The importance of context has been stressed several times, for example in the phrase *draw blood from an arm vein and*. This was already mentioned in example (5.21) and will be revisited below. The second phrase which will be analyzed in more detail is *creates a force which compresses*. In both phrases the context demands the indefinite article. As has been mentioned before, it is not important which arm vein the blood was drawn from, but that it was not from a neck vein. Moreover, in the second phrase, it is not a specified force but a general one, therefore the indefinite article is needed. Table 45 lists the two phrases and the four systems used to correct them. For the machine learning tool, the balanced models were used.

	RB	**Brown**	**AmE06**	**Both**
. . . an arm vein . . .	the/∅	the	the	the
. . . creates a force which . . .	the/∅	the	the	the

Table 45: Two indefinite Phrases and the Predictions by all systems

It quickly becomes apparent that none of the corrections are accurate. All machine

learning models suggest the definite article, while the rule-based system suggests the definite or ∅-form. This is the case, as *force* is a singular countable noun, and it is not followed by the preposition *of* nor is it part of any of the specified constructions. For the machine learning, the problem likely lies in the fact that too few contextual features are given, and so the models cannot learn the necessary patterns even if the data is balanced. Table 46 shows two phrases where the definite article or the ∅-from is correct though the systems suggested an indefinite article. The full phrase for the first example is *by the linear compressing action between the rollers and* and the second phrase reads *Information regarding the causal direction.* Both contexts do not allow for an indefinite article, as the type of *compressing action* is specified, and in the second example *information* is an uncountable noun.

	RB	Brown	AmE06	Both
. . . the linear compression action . . .	the/∅	a	a	the
Information regarding the causal direction . . .	an/the	an	∅	the

Table 46: Two definite Phrases and the Predictions by all systems

The correction systems are far less unanimous in their suggestions than in the previous example. The rule-based suggestion in the first phrase stems from the same reasoning as was outlined above; action is a singular countable noun and therefore the two solutions are given. Brown and AmE06 predict an indefinite article, while the combined data set chose the correct definite article. The casual relationship behind these differing predictions is unclear, though it is most probably related to the instances used during the training. In the second phrase, *information* is an uncountable noun, and so usually does not take an article. However, the rule-based system treats it as a countable singular noun, as the list of predefined uncountable nouns is not exhaustive. The models trained on Brown and the combined systems did not take into account that the noun is uncountable. This information was not explicitly taught to the models through features, but it was hypothesized the system would learn it through pattern recognition. Though arguably the definite article, suggested by the Both model, could technically be grammatically correct, it nevertheless reads somewhat awkwardly given the context. A possibility to add semantic context would be to add real co-reference resolution. As explained previously, the purpose of the resolution would be to add enough information to know if the concept in question has already been introduced. Other semantic analysis could also help provide needed context, for example, an automatic extraction of the discourse topic. However, additional features are liable to result in an excess of information, which will ultimately lead to more noise than accurate predictions.

The last article choice which will be looked at in more detail is between the definite article and the ∅-form. Before this thesis, it was hypothesized that differentiating between these choices would be relatively straightforward, as it depends less on context than the binary of definite/indefinite. However, as the rule-based approach has illustrated nicely, the decision is far more complex than was first assumed. In Table 47, two phrases for ∅-forms mistaken for definite forms, and definite articles mistaken for ∅-forms, respectively, are listed. In the first phrase *a desire for secondary gain drive this deception*, the rule-based system provides two options, as was expected. The definite article predicted by the the Brown and AmE06 model sounds very strange. Moreover, the expression 'desire for something' does not require an article. The second phrase, *the Court corrects its own error*, is interesting as the position of the article is taken by a possessive pronoun. A similar case was discussed with example (5.16). The systems do not recognize that *its* takes the position of the article, and therefore no other element is needed. Nonetheless, the Brown and AmE06 models correctly predict the ∅-form, while the Both model and the rule-based system suggest the definite article. The root of the rule-based model's error is very simple: *error* is a singular countable noun, therefore the definite article or the ∅-form is suggested. In regards to the machine learning models, it is nearly impossible to determine exactly which feature is responsible for the final suggestion. Other parts-of-speech that can take the function of articles will have to be incorporated in further systems.

	RB	Brown	AmE06	Both
... for secondary gain drive...	the/∅	the	the	∅
... its own error.	the/∅	∅	∅	the
... the two rollers...	the/∅	∅	∅	∅
...: the basic elements...	the/∅	the	the	∅

Table 47: Phrases with Erroneous Suggestions for definite/∅

The phrase *and the juncture of the two rollers* is predicted wrong by all machine learning systems and the rule-based system would also allow for the wrong ∅-form. While grammatically both are correct, it is once more the context that renders the ∅-form less optimal. The *two rollers* have been described in the previous sentences, therefore, it would be strange to use the ∅-form, which implies a certain indefiniteness. The last phrase *the basic elements are* demonstrates again that reasons leading to this prediction are complex. The ∅-form is correct, however, it makes the statement weaker, suggesting that not all *basic elements* were listed in the following phrase. This nuance is almost impossible to grasp for both the rule-based and machine learning-based algorithms. The classifier used here were trained for all

decisions about articles. It might be fruitful to train classifiers solely on the decision between article and ∅-form. It has been seen in this discussion that the indefinite articles produce noisy data and complicate the decision between ∅-form and definite article. Another classifier could then be used to better inform which article is needed, after the first classifier has concluded that an article is indeed necessary.

