AN INVESTIGATION OF TENSE, ASPECT AND OTHER VERB GROUP FEATURES FOR ENGLISH PROFICIENCY ASSESSMENT ON DIFFERENT ASIAN LEARNER CORPORA

ALEXANDRA PANAGIOTOPOULOS

A THESIS
IN
THE DEPARTMENT
OF
COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

APRIL 2015

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Alexandra Panagiotopoulos**

Entitled: **An investigation of tense, aspect and other verb group features for English proficiency assessment on different Asian learner corpora**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

_____ Chair

Dr. Rajagopalan Jayakumar

_____ Examiner

Dr. Leila Kosseim

_____ Examiner

Dr. Olga Ormandjieva

_____ Thesis Supervisor

Dr. Sabine Bergler

Approved by _____

Chair of Department or Graduate Program Director

_____ 2015

_____

Dr. Amir Asif, Dean

Faculty of Engineering and Computer Science

Abstract

An investigation of tense, aspect and other verb group features for English proficiency assessment on different Asian learner corpora

Alexandra Panagiotopoulos

Recent interest in second language acquisition has resulted in studying the relationship between linguistic indices and writing proficiency in English. This thesis investigates the influence of basic linguistic notions, introduced early in English grammar, on automatic proficiency evaluation tasks. We discuss the predictive potential of verb features (*tense*, *aspect*, *voice*, *type* and *degree of embedding*) and compare them to word level n-grams (*unigrams*, *bigrams*, *trigrams*) for proficiency assessment. We conducted four experiments using standard language corpora that differ in authors' cultural backgrounds and essay topic variety. *Tense* showed little variation across proficiency levels or language of origin making it a bad predictor for our corpora, but *tense* and *aspect* showed promise, especially for more natural and varied datasets. Overall, our experiments illustrated that verb features, when examined individually, form a baseline for writing proficiency prediction. Feature combinations, however, perform better for these verb features, which are grammatically not independent. Finally, we investigate how language homogeneity due to corpus design influences the performance of our features. We find that the majority of the essays we examined use present tense, indefinite aspect and passive voice, thus greatly limiting the discriminative power of *tense, aspect*, and *voice* features. Thus linguistic features have to be tested for their interoperability together with their effectiveness on the corpora used. We conclude that all corpus-based research should include an early validation step that investigates feature independence, feature interoperability, and feature value distribution in a reference corpus to anticipate potentially spurious data sparsity effects.

Acknowledgments

I would like to express my sincere gratitude to my supervisor Dr. Sabine Bergler for the continuous support of my masters study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. Additionally, I would like to thank my fellow lab mates for their support and in particular a big thank you to my friends Canberk Ozdemir and Michelle Khalife for their sound advice and for all the fun we have had in the last four years. Last but not the least, I would like to thank my partner in life Nihat Tartal for his constant support and especially my parents Vassiliki Andreou and Nikolaos Panagiotopoulos for encouraging me and helping me with all means at their disposal. Without them, none of this would have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Second Language Learning

Learning a second language (L2) involves mastering reading, listening, speaking and writing. Second language acquisition (SLA) is the research field that monitors which of these skills are acquired and into what extent by non native speakers (NNS) (VanPatten and Benati, 2010). SLA researchers have been interested in the emergence of proficiency in second language, its characteristics and the possible stages that language learners pass through on their way to language competence.

To enhance this process, SLA researchers investigate whether they can associate L2 learner's proficiency stage with specific linguistic devices. This requires analysis of native speakers data on a variety of aspects, including lexical and grammatical features (e.g., *tense*, *aspect*, *syntactic structure* , *word usage*). For example, learners tend to use more modality markers (e.g., *hedges*, *booster*) with increasing proficiency, approaching a level closer to that of native speakers (Trosborg, 1987). Results from these analyses are often employed by international communities of language testing such as TOEIC [1] (Test of English for International Communication) in an attempt to automatically evaluate L2 learner's language cognition. However, associating non native speakers' language proficiency to linguistic devices is not straightforward because of external factors that influence the evaluation process.

Out of four communication skills, learning how to write is the most challenging task for non native speakers (Bell and Burnaby, 1984, Kitao and Saeki, 1992). Second language acquisition researchers also have difficulty examining L2 learner's writing text and attributing linguistic features to their writing acquisition stage. Relating lexical and grammatical features to proficiency stages depends on many external factors such as author's cultural background, topic variety and type of written text. Because of these factors, obtaining consistent results is not always possible. *Tense* for

[1] http://www.etscanada.ca/toeic

instance, is a feature with unclear outcome regarding its relation to writing acquisition in English. We report further the controversial results of *tense* in Section 1.2. At this point, it is important to define what is meant by the terms "language proficiency" and "writing proficiency" since we are going to use them extensively in this thesis.