The discussion of the three most prominent sources of errors has shown that the rule-based system is competent at making basic decisions, while the machine learning approach is able to make better predictions in more nuanced cases. Regardless, training more classifiers on how to respond to specific cases, as was suggested in the conclusion to Chapter 5, is a promising first step in the further development of these systems. One possibility would be to make a basic triage, based on set rules, and then use specific classifiers to predict single options in unclear cases. An exploration of more fixed constructions for articles and nouns, for example through a Construction Grammar approach, could also prove to be fruitful.

7 Conclusion

In this thesis, two systems designed to automatically correct article usage in academic texts were successfully engineered. Using pre-determined, manually chosen rules to guide the algorithm's choices proved that, with a sufficient amount of time and language expertise, respectable results can be achieved. Extrapolating from this conclusion, a Construction Grammar approach, following for example Hilpert (2014), is likely to produce a number of effective rules for article correction. A significant challenge to the rule-based approach has been an over-reliance on rules which result in multiple answers as potentially correct. In the majority of cases, the rules allowed for either the definite article *the* or the ∅-form. While in many cases one of these suggestions was correct, one almost always fit the context better. With the machine learning approach, it was hypothesized that the sheer volume of data used to train the program would help simulate context, and therefore, encourage the algorithm to make better predictions in instances where the context is pivotal for a correct choice. In order for the machine learning to work effectively, suitable features and great training data are needed. Over the course of this project useful basic features were extracted. Using academic texts written by native speakers ensured the training data was also of high quality. However, the evaluation using unseen data exemplified how volatile results can be based on how similar the tested data is to what the model encountered during training. A major hurdle was the unequal distribution of the labels in the training data. It was a significant challenge for the models to deduce distinct patterns for the comparatively infrequent indefinite article than it is for the other two article forms.

Both approaches encountered similar difficulties. Context, of course, plays a major role in the decision between definite and indefinite articles as well as between definite article and ∅-form. The semantics involved in such nuanced decisions were not sufficiently captured in the set of rules nor in the features describing the article usage for the algorithm. For example, proper nouns have a special role in this discussion, as their article usage depends heavily on contextual information. Several suggestions to improve the prediction for proper nouns have been made, among them the application of Named Entity Recognition tools. Further suggestions of

improving the automatic article correction systems include to break the correction process down into smaller decisions. Specific classifiers or rules for the decision between definite and indefinite article could be trained or written, respectively, once it has been established that an article is necessary. Here, a hybrid system might be interesting to implement; rule-based techniques would be used to make basic decisions and categorizations, followed by machine learning tools to make the more nuanced predictions.

The aim of the project was to have two functioning systems automatically correct articles. This has been successfully achieved, as elaborated above. The second aim was to deepen my understanding of machine learning techniques and writing a grammar for language correction. Employing different workbenches for machine learning has confirmed that the data and algorithms one works with are crucial to the results. Moreover, the implementation of the algorithm can have effects on the quality of predictions as well. Furthermore, understanding the internal calculations of the workbench is crucial to be able to correctly interpret the output. For both approaches, it was interesting to see how very simple rules and features lead to good results, while more complex ideas often introduced more noise than necessary. Moreover, the juxtaposition between expert knowledge and big data has proved to be far less significant than expected in the fuzzy and subjective world of article usage.

References

Bayes, M. and Price, M. (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philosophical Transactions (1683-1775)*, pages 370–418.

Behera, B. and Bhattacharyya, P. (2013). Automated grammar correction using hierarchical phrase-based statistical machine translation. In *IJCNLP*, pages 937–941.

Berezowski, L. (2009). *The myth of the zero article*. Bloomsbury Publishing.

Berry, R. (2013). *English grammar: a resource book for students*. Routledge.

Bhaskar, P., Ghosh, A., Pal, S., and Bandyopadhyay, S. (2011). May i check the english of your paper!!! In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 250–253. Association for Computational Linguistics.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). *Longman grammar of spoken and written English*, volume 2. MIT Press.

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M. A., and Schasberger, B. (1995). Bracketing guidelines for treebank ii style penn treebank project. *University of Pennsylvania*, 97:100.

Christophersen, P. (1939). *The articles: A study of their theory and use in English*. Eubar Munksgaard.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 133–140. Association for Computational Linguistics.

Dahlmeier, D. and Ng, H. T. (2011). Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 915–923. Association for Computational Linguistics.

Davis, M. (2008). The corpus of contemparary american english: 520 million words, 1990-present. available online at http://corpus.byu.edu/coca/.

Davis, M. (2010). The corpus of contemparary american english: 400 million words, 1810-2009. available online at http://corpus.byu.edu/coha/.

Doran, R. M. (2006). The starting point of systematic theology. *Theological Studies*, 67(4):750–776.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available from: `http://wals.info/`.

Foth, K. A. (2007). *Hybrid Methods of Natural Language Analysis*. Shaker Aachen,, Germany.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Belenko, D., and Vanderwende, L. (2008). Using contextual speller techniques and language modeling for esl error correction. In *IJCNLP*, volume 8, pages 449–456.

Grammarly, Inc. (2016). Grammarly.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Han, N.-R., Chodorow, M., and Leacock, C. (2004). Detecting errors in english article usage with a maximum entropy classifier trained on a large, diverse corpus. In *LREC*.

Heidorn, G. (2000). Intelligent writing assistance. *Handbook of natural language processing*, pages 181–207.

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh University Press.

Hudson, R. (2010). *An introduction to word grammar*. Cambridge University Press.

Jacoby, T. (2006). Immigration nation. *Foreign Affairs*, pages 50–65.

Kałuża, H. (1981). *The use of articles in contemporary English*, volume 6. Groos.

Kempe, A. (1993). A probabilistic tagger and an analysis of tagging errors. *Rapport technique, Institut für maschinelle sprachverarbeitung, Universität stuttgart*.