Language Proficiency is defined by Thomas (1994) as a person's overall competence and ability to perform in L2. However, this definition is considered ambiguous by researchers in the field (Higgs, 1984, Hulstijn, 2011) because it raises questions on how to define *competence* and *ability*. A more recent and less controversial definition was given by Hulstijn (2011) which covers both native speakers' and learners' language proficiency in both linguistic and cognitive competence.

> language proficiency is the extent to which an individual possesses the linguistic cognition necessary to function at a given communicative situation, in a given modality (listening, speaking, reading or writing). Linguistic cognition is the combination of the representation of linguistic information (knowledge of form-meaning mapping) and the ease with which linguistic information can be processed (skill)

Above, we provide the original definition given by Hulstijn (2011) on language proficiency which we believe is the most relevant for this thesis. The reason falls into the two-dimension model that Hulstijn (2011) proposes to describe second language (L2) proficiency. This consists of a two--dimensional grid, with components of linguistic knowledge along one axis (knowledge of lexis, morphology, syntax, grammar and phonology/orthography), crossed with the four language skills (listening, reading, speaking, and writing). It should be noted that in this thesis, the term *proficiency* will sometimes be used to refer to overall language proficiency, as in the above definition, and at other times to proficiency with respect to the specific component this thesis focuses on, *writing proficiency*.

Writing proficiency refers to expressing ideas effectively in written English, recognize writing errors in usage and structure and use language in a way that exemplifies linguistic knowledge (White, 1994). Writing proficiency is a term that can be hard to conceptualize and even harder to define because it is a "slippery term" that hides "an even more slippery concept" (White, 1994). Proficiency may be thought of as skill, adequacy, sufficiency for a defined purpose, or capability. Regardless of how proficiency is specifically defined, it can be seen as something that "is socially determined by communities of readers and writers". In this thesis we examine writing proficiency from the point of linguistic acquisition and we ignore the other two components of structural writing errors and competence of expressing ideas properly.

1.2 Motivation

The rate of acquisition of tense in second language learners has been rather controversial. In the literature, there is no consensus on whether this particular feature can distinguish L2 writing proficiency in English. Ferris (1994), for instance, analyzed a corpus of 160 texts written by L2 students of different origin (Chinese, Japanese, Spanish, Arabic) at three proficiency levels. She experimented on 62 quantitative, lexical and syntactic features (such as word length, relative clauses and pronouns) and observed that only 23 of those were directly related with the level of proficiency of the L2 writers and suggested that the use of tense and more specifically the use of *present* and *past* tense were not related to the writing performance.

Bardovi-Harlig and Reynolds (1995), on the other hand, examined only tense by performing a cross-sectional investigation of 182 adult learners of English as a second language (Arabic, Korean, Japanese, Spanish, Chinese, Portuguese, Thai, Italian, and Russian) at six levels of proficiency and showed that the acquisition of *past* tense in English proceeds in stages. These two SLA research attempts yielded two completely different conclusions regarding the relation of *tense* to writing proficiency. Giving a crisp answer to whether tense can predict writing proficiency is not as easy as it may sound.

Examining tense and its relation to writing proficiency depends on many factors such as the type of writing sample (essay, letter, ...) and author's cultural background. Regarding the first factor, the choice of verb's tense is directly related to the category the text belongs to. For example, in letters and exposition essays the use of *present tense* is more frequent than in narratives. This occurs because L2 writers in exposition essays build their arguments by describing specific events and by providing generalizations and generalizable statements or describing events that are considered general truths to the reader (Beason and Lester, 2010, Hunston, 2006). These require the use of *present* tense, whereas in writing an article or a story the occurrence of *past* tense is more frequent since it requires reporting events that happened in the past (Paltridge, 1996).

Author's cultural background and tense usage are also strongly related. For instance, Japanese and Chinese learners of English have the tendency to use modal verbs when they write English essays because it is part of their rhetorical tradition, given the fact that they use modals in writing essays in their mother tongue (Hinkel, 2002). Speakers of tense-less languages, such as Chinese, Japanese, and Indonesian, used significantly higher rates of past-tense verbs, while the texts of the speakers of Arabic included significantly lower rates of this tense. Among the languages whose speakers wrote the essays, only English and Arabic have a developed morphological system of marking tenses, and in fact, the past tense in these two languages is often used in similar contexts. For example, speakers of Arabic have less difficulty with the English tense system and the *past* tense in particular, than speakers of tense-less languages, such as Chinese, Japanese, and Vietnamese

(Hinkel, 2004).