Kunchukuttan, A., Shah, R. M., and Bhattacharyya, P. (2013). Iitb system for conll 2013 shared task: A hybrid approach to grammatical error correction. In *CoNLL Shared Task*, pages 82–87.

Lewis, M. Paul, G. F. S. and Fenning, C. D., editors (2016). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 19th edition.

Libermann, M. (2016). "the case of disapearing determiners". `http://languagelog.ldc.upenn.edu/nll/?p=23277`. Accessed: 26. August 2016.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Mayfield, E. and Rosé, C. (2013). Lightside: Open source machine learning for text accessible to non-experts. invited chapter in the handbook of automated essay grading.

Mitkov, R. and Carroll, J. (2012). Parsing.

Mitkov, R. and Mooney, R. J. (2012). Machine learning. Available from: `//www.oxfordhandbooks.com/10.1093/oxfordhb/9780199276349.001.0001/oxfordhb-9780199276349-e-20`.

Murphy, R. (2004). *English grammar in use-With answers*. Cambridge University Press.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*. Citeseer.

Nivre, J. and Scholz, M. (2004). Deterministic dependency parsing of english text. In *Proceedings of the 20th international conference on Computational Linguistics*, page 64. Association for Computational Linguistics.

Oxford English Dictionary (2003a). "algorithm, n.". Available from: `http://www.oed.com/view/Entry/4959?redirectedFrom=algorithm`.

Oxford English Dictionary (2003b). "article, n.". Available from: http://www.oed.com/view/Entry/11179?rskey=tdBp8Z&result=1&isAdvanced=false#eid.

Patenaude, B., Zitsch III, R., and Hirschi, S. D. (2006). Blood–but not bleeding–at a tracheotomy site: a case of munchausen's syndrome. *Ear, Nose and Throat Journal*, 85(10):677–680.

Peltason, J. W. (1961). *Fifty-eight lonely men: Southern federal judges and school desegregation*, volume 74. University of Illinois Press.

Potts, A. and Baker, P. (2012). Does semantic tagging identify cultural change in british and american english? *International journal of corpus linguistics*, 17(3):295–324.

Quirk, R., Crystal, D., and Education, P. (1985). *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press.

Rozovskaya, A. and Roth, D. (2010). Training paradigms for correcting errors in grammar and usage. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 154–162. Association for Computational Linguistics.

Sakaguchi, K., Mizumoto, T., Komachi, M., and Matsumoto, Y. (2012). Joint english spelling error correction and pos tagging for language learners writing. In *COLING*, pages 2357–2374. Citeseer.

Schmid, H. (1995). Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Scholkopf, B. and Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shalizi, C. (2013). *Advanced data analysis from an elementary point of view*. Citeseer.

Sherwin, C. W. and Fano, U. (1961). Basic concepts of physics. *American Journal of Physics*, 30(5):392–393.

Shillington, A. M., Reed, M. B., Lange, J. E., Clapp, J. D., and Henry, S. (2006). College undergraduate ritalin abusers in southwestern california: Protective and risk factors. *Journal of Drug Issues*, 36(4):999–1014.

Siepmann, D. (2008). *Writing in English: A Guide for Advanced Learners*. UTB. Schlüsselkompetenzen, Sprache und Literatur. Francke. Available from: `https://books.google.ch/books?id=SgcFbUmgr64C`.

Solinger, J. (1961). Apparel manufacturing analysis.

Sweet, H. (1898). *A NEW ENGLISH GRAMMAR LOGICAL AND HISTORICAL: PART 2: SYNTAX.* Oxford: Clarendon Press.

Symonds, P. M. and Jensen, A. R. (1961). From adolescent to adult.

Tajiri, T., Komachi, M., and Matsumoto, Y. (2012). Tense and aspect error correction for esl learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 198–202. Association for Computational Linguistics.

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232. accessed online 15.07.16.

Wann, D. L. (2006). Examining the potential causal relationship between sport team identification and psychological well-being. *Journal of Sport Behavior*, 29(1):79.

Widdows, D. (2004). *Geometry and meaning*, volume 773. CSLI publications Stanford.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc.

Xiang, Y., Zhang, Y., Wang, X., Wei, C., Zheng, W., Zhou, X., Hu, Y., and Qin, Y. (2013). Grammatical error correction using feature selection and confidence tuning. In *IJCNLP*, pages 1067–1071.

Yamada, H. and Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In *Proceedings of IWPT*, volume 3, pages 195–206.

Yotsukura, S. (1970). *The articles in English: a structural analysis of usage*, volume 49. Mouton.

A Tables

A.1 Confusion Matrices for First Training Cycle

Data Brown
Features Noun, Noun Tag, Modifier Tag, Modifier Relation, Preposition
Categories definite, indefinite, ∅

Act\Pred	definite	indefinite	∅
definite	**9384**	794	3189
indefinite	2837	**1026**	948
∅	5967	604	**13070**

Table 48: Confusion Matrix for Naïve Bayse, First Training Cycle

Act\Pred	definite	indefinite	∅
definite	**7683**	785	4899
indefinite	2210	**979**	1622
∅	4022	601	**15018**

Table 49: Confusion Matrix for SVM, First Training Cycle

Act\Pred	definite	indefinite	∅
definite	**7866**	518	4983
indefinite	2430	**776**	1605
∅	4007	356	**15278**

Table 50: Confusion Matrix for Logistic Regression, First Training Cycle

A.2 Confusion Matrices for Second Training Cycle

Data Brown
Features Noun, Noun Tag, Modifier Tag, Modifier Relation, Preposition, Trigram-ONE, Trigram-TWO, Trigram-THREE
Categories definite, indefinite, ∅