Coming to a conclusion whether tense and in general linguistic features are related to writing proficiency requires consideration of the factors: *type of written text* and *author's cultural background*. This also explains why the two research attempts by Bardovi-Harlig and Reynolds (1995), Ferris (1994) yielded two different statements regarding the predictive potential of *tense*. Both of them contained the same type of text but examined authors of different cultural backgrounds. Additionally, the number of proficiency groups examined is also a factor that can influence the research process. For example applying features to distinguish between beginners and advanced L2 learners will not yield the same results as with examining three proficiency groups, beginners, intermediate and advanced. Going back to Bardovi-Harlig and Reynolds (1995), Ferris (1994) we notice that in the first analysis six proficiency levels were considered, while in the second one, three. This also highlights the fact that SLA results depend on data design criteria.

We believe that writing proficiency and its association with linguistic factors strongly depends on the design criteria of the analyzed data. Failure to take all parameters into account may lead to inconsistent results. To investigate this hypothesis, we extend our examination from tense to other verb characteristics such as *aspect*, *voice* and *type of subordination*. We also introduce a new feature that describes the complexity of the sentence structure based on the position of the verb in the parse tree (*degree of embedding*). Finally we examine the use of word combinations (word level n-grams) and their relation to L2 proficiency. The reason behind choosing these features lies in their simplicity. All of them are basic linguistic notions introduced at an early stage of English learning. It turns out, their relation to writing proficiency is not that straight forward.

1.3 Objectives

Given the conflicting results regarding the relation of *tense* in English writing proficiency discussed in Section 1.2 we re-examine its potential aiming to obtain a greater insight regarding this feature. The research hypothesis we explore regards the depended relation of *tense* to learner corpora criteria and how it affects its relation to English writing proficiency. By analyzing the verb features *tense*, *aspect*, *voice*, *type of subordination* and *syntactic position of the verb in the parse tree* we want to explore in what extend these basic linguistics notions are related to second language learners writing proficiency and report any emerging trends regarding their individual or combination potential. Finally we aim to obtain better insight regarding the influence of data sparsity and homogeneity when associating our features to writing proficiency.

1.4 Methodology

We explore the relation of our verb features (*tense*, *aspect*, *voice*, *type of subordination* and *syntactic position of the verb in the parse tree*) and word level n-grams to writing proficiency using the following steps: Firstly, we extract verb features and word level n-grams from essays written by second language learners of English. Secondly, we apply those features and their combinations in four text classification tasks. Thirdly, we analyze the relation of our combined and individual features across the four experiments in order to determine any emerging trends across all attempts. Finally, we examine the occurrence of our features in the learner corpora used, so as to report the effect of data idiosyncrasy.

1.5 Contribution

In this thesis, we present the relation of feature sets *unigrams*, *bigrams*, *trigrams*, *tense*, *aspect*, *voice*, *type* and *degree of embedding* to writing proficiency. We found that feature value distribution in a reference corpus has to be taken into consideration when associating linguistic indices to a proficiency group. Using comprehensive analysis of all feature combinations confirms the intuitive hypothesis that the features *tense* and *aspect* have greater performance in combination than individually and in fact this combination is further enhanced by other grammatical features (*voice* and *degree of embedding* in particular). Additionally, we showed that linguistic features have to be tested for their interoperability together with their effectiveness on the corpora used. Overall, we believe that for all corpus based studies results need to be validated by carefully understanding the corpus and the relationship between features. These contributions are built on the initial work detailed in Panagiotopoulos and Bergler (2014).

1.6 Thesis Summary

The thesis is organized as follows: In Chapter 2 we introduce the related work and how linguistic features are applied in real life applications to determine writing proficiency automatically. Chapter 3 presents a background on how learner corpora are used in second language acquisition research. Chapter 4 describes our features in detail by providing definitions and how we implemented them. Chapter 5 introduces our approach in relating our features to L2 writing proficiency and the evaluation metrics we used to analyze the obtained results. In Chapter 6, we provide a detailed analysis of our results obtained on four experiments using the same features. Finally, Chapter 7 presents points learned within the course of this research and proposes some future work for further research on the proposed method.

Chapter 2

Related Work

2.1 Second Language Writing Analysis

Many studies focus on the writing process of second language learners and examine effects of variables such as language background and writing medium. Jiang et al. (2014) investigated a variety of linguistic features to determine the native language (Brazilian, Chinese and Russian) of L2 English learners by examining their written texts. They captured the occurrence of word level n-grams (a contiguous sequence of n words), character n-grams (a contiguous sequence of n characters), Part of Speech n-grams, production rules (in a formal grammar, a production rule is a rewrite rule that specifies a symbol substitution for generating new symbol sequences, e.g. $S \rightarrow NP + VP$) and dependencies (grammatical relationships between constituents in a clause, such as nsubj for non-clausal subject relations and dobj for direct object relations) on argumentative essays from the EFCamDat corpus (Geertzen et al., 2014). Their most predictive features were word level and character n-grams.

The fact that certain punctuation marks and phrases are used more frequently by English learners from one country to another justifies that word level and character n-grams were the most predictive features. For instance, Chinese students do not use dashes as frequently as Russians or Brazilians do. Additionally, phrases such as "*as for me*" and "*to my mind*" are featured in the essays of Russian students, and phrases such as "*try my best*" and "*what's more*" are commonly used by Chinese students, perhaps due to the frequent use of the same expression in Russian and Chinese languages. Finally they reported that the use of prepositional phrases (PPs) is a useful feature for distinguishing among Chinese, Russian and Brazilian students. For example, Chinese students tend to put time references at the beginning of a clause to emphasize its tense, e.g. "On Sunday, he goes to the park and meets friends, and at half past eleven he plays tennis with his friends." Russian's on the other hand still use PPs for temporal reference but their phrases are more complex, including

often two points of temporal reference, e.g. “On Saturday at eleven thirty”.

Biesenbach-Lucas et al. (2000) determined whether the writing medium has an effect on the language students produce. They performed a discourse analysis by comparing word-processed and e-mail writing assignments of non-native speakers of English from largely Asian and Arab countries. The students involved in the study were enrolled in a higher-intermediate English as a Foreign Language course at a university in the United States. For the discourse analysis they focused on text length and eleven cohesive features (see Table 1). Apart from text length, only demonstrative pronouns and sentence connectors appeared to be used differently across media. While demonstrative noun phrases distinguished e-mails from word-processed texts as expected, sentence connectors also distinguished text types but occurred more often in e-mail. They also observed that Arab students tended to use some of the cohesive features of Table 1 more often than Asian students. More specifically, Arab students used more the demonstrative pronouns *these* and *that*, sentence connectors such as *however*, *in addition*, *in contrast*, *also* and the clause subordinator *because*. Whereas Chinese students made more frequent use of pronouns such as *them* and *they*.

demonstrative pronouns (eg., this, that)
demonstrative noun phrases (eg., this policy)
sentence connectors (eg., however, moreover)
clause co-ordinators (eg., and, but, or)
clause subordinators (eg., when, although)
phrase subordinators (eg., because of)
discourse particles (eg., well)
summative expressions (eg., as stated above)
pronouns (eg., I,them, us)
lexical repetition
synonyms

Table 1: Discourse analysis features used by Biesenbach-Lucas et al. (2000)

In contrast, other studies focus on analyzing second language written texts in terms of proficiency. Recent developments in natural language processing allow us to consider deeper-level linguistic features and their relation to writing proficiency. Thus we can examine how differences in perceived writing proficiency are related to linguistic features present in the writers’ texts. Our premise is that linguistic features are indicators of writers’ language abilities, which likely result from their exposure to the language and the amount of experience and practice they have in understanding and communicating in the second language (Kubota, 1998).

McNamara et al. (2010) examined the role of a collection of linguistic features in distinguishing between low and high proficiency undergraduate student essays. Driven by the assumption that cohesion is related to essay quality, they investigated whether cohesive cues (e.g., coreference and connectives) are more predominant in essays judged to be of high proficiency as opposed to those

of lower proficiency. Additionally, they examined three other types of linguistic features, *syntactic complexity* (e.g., number of words before the main verb, sentence structure overlap), *lexical diversity* and *lexical characteristics* (e.g., frequency, concreteness, imagability). They used Coh-Metrix[1], a system for computing cohesion and coherence metrics for written and spoken texts, to calculate the scores for each essay on linguistic indices categorized in five classes, *coreference*, *connectives*, *syntactic complexity*, *lexical diversity* and *word characteristics*. They reported that the most predictive indices of essay quality were *syntactic complexity*, *lexical diversity* and *word frequency*.

More specifically, McNamara et al. (2010) assessed *syntactic complexity* by considering the mean number of higher level constituents (i.e., noun phrase, verb phrase) per word and the number of words before the main verb. The index of syntactic complexity that showed the largest difference between high and low proficiency essays was the number of words before the main verb. For example, one type of simple sentence structure is noun phrase + verb (e.g., "The dog ate"; "The girl walked"; "She laughs"). These simple sentences contrast with the sentence, " Thus, in syntactically simple English sentences there are few words before the main verb", for which there are seven words before the main verb (e.g., are). Their results indicated that high-proficiency writers use more complex syntax than low-proficiency writers.