Act\Pred	definite	indefinite	∅
definite	**10014**	2061	1292
indefinite	2302	**2189**	320
∅	4945	1336	**13360**

Table 51: Confusion Matrix for Naïve Bayse, Second Training Cycle

Act\Pred	definite	indefinite	∅
definite	**9487**	1134	2746
indefinite	2231	**1555**	1025
∅	3373	690	**15578**

Table 52: Confusion Matrix for SVM, Second Training Cycle

Act\Pred	definite	indefinite	∅
definite	**9775**	793	2799
indefinite	2464	**1282**	1065
∅	3337	422	**15882**

Table 53: Confusion Matrix for Logistic Regression, Second Training Cycle

A.3 Confusion Matrices for Third Training Cycle

Data Brown

Features Noun, Noun Tag, Modifier Tag, Modifier Relation, Preposition, Trigram-ONE, Trigram-TWO, Trigram-THREE, Resolution, *LightSide* UNIGRAM

Categories definite, indefinite, ∅

Act\Pred	**definite**	**indefinite**	∅
definite	**9481**	1789	2097
indefinite	2098	**2122**	691
∅	4120	1313	**14208**

Table 54: Confusion Matrix for Naïve Bayes, Third Training Cycle

Act\Pred	**definite**	**indefinite**	∅
definite	**9258**	1406	2703
indefinite	1725	**2100**	986
∅	2854	851	**15936**

Table 55: Confusion Matrix for SVM, Third Training Cycle

Act\Pred	**definite**	**indefinite**	∅
definite	**9750**	1033	2584
indefinite	1909	**1923**	979
∅	2731	606	**16304**

Table 56: Confusion Matrix for Logistic Regression, Third Training Cycle

A.4 Confusion Matrices for Forth Training Cycle

A.4.1 Brown

Data Brown

Features Noun, Noun Tag, Modifier Tag, Modifier Relation, Preposition, Trigram-ONE, Trigram-TWO, Trigram-THREE, Resolution, *LightSide* UNIGRAM, Vowel

Categories definite, indefinite-consonant, indefinite-vowel, ∅

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**9737**	1399	143	2088
indefinite-consonant	1898	**1528**	39	510
indefinite-vowel	489	181	**55**	111
∅	4235	1064	156	**14186**

Table 57: Confusion Matrix for Naïve Bayes, Forth Training Cycle

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**9326**	1110	230	2701
indefinite-consonant	1563	**1528**	133	855
indefinite-vowel	332	161	**179**	164
∅	2861	705	131	**15944**

Table 58: Confusion Matrix for SVM, Forth Training Cycle

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**9886**	768	93	2620
indefinite-consonant	1655	**1395**	60	866
indefinite-vowel	394	146	**103**	193
∅	2782	471	49	**16339**

Table 59: Confusion Matrix for Logistic Regression, Forth Training Cycle

A.4.2 AmE06

Data AmE06

Features Noun, Noun Tag, Modifier Tag, Modifier Relation, Preposition, Trigram-ONE, Trigram-TWO, Trigram-THREE, Resolution, *LightSide* UNIGRAM, Vowel

Categories definite, indefinite-consonant, indefinite-vowel ∅

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**8218**	1347	117	1966
indefinite-consonant	1737	**1528**	41	509
indefinite-vowel	435	217	**77**	118
∅	4521	1598	284	**20603**

Table 60: Confusion Matrix for Naïve Bayes, Forth Training Cycle

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**7778**	954	184	2732
indefinite-consonant	1189	**1648**	125	976
indefinite-vowel	152	161	**222**	231
∅	2682	776	155	**23393**

Table 61: Confusion Matrix for SVM, Forth Training Cycle

Act\Pred	definite	indefinite-consonant	indefinite-vowel	∅
definite	**7976**	730	86	2856
indefinite-consonant	1313	**1478**	67	1080
indefinite-vowel	294	140	**147**	266
∅	2330	487	71	**24118**

Table 62: Confusion Matrix for Logistic Regression, Forth Training Cycle

A.5 Confusion Matrices for Fifth Training Cycle

Data AmE06

Features Noun, Noun Tag, Modifier Tag, Modifier Relation, Preposition, Trigram-ONE, Trigram-TWO, Trigram-THREE, Resolution, *LightSide* UNIGRAM, Vowel

Categories definite, indefinite, ∅

Act\Pred	**definite**	**indefinite**	∅
definite	**7945**	1752	1951
indefinite	1756	**2421**	608
∅	4354	1966	**20686**

Table 63: Confusion Matrix for Naïve Bayse, Fifth Training Cycle

Act\Pred	**definite**	**indefinite**	∅
definite	**7721**	1200	2727
indefinite	1385	**2235**	1165
∅	2659	970	**23377**

Table 64: Confusion Matrix for SVM, Fifth Training Cycle

Act\Pred	**definite**	**indefinite**	∅
definite	**7878**	949	2821
indefinite	1525	**3023**	1237
∅	2298	645	**24063**

Table 65: Confusion Matrix for Logistic Regression, Fifth Training Cycle

A.6 Confusion Matrices Sixth Training Cycle

Data Brown and AmE06

Features Noun, Noun Tag, Modifier Tag, Modifier Relation, Preposition, Trigram-ONE, Trigram-TWO, Trigram-THREE, Resolution, *LightSide* UNIGRAM, Vowel

Categories definite, indefinite, ∅

Act\Pred	definite	indefinite	∅
definite	**7878**	949	2097
indefinite	2098	**2122**	691
∅	4120	1313	**14208**

Table 66: Confusion Matrix for Naïve Bayse, Sixth Training Cycle

Act\Pred	definite	indefinite	∅
definite	**7721**	1200	2727
indefinite	1385	**2235**	1165
∅	2659	970	**23377**

Table 67: Confusion Matrix for SVM, Sixth Training Cycle

Act\Pred	definite	indefinite	∅
definite	**17763**	1991	5261
indefinite	3470	**3981**	2145
∅	5058	645	**40340**

Table 68: Confusion Matrix for Logistic Regression, Sixth Training Cycle

B Evaluation Details

B.1 Evaluation Texts

Text Nr. 1 1961 – original

Characterizing an age group in this way prevents one from attending to all the variations in the personalities of adolescents and the factors that are responsible for these differences. It is much more helpful to think of personality as a form of adjustment that has been acquired from responding to the vicissitudes of living. So the personality of the adolescent is an outgrowth of experiences of infancy and childhood, and it is not so much a matter of growing out of adolescence as it is reacting to life's experiences with the personality equipment and reaction tendencies with which the earlier years have equipped one. It is possible, of course, that the conditions of the 1940 study did not make it possible to observe the chameleon-like changes that Anna Freud describes. One would have to live with an adolescent day in and day out to be able to observe the shifts in mood that she describes.
(Symonds and Jensen, 1961).

Text Nr. 1 1961 – jumbled

Characterizing age group in this way prevents one from attending to all the variations in the personalities of the adolescents and factors that are responsible for these differences . It is much more helpful to think of personality as form of adjustment that has been acquired from responding to vicissitudes of living . So personality of adolescent is outgrowth of experiences of infancy and childhood , and it is not so much matter of growing out of the adolescence as it is reacting to life 's experiences with personality equipment and reaction tendencies with which earlier years have equipped one . It is possible , of course , that conditions of 1940 study did not make it possible to observe the chameleon-like changes that the Anna the Freud describes . One would have to live with adolescent day in and day out to be able to observe shifts in mood that she describes .

Text Nr. 2 1961 – original

Unfortunately, there are some snakes that have fangs connected to poison sacs and are capable of inflicting a poisonous bite, which in some cases may be venomous enough to cause the death of a human being. In spite of whatever benefit they might be otherwise, such snakes are undesirable members of the earth's population and should be destroyed in areas of human population. However, they make up only a small proportion of the total number of snakes in the world, and it is certainly not fair to condemn an entire group of animals just because there are some undesirables in their midst. It is much more sensible to learn // which are not desired and destroy those and preserve the others which are ue. It is easy to learn the poisonous snakes in the United States, for there are only r types that can bite a person, and these have easily recognized characteristics.

Text Nr. 2 1961 – jumbled

Unfortunately , there are some snakes that have fangs connected to poison sacs and are capable of inflicting a poisonous bite , which in some cases may be venomous enough to cause death of human being . In spite of benefit they might be otherwise , such snakes are undesirable members of the earth 's population and should be destroyed in areas of human population . However , they make up only a small proportion of total number of snakes in world , and it is certainly not fair to condemn the entire group of a animals just because there are some undesirables in their midst . It is much more sensible to learn // which are not desired and destroy those and preserve the others which are ue . It is easy to learn poisonous snakes in United States , for there are only r types that can bite person , and these have easily recognized the characteristics .

(no comprehensible source was provided by COHA)

Text Nr. 3 1961 – original

The law of force readily formulated on the basis of these observations states: Like charges repel, and unlike charges attract. Because opposite signs of charge exist, it is possible in many cases to cancel out the direct electrostatic force demonstrated by the pith balls and to leave exposed to easy observation the relativistic phenomena associated with velocity and acceleration. Of the basic forces in nature? gravitational, electric, and nuclear? only the electric exhibits the phenomenon of the two opposite signs of charge, and it is only with respect to electrical phenomena, therefore, that relativistic effects can be easily observed even at small velocities.
(Sherwin and Fano, 1961)

Text Nr. 3 1961 – jumbled

the law of the force readily formulated on the basis of these observations states : Like the charges repel , and unlike a charges attract . Because opposite signs of the charge exist , it is possible in many cases to cancel out the direct electrostatic force demonstrated by the pith balls and to leave exposed to easy observation the relativistic phenomena associated with velocity and a acceleration . Of the basic forces in nature ? gravitational , electric , and nuclear ? only electric exhibits phenomenon of two opposite signs of charge , and it is only with respect to electrical phenomena , therefore , that relativistic effects can be easily observed even at small velocities .

Text Nr. 4 1961 – original

This compression is induced by passing the fabric between two parallel rollers rotating at different speeds. Before the fabric: passes between the rollers, it passes through an acute angle formed by the plate which services the fabric into contacting the upper roller, which is the faster of the two rollers. This, plus the roller speed differential, creates a force which compresses the fabric between the juncture of upper roll and service plate, and the juncture of the two rollers. The fabric is steamed before it is serviced to the rollers and it is dried after the compaction. The steam plasticizes the fabric to make it more susceptible to linear compaction (shrinkage) by the linear compressing action between the rollers and service plate.
(Solinger, 1961)

Text Nr. 4 1961 – jumbled

This compression is induced by passing fabric between two parallel rollers rotating at different speeds . Before fabric : passes between rollers , it passes through acute angle formed by plate which services fabric into contacting upper roller , which is 0 the faster of two rollers . This , plus the roller speed differential , creates force which compresses fabric between juncture of upper roll and service plate , and juncture of the two rollers . fabric is steamed before it is serviced to rollers and it is dried after compaction . steam plasticizes fabric to make it more susceptible to linear compaction (shrinkage) by linear compressing action between a rollers and service plate .