McNamara et al. (2010) measured *lexical diversity* using Coh-Metrix. Lexical diversity refers to how many different words occur in a text in relation to its total number of words. When lexical diversity is at a maximum (all words are different), then the text is likely to be either very low in cohesion or very short. Their results demonstrated that more proficient writers use a greater range of lexical diversity in their essays. Word frequency showed the largest difference between high and low proficiency. They measured word frequency by searching in CELEX[2] the reported frequency of each word. CELEX is a database that consists of frequencies taken the early 1991 version of the COBUILD corpus, a 17.9 million-word corpus. Their results suggest that high proficiency writers use words that occur less frequently in CELEX.

Overall a considerable amount of research has been conducted on the role of linguistic features in second language (L2) writing proficiency (Connor, 1990, Ferris, 1994, Ortega, 2003). For example, several Test of English as a Foreign Language (TOEFL) research investigations aimed at better understanding variation in writing quality. These studies have established that in large scale testing and university-level assessments of student essays, many characteristics of simple or sophisticated uses of language are considered to be markers of L2 writers' proficiency in English. *Syntactic complexity*, use of specific *grammatical constructions* and *lexical features* are three of those linguistic characteristics (Frase et al., 1999). We report how other researchers applied those linguistic indices in an attempt to relate them with L2 writing proficiency before we show our approach.

[1] `http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html`
[2] `http://celex.mpi.nl/`

2.1.1 Syntactic Complexity

Syntactic complexity (also called syntactic maturity or linguistic complexity) refers to the range of forms that surface in language production and the degree of sophistication of such forms. This indice is important in second language research because of the assumption that language development entails, among other processes, the growth of an L2 learner's syntactic skills and her or his ability to use them appropriately in a variety of situations (Ortega, 2003). A large variety of measures have been proposed for characterizing syntactic complexity in the second language writing development literature. Table 2 shows some of those statistical measures.

Measure	**Acronym**	**Definition**
Mean length of clause	MLC	# of words/# of clauses
Mean length of sentence	MLS	# of words/# of sentences
Mean length of T-unit	MLT	#of words/# of T-units
Clause per T-unit	CT	# of clauses/# of T-units
Dependent clauses per clause	DC/C	# of dependent clauses/# of clauses
Dependent clauses per T-unit	DC/T	# of dependent clauses/# of T-units
T-units per sentence	T/S	# of T-units/# of sentences
Complex nominals per clause	CN/C	# of complex nominals/# of clauses
Complex nominals per T-unit	CN/T	# of complex nominals/# of T-units

Table 2: Syntactic complexity measures

A definition of the six productions units and syntactic structures involved in the measures of Table 2 are recaptured below.

Sentence: A sentence is defined as a group of words (including sentence fragments) punctuated with a sentence final punctuation mark, including a period, exclamation mark, question mark, and occasionally elliptical marks or closing quotation marks (In(1) we present a sentence example).

(1) I didn't sleep yesterday night because I was sick.

Clause: A clause is a structure with a subject and a finite verb, including independent, adjective, adverbial, and nominal clauses, but not non-finite verb phrases, which are included in the definition of verb phrases instead (In (1) *I didn't sleep yesterday* is the independent clause). Dependent clause: A dependent clause is defined as a finite adverbial, adjective, or nominal clause (In (1) *because I was sick* is the dependent clause).

T-unit: A T-unit consists of a main clause and any dependent clause or non clausal structure attached or embedded. For example the sentence in (2) contains two independent (main) clauses; thus it has two T-units. The sentence in (1) is one T-unit by itself.

(2) There was a woman next door and she was a singer.

Complex nominal: Complex nominals include 1) noun phrases with one or more of the following pre- or post-modifiers: adjective, possessive, prepositional phrase, adjective clause, participle, or appositive; 2) gerunds and infinitives in subject position and 3) nominal clauses. A nominal clause is a subordinate clause that functions as a noun phrase. For example in (3a) the phrase *where I stood* is a nominal clause, similar case is the phrase *that he is here* in in (3b)

(3) (a) From *where I stood*, I saw the horse.
(b) I know *that he is here*.

Ortega (2003) examined syntactic complexity and it's relation to second language (L2) proficiency across twenty one studies. She focused on the six most frequently used syntactic complexity measures across those studies (MLS, MLT, MLC, CT, DC/C and T/S), defined at Table 2. She concluded that the most statistically significant measure in all the twenty one studies where MLS and more precise 4.5 words or more per sentence, MLT (two or more words per T-unit) and MLC (one or more words per clause). Overall these measures are related to writing proficiency however, should not be considered as absolute developmental indices or as direct indices of language ability. She notes that, "more complex" may mean "more developed" in many different ways, and the nature of L2 development cannot be sufficiently investigated by means of these global measures alone. They can only provide a start for the analyst to search further for evidence relating syntactic complexity to language proficiency.