Text Nr. 5 1961 – original

The Little Rock Chamber of Commerce polled its members on February 23: they voted 819 to 245 in favor of reopening the high schools " on a controlled minimum plan of integration acceptable to the Federal Courts. " In large advertisements the Chamber of Commerce told Little Rock citizens: " The decision of the Supreme Court of the United States, however much we dislike it, is the declared law and it is binding upon us. We think the decision was erroneous... but we must in honesty recognize that, because the Supreme Court is the court of last resort in this country, what it has said must stand until there is a correcting constitutional amendment or until the Court corrects its own error. We must live and act now under the decision of that Court. We should not delude ourselves about that.
(Peltason, 1961)

Text Nr. 5 1961 – jumbled

The Little a Rock Chamber of a Commerce polled its members on February 23 : they voted 819 to 245 in favor of reopening high schools " on the controlled minimum plan of the integration acceptable to Federal a Courts . " In large advertisements Chamber of Commerce told Little Rock citizens : " the decision of the Supreme Court of United States , however much we dislike it , is declared law and it is binding upon us . We think a decision was erroneous ... but we must in a honesty recognize that , because the Supreme Court is the court of last resort in this country , what it has said must stand until there is a correcting constitutional amendment or until Court corrects its own error . We must live and act now under decision of that Court . We should not delude ourselves about that .

Text Nr. 1 2006 – original

This study replicates other single campus epidemiological assessments of Ritalin/Adderall use. The study had a large sample but with a low response rate of 32%, these data should be viewed with caution. A major strength of the study is our geographic setting. To date, no studies have examined this issue with a west coast sample of college students. Further studies separating the abuse of Ritalin and Adderall may also yield interesting findings. Additionally, in this study, we did not ask respondents in our sample to provide reasons for using these drugs, and it is possible the context and reasons for the nonmedical use of these substances may differ from region to region. In a sample of Midwestern college students, the top reasons listed by respondents for the use of these drugs were linked to academic motives (i.e., better concentration, increased alertness) and feeling a " high " (Teter et al., 2005).
(Shillington et al., 2006, 16)

Text Nr. 1 2006 – jumbled

This study replicates other single campus epidemiological assessments of Ritalin/Adderall use . the study had large sample but with low response rate of 32 % , these data should be viewed with caution . a major strength of study is our geographic setting . To the date , studies have examined this issue with west coast sample of college students . Further studies separating abuse of Ritalin and Adderall may also yield interesting findings . Additionally , in this study , we did not ask respondents in our sample to provide reasons for using these drugs , and it is possible the context and reasons for the nonmedical use of these substances may differ from region to region . In the sample of Midwestern college students , top reasons listed by respondents for use of these drugs were linked to academic motives (i. e. , better concentration , increased alertness) and feeling 0 a " high " (Teter et al. , 2005)

Text Nr. 2 2006 – original

This is part of the significance of accepting the four-point hypothesis as the special-categorial component of the unified field structure of systematic theology. Obviously, the second and third of these supernatural analogues will be the most prominent, and it is to them that Lonergan refers when he sets forth his later psychological analogy: The psychological analogy. has its starting point in that higher synthesis of intellectual, rational, and moral consciousness that is the dynamic state of being in love. Such love manifests itself in its judgments of value. And the judgments are carried out in decisions that are acts of loving. Such is the analogy found in the creature.
(Doran, 2006, 27)

Text Nr. 2 2006 – jumbled

This is part of significance of accepting four-point hypothesis as special-categorial component of unified field structure of systematic theology . Obviously , a the second and third of these supernatural analogues will be 0 the most prominent , and it is to them that Lonergan refers when he sets forth later psychological analogy : sychological analogy . has its starting point in that higher synthesis of intellectual , rational , and moral consciousness that is the dynamic state of being in love . Such love manifests itself in its the judgments of value . And judgments are carried out in the decisions that are acts of loving . Such is analogy found in creature .

Text Nr. 3 2006 – original

That is, the literature to date supports the notion that strong psychological ties to a local sport team are positively related to psychological health. However, this body of evidence is only tentative support for the Prediction 1 because, as noted by Wann and his associates (Wann, 1994; Wann et al., 1999), the research to date has been correlational in nature. Information regarding the causal direction of the relationship is lacking and it is not possible to determine if higher levels of identification cause better psychological health, vice versa, or if the relationship tends to be bi-directional (i.e., circular). This is an important point, and a vital piece missing in the research to date, because the model being proposed here explicitly hypothesizes a causal pattern in which identification with a local team has a direct and positive effect on one's psychological health.
(Wann, 2006, 17)

Text Nr. 3 2006 – jumbled

That is , literature to the date supports the notion that strong psychological ties to a local sport team are positively related to psychological health . However , this body of the evidence is only tentative support for the Prediction 1 because , as noted by Wann and his associates (Wann , 1994 ; the Wann et al. , 1999) , research to the date has been correlational in nature . the Information regarding causal direction of a relationship is lacking and it is not possible to determine if higher levels of identification cause better psychological health , vice versa , or if the relationship tends to be bi-directional (i. e. , circular) . This is important point , and the vital piece missing in research to date , because the model being proposed here explicitly hypothesizes causal pattern in which a identification with the local team has a direct and positive effect on one 's psychological health .