Similar complexity measures were identified by Haiyang and Xiaofei (2013). They analyzed 600 essays using ten syntactic complexity measures (MLC, MLT, DC/C, DC/T, T/S, CN/C, CN/T) to investigate whether low proficiency level non native speakers of English (NNS) differed with respect to writing complexity with advanced second language learners. Results showed significant differences in syntactic complexity measures MLS, MLT and CN/T which indicate that higher proficiency level writers construct more complex sentences than those at lower proficiency levels. This is due to the false assumption that native speakers of English write text with increased syntactic complexity; thus advanced learners try to imitate this false belief (Hinkel, 2003).

Using statistical measures to capture syntactic complexity is the quick and easy way but has drawbacks: sentence length is not an accurate measure of syntactic complexity, and syllable count does not necessarily indicate the difficulty of a word. Additionally, a student may be familiar with a few complex words (e.g. dinosaur names) but unable to understand complex syntactic constructions. In contrast to these traditional measures of text L2 writing research has been conducted using deeper level linguistic measures. This includes examining students essays in terms of using specific grammatical structures which can distinguish NNS writers of different proficiency levels.

2.1.2 Grammatical Structure

In the field of second language acquisition, researchers examine how L2 learners acquire grammatical phenomena and report their relation to writing proficiency (Granger, 1998). This includes both single grammatical targets (i.e., *articles*) and a broader range of grammatical structures (i.e., *articles*, *copula "be"*, *regular past tense*, *irregular past tense* and *preposition phrases*).

Mattar (2003) focused on English learners of Arabic origin and studied the use of subordinating constructions expressing contrast: *although + clause*, *because +clause* and *despite + gerund/NP*. He compared 89 university students of Arabic origin in three different groups according to their English proficiency (low, intermediate, advanced). The results showed that the frequency of using the subordinating adverbs *despite + gerund/NP* and *because of + NP* was much higher in the group of advanced students while the lower-level students avoid using those adverbs and preferred *although + clause* and *because +clause* instead.

The use of subordinate clauses (*noun clauses*, *adverbial clauses* and *relative clauses*) was examined by Grant and Ginther (2000). Their study examined L2 opinion essays distributed across three levels of proficiency (beginner, intermediate and advanced). They report that writers use more *subordination* as they become more proficient. More specifically, they noticed that as proficiency level increased so was the use of *adverbial clauses* and *noun clauses*. Additionally, they attributed the fact that *relative clauses* were not a distinguishing factor across proficiency levels to the type of text they examined. In opinion essays writers are asked to discuss and give reasons for supporting their opinions; thus the use of *relative clauses* (that-complement) is likely to occur.

Ishikawa (2010) examines the use of linking adverbials in essays written by Japanese university students of four proficiency levels (lower, middle, semi-upper, upper). He defines *linking adverbials* as adverbs that connect two independent clauses or sentences and provide transition between ideas (i.e. moreover, furthermore). The results suggest a direct relation between the linguistic feature and the lower writing proficiency class. Students of this level in English have the tendency to use specific linking adverbials such as *also*, *too* and *again*. However, no differences were observed in the use of linking adverbials among the other three levels of proficiency.

Hawkins and Buttery (2010) explored the use of learner corpora to chart grammatical development with increasing proficiency, using the notion of criterial features. They reported the occurrence frequency of each feature in relation to the level of proficiency. If the occurrence of a feature at one level is significantly different from the level below, it is criterial to that level. More specifically, they investigated linguistic properties which examiners use as markers when they assess the L2 learners' level of proficiency. They focused on the six proficiency levels of the Common European Framework of Reference (CEFR) (C2 Mastery, C1 Effective Operational Proficiency, B2 Vantage,

B1 Threshold, A2 Waystage, A1 Breakthrough) and studied two grammatical categories, verb co-occurrence patterns and relative clauses.

Verb co-occurrence patterns refer to grammatical constructions of English defined in terms of the verb and its co-occurring phrases. To identify those patterns they used the Robust Accurate Statistical Parsing (RASP) parser (Briscoe et al., 2006) in combination with the verb sub-categorization lexicon of Korhonen et al. (2006), which contains an extensive list of valid grammatical constructions including verbs. For example if a verb is intransitive, it should not be followed by an object; thus a simple valid grammatical construction is the one in (4). This construction is identified as NP-V (a noun phrase followed by the intransitive verb).

(4) They went.

Although, they did not find verb co-occurrence patterns that can distinguish all six proficiency levels from A1 to C2, they noticed that distinguishing between two proficiency groups beginner (levels A1, A2, B1) and advanced (levels B2, C1, C2) instead of six yielded specific verb co --occurrence patterns (see Table 3). These patterns appear more frequently in essays written by advanced learners than beginners and their frequency difference is statistically significant.

Pattern	**Example**
NP-V-NP-AdjP (Obj Control)	He painted [the car] red
NP-V-NP-as-NP (Obj Control)	I sent him as [a messenger]
NP-V-NP-S	He told [the audience] [that he was leaving]
NP-V-P-NP-V (+ing)(Obj Control)	They worried about him drinking
NP-V-P-VPinfin (Wh-move)(Subj Control)	He thought about [what to do]
NP-V-S (Wh-move)	He asked [what he should do]
NP-V-Part-VPinfin (Subj Control)	He set out to win

Table 3: Verb co-occurrence patterns distinguishing [B2, C1, C2] from [A1, A2, B1] levels of proficiency according to Hawkins and Buttery (2010)

2.1.2.1 Tense and Aspect

The use of tense and aspect and their relation to second language writing proficiency has been a major topic in second language acquisition research. Patanasorn (2013) investigated the claim that present perfect emerges after L2 learners demonstrate a stable rate of accurate use of the simple past. Fifteen Thai learners of English participated in the study, nine were considered high proficiency learners and six were considered low proficiency learners. Participants were administered a writing test that was designed to elicit the use of the simple past and present perfect. Every essay was checked for attempts of *simple past* and *present perfect* use in their appropriate context. The unit of analysis was types of verbs. The verb type was labeled as either appropriate *present perfect* usage or *simple past* usage. Appropriate was defined as correct choice of tense and aspect regardless

of its form. Thus, mistakes on spelling and grammatical inflections were ignored. To determine emergence of the *present perfect*, emergence was defined as appropriate use of the *present perfect* with at least three types of verbs. Thus, participants who used three or more different types of verbs in the *present perfect* in its appropriate context were considered to have demonstrated emergence. Using three types of verbs as a criteria was to ensure that the use was not achieved by mere chance.

The findings from this study suggest that L2 learners of English acquire the simple past before the present perfect tense. More specifically, there is a direct relation between the accurate use of simple past and the usage of present perfect. High proficiency learners of English use both tenses more often and appropriately, in contrast writers of low proficiency make less accurate use of simple past and almost never use *present perfect*.

Similarly, Hinkel (2004), concentrating on advanced learners of English and native speakers only, reported on the use of English tenses, aspect and passive voice in academic texts. She reported that advanced L2 writers showed lower frequency of *present perfect* and high frequency of *simple past* in their papers compared with L1 writers. Moreover, L2 writers showed reduced use of *passive voice* constructions, possibly due to lack of familiarity. Overall she highlights advanced students' difficulty with the conventionalized uses of *tense*, *aspect* and *passive voice* in written academic discourse despite several years of second language learning and use. Her study reports the majority of these students avoiding "complex verb phrase constructions as passive voice, the perfective aspect, or predictive/hypothetical would".

Min (2013) examined the relationship of second language writing proficiency with the usage of verb tense and aspect. His study focused on examining English verb tense and aspect combinations (see Table 4) in 120 argumentative essays corresponding to three proficiency levels (intermediate L2, advanced L2, and native speakers). His findings suggested that the use of specific tense-aspect patterns was relevant to the students' L2 writing proficiency because advanced students showed their grammatical knowledge in their essay's purpose, content, and discourse register.

Tense	**Aspect**
Simple	Present
	Past
Perfect	Present
	Past
Progressive	Present
	Past
Predictive	will
	would
	may/might

Table 4: Taxonomy of verb tense/aspect according to Min (2013)

More specifically, regarding the essay's purpose and content, he assessed each tense and aspect combination by examining its correct use in the text. For example, he observed that advanced L2 learners made more but appropriate use of *present perfect*, as opposed to intermediate learners who used more *simple past* even in situations where *present perfect* was more relevant. In particular advanced students not only used the *present perfect* to express an event that started in the past and continued in the present but also to express current relevance (current relevance indicates the result or effect of a situation it still holds at the moment of speaking). For example a beginner student would write "Although the FDA *introduced* the GMO food as a safe food" the advanced student would use *has introduced* which is more appropriate.

As far as discourse register is concerned, he evaluated the tense and aspect shifts in a text. Advanced learners showed natural shifts in the text but beginners changed the tense in a paragraph with no reason. For example the tense shifts in (5) have no logical reasoning.

(5) In the lecture, opposite side people *were worry* about that consumers *will* consider GM foods as something harmful and wrong after labeling even if there *are* no risk.