Text Nr. 4 2006 – original

In actuality, she had drawn the blood from an arm vein and spattered it on her neck. A factitious illness is one in which a patient consciously and deliberately presents with a self-induced injury or a false history in order to mislead a physician into making an erroneous diagnosis and administering some type of treatment. An underlying psychological disorder and a desire for secondary gain drive this deception. (n1) Factitious illness is rare in otolaryngology, and it is different from malingering, which is not uncommon in otolaryngologic practice. Malingerers mislead doctors in order to acquire tangible gains, such as money and narcotics. Factitious illness is also different from a somatoform disorder, in which symptoms are involuntary. (n2) One of the more well-known factitious illnesses is Munchausen's syndrome, in which affected patients repeatedly seek treatment for nonexistent acute illnesses while reporting a dramatic yet plausible history.
(Patenaude et al., 2006, 3)

Text Nr. 4 2006 – jumbled

In the actuality , she had drawn a blood from arm vein and spattered it on her neck . Introduction the factitious illness is one in which patient consciously and deliberately presents with the self-induced injury or false history in the order to mislead physician into making the erroneous diagnosis and administering some type of treatment . underlying psychological disorder and desire for secondary gain drive this deception . (n1) Factitious illness is rare in otolaryngology , and it is different from malingering , which is not uncommon in otolaryngologic practice . Malingerers mislead doctors in the order to acquire tangible gains , such as money and a narcotics . Factitious illness is also different from somatoform disorder , in which the symptoms are involuntary . (n2) One of more well-known factitious illnesses is Munchausen 's syndrome , in which affected patients repeatedly seek the treatment for nonexistent acute illnesses while reporting dramatic yet plausible history .

Text Nr. 5 2006 – original

Fourth, some of the most charged disagreements of the past year were about enforcement issues: whether or not to build a fence, whether to make felons of unauthorized workers or of those who provide them with humanitarian assistance. But in fact, of the three essential elements of comprehensive reform, enforcement is the least controversial, at least among policymakers serious about fixing the system. It is well known what works best on the border: little can be done that is not done already, although it could be augmented by more technology. And it is well known what is needed in the workplace: a national, mandatory, electronic employment-verification system that informs employers in a timely way whether the job applicants standing before them are authorized to work in the United States or not. Such a system need not be Orwellian: the basic elements are biometric identity cards and a computer database. And the process should operate much like ordinary credit card verification but be backed up by significantly stepped-up sanctions against employers who fail to use the system or who abuse it.

(Jacoby, 2006, 16)

Text Nr. 5 2006 – jumbled

Fourth , some of the most charged disagreements of past year were about enforcement issues : whether or not to build fence , whether to make a felons of unauthorized workers or of those who provide them with humanitarian assistance . But in a fact , of three essential elements of comprehensive reform , the enforcement is the least controversial , at least among policymakers serious about fixing system . It is well known what works best on border : little can be done that is not done already , although it could be augmented by more technology . And it is well known what is needed in workplace : national , mandatory , electronic employment-verification system that informs employers in timely way whether job applicants standing before them are authorized to work in United States or not . Such system need not be Orwellian : the basic elements are biometric identity cards and computer database . And the process should operate much like ordinary credit card verification but be backed up by significantly stepped-up sanctions against the employers who fail to use system or who abuse it .

B.2 Detailed Evaluation Results – Rule-Based Approach

Table 69 and 70 list every combination of articles found in the evaluation data. All the determiners have not been listed, as the rules do not touch them.

	original	jumbled	correction	# of instances
correct	∅	∅	∅	10
correct	∅	a	∅	2
correct	∅	the	∅	3
wrong	∅	the	a/the	2
wrong	∅	the	an/the	2
wrong	∅	a	an/the	1
partially correct	∅	∅	the/∅	45
partially correct	∅	a	the/∅	4
partially correct	∅	the	the/∅	3
partially correct	a	a	a/the	1
wrong	a	a	an/the	1
wrong	a	a	the/∅	1
partially correct	a	the	a/the	1
wrong	a	∅	∅	1
wrong	a	∅	the/∅	4
wrong	an	∅	the/∅	3
wrong	an	the	a/the	1
wrong	the	∅	∅	7
partially correct	the	∅	the/∅	30
partially correct	the	a	the/∅	1
partially correct	the	the	the/∅	10
partially correct	the	the	an/the	1
wrong	the	the	∅	2
partially correct	the	the	a/the	5

Table 69: Detailed Evaluation Texts from 1961

	original	**jumbled**	**correction**	**# of instances**
wrong	the	∅	∅	2
partially correct	the	the	a/the	7
partially correct	the	the	the/∅	4
partially correct	the	the	an/the	1
partially correct	the	a	a/the	2
partially correct	the	∅	a/the	24
partially correct	∅	∅	the/∅	56
wrong	∅	a	an/the	1
correct	∅	a	∅	1
partially correct	∅	a	the/∅	2
correct	∅	the	∅	6
wrong	∅	the	a/the	1
wrong	∅	the	an/the	3
partially correct	∅	the	the/∅	5
partially correct	a	the	a/the	4
wrong	a	the	the/∅	1
partially correct	a	the	an/the	1
partially correct	a	a	a/the	2
partially correct	a	a	an/the	1
partially correct	a	∅	a/the	1
wrong	a	∅	∅	17
correct	∅	∅	∅	10

Table 70: Detailed Evaluation Texts from 2006

B.3 Confusion Matrices for the Evaluation of the Machine Learning Approach

Act\Pred	definite	indefinite	∅
definite	**37**	7	15
indefinite	9	**4**	1
∅	8	9	**63**

Act\Pred	definite	indefinite	∅
definite	**25**	6	6
indefinite	14	**11**	2
∅	17	16	**70**

Table 71: Confusion Matrices for AmE06-Balanced & Logistic Regression, re-evaluated on Evaluation Text 1961 (left) and 2006 (right)

Act\Pred	definite	indefinite	∅
definite	**41**	0	18
indefinite	13	**0**	1
∅	14	0	**66**

Act\Pred	definite	indefinite	∅
definite	**34**	0	3
indefinite	25	**0**	2
∅	18	0	**85**

Table 72: Confusion Matrices for AmE06 & Logistic Regression, re-evaluated on Evaluation Text 1961 (left) and 2006 (right)

Act\Pred	definite	indefinite	∅
definite	**31**	7	21
indefinite	11	**3**	0
∅	8	2	**70**

Act\Pred	definite	indefinite	∅
definite	**30**	1	6
indefinite	12	**13**	2
∅	8	14	**81**

Table 73: Confusion Matrices for Brown-Balanced & Logistic Regression, re-evaluated on Evaluation Text 1961 (left) and 2006 (right)

Act\Pred	definite	indefinite	∅
definite	**39**	1	19
indefinite	13	**0**	1
∅	16	0	**64**

Act\Pred	definite	indefinite	∅
definite	**31**	0	6
indefinite	26	**0**	1
∅	19	0	**84**

Table 74: Confusion Matrices for Brown & Logistic Regression, re-evaluated on Evaluation Text 1961 (left) and 2006 (right)

Act\Pred	definite	indefinite	∅
definite	**10**	25	24
indefinite	2	**12**	0
∅	2	7	**71**

Act\Pred	definite	indefinite	∅
definite	**10**	19	8
indefinite	3	**23**	1
∅	2	24	**77**

Table 75: Confusion Matrices for Both-Balanced & Logistic Regression, re-evaluated on Evaluation Text 1961 (left) and 2006 (right)

Act\Pred	definite	indefinite	∅
definite	**46**	0	13
indefinite	13	**0**	1
∅	16	0	**64**

Act\Pred	definite	indefinite	∅
definite	**34**	0	6
indefinite	25	**0**	2
∅	18	0	**85**

Table 76: Confusion Matrices for Both & Logistic Regression, re-evaluated on Evaluation Text 1961 (left) and 2006 (right)

C Miscellaneous

C.1 Equations

Logistic Regression with more than two classes can be formalized as

$$Pr(Y = c|\overrightarrow{X} = x) = \frac{e^{\beta_0^{(c)} + x \cdot \beta^{(c)}}}{\sum_c e^{\beta_0^{(c)} + x \cdot \beta^{(c)}}} \tag{C.1}$$

C.2 Genre Overview Brown Corpus

LABEL	GENRE	# OF TEXTS
A	Reportage	44
B	Editorial	27
C	Reviews	17
D	Religion	17
E	Skills, trades and hobbies	36
F	Popular lore	48
G	Belles letters, biographies, essays	75
H	Miscellaneous	30
J	Science	80
K	General Fiction	29
L	Mystery and Detective Fiction	24
M	Science fiction	6
N	Adventure and Western	29
P	Romance and love story	29
R	Humor	9

D Selbstständigkeitserklärung

Universität Zürich[UZH]

Englisches Seminar

Universität Zürich
Englisches Seminar
Plattenstrasse 47
CH-8032 Zürich
Telefon +41 44 634 35 51
Telefax +41 44 634 49 08
www.es.uzh.ch

Selbstständigkeitserklärung zur wissenschaftlichen Arbeit am Englischen Seminar der Universität Zürich

Originalarbeit
Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten schriftlichen Arbeit mit dem Titel

.....

.....

um eine von mir selbst und ohne unerlaubte Beihilfe sowie *in eigenen Worten* verfasste Originalarbeit handelt.

Sofern es sich dabei um eine Arbeit von mehreren Verfasserinnen oder Verfassern handelt, bestätige ich, dass die entsprechenden Teile der Arbeit korrekt und klar gekennzeichnet und der jeweiligen Autorin oder dem jeweiligen Autor eindeutig zuzuordnen sind.

Ich bestätige überdies, dass die Arbeit als Ganze oder in Teilen weder bereits einmal zur Abgeltung anderer Studienleistungen an der Universität Zürich oder an einer anderen Universität oder Ausbildungseinrichtung eingereicht worden ist noch inskünftig durch mein Zutun als Abgeltung einer weiteren Studienleistung eingereicht werden wird.

Verwendung von Quellen
Ich erkläre ausdrücklich, dass ich *sämtliche* in der oben genannten Arbeit enthaltenen Bezüge auf fremde Quellen (einschliesslich Tabellen, Grafiken u. Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich *ausnahmslos* und nach bestem Wissen sowohl bei wörtlich übernommenen Aussagen (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen anderer Autorinnen oder Autoren (Paraphrasen) die Urheberschaft angegeben habe.

Sanktionen
Ich nehme zur Kenntnis, dass Arbeiten, welche die Grundsätze der Selbstständigkeitserklärung verletzen – insbesondere solche, die Zitate oder Paraphrasen ohne Herkunftsangaben enthalten –, als Plagiat betrachtet werden und die entsprechenden rechtlichen und disziplinarischen Konsequenzen nach sich ziehen können (gemäss §§ 7ff der Disziplinarordnung der Universität Zürich sowie § 36 der Rahmenordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich).

Ich bestätige mit meiner Unterschrift die Richtigkeit dieser Angaben und erlaube es der Kursleitung, die Arbeit mit einer Plagiatssoftware zu prüfen, welche die Arbeit für künftige Vergleiche speichert.

Name: ..... Vorname: .....

Matrikelnummer: .....

Datum: ..... Unterschrift: .....

Seite 1/1 04.04.2014