2.1.2.2 Modal Verbs

The use of modal auxiliaries and their association with second language writing proficiency has been approached by many researchers in second language acquisition. McDoual (2010) reports the role of modal auxiliaries (*can*, *would*, *will*, *may*, *could*, *should*, *must*, *might* and *shall*) in distinguishing L2 learners writing proficiency. McDoual (2010) focuses on the functional use of modal auxiliaries in opinion essays written by Korean advanced and intermediate learners of English. He divided modals in two functional categories: *propositional modality* and *event modality*. His findings indicate that *event modality* is acquired earlier than *propositional modality* and that with increasing proficiency L2 learners use propositional modals more frequently.

A similar attempt performed by Begi et al. (2013) examined the use of English modal auxiliaries (*can*, *could*, *may*, *might*, *must*, *should*, *will*, *would*, *have to*, *need to*) in terms of frequency and function on argumentative essay. They examined samples written by Malaysian learners of English on two proficiency levels (beginners, advanced). The findings of their study showed that beginner L2 learners use the present tense form of *can* and *will* more frequently than advanced learners. Finally, they examined the functional use of modal auxiliaries by dividing them in: modals of ability (*can*, *could*), modals of probability (*will*, *would*, *may*, *might*) and modals of necessity/obligation (*should*, *must*, *have to*, *need to*). They proved that the modals of ability are mostly found in beginner level students' essays than advanced.

Vethamani et al. (208) investigated the use of modal auxiliaries in distinguishing low and advanced Malaysian second language learners of English. The aim of their study was to investigate the distribution and functions of modals used in the students' writing. Their findings showed that modals expressing ability and certainty *can*, *will* and *could* were used by both levels equally. Modals of probability/possibility (*will*, *would*, *may*, *might* and *shall*) showed lower frequencies of use in the writing. Also, students at the lower level used present tense modals (*will*, *can*, *shall*, *may*) more than advanced L2 learners, whereas past tense form modals (*would*, *could*, *should*, *might*) are more often in advanced learners writing.

Many second language acquisition researchers concentrated on grammatical competence and its relation to writing proficiency because grammar is considered to be more fundamental and creative, and to consist of elements of the generative system of language. In addition to grammatical structures a lot of attention is paid to the importance of lexical features as a predictor of writing proficiency.

2.1.3 Lexical Features

In second language acquisition studies there is an association between writing proficiency and lexical features. One of the most common lexical indices that are examined is related to vocabulary size (Crossley et al., 2013). Vocabulary size relates to how many words a learner uses.

For instance, Engber (1995) found that more proficient L2 writers use a more diverse range of words, and thus show greater lexical diversity. He examined 66 essays written from students of mixed cultural backgrounds (Arabic, French, Italian, Japanese, Korean, Russian, Spanish and Thai) distributed in four levels of proficiency. The lexical features he used included lexical variation (a type/token ratio, expressed as the ratio of the number of different lexical items to the total number of lexical items in the essay), error-free variation (lexical variation without lexical errors) and percentage of lexical errors. The results showed that lexical variation with or without errors was highly related to the writing proficiency. More specifically, intermediate to high-intermediate writers used a greater variety of lexical choices in the correct lexical form.

For instance, Crossley et al. (2011) studied 100 writing samples from 100 L2 learners of different cultural backgrounds. The samples were analyzed for lexical indices such as word frequency and correctness by the computational tool Coh-Metrix. The L2 writing samples were categorized into beginning, intermediate, and advanced groupings. The results indicated that automated, lexical indices can be used to predict the language proficiency levels of second language learners based on their writing samples. They found that the more specific words a writer uses, the more proficient he is. This contrasts with other studies which state that L2 learners' word use becomes less specific as time spent studying a language increases. Crossley et al. (2011) counted the number of different

words, which resulted in the conclusion that advanced learners use more specific words and different ones, where beginners use general words and have limited lexical variety.

The link between word associations and language acquisition is supported in several recent studies. Word association relates to the meaningfulness of the word. (Toglia and Battig, 1978). Words with high meaningfulness include words like *food*, *music*, and *people* while words with low meaningfulness include *acumen*, *cowl*, and *oblique*. Words in the first list invoke multiple word associations, while those in the second list have fewer associations. Zareva (2007) found that higher proficiency learners provide significantly more word associations than intermediate and beginning level learners. She argued that larger vocabularies allow for a greater number of word associations. Her study involved written text from Native Speakers, L2 advanced and L2 intermediate learners of English. The analysis of the size and the diversity of the intermediate learners word associations domains showed that they had a repertoire much smaller in size and less diverse than the Native Speakers' and the advanced learners', whereas the advanced learners' associative domain was similar to the NSs in size, but showed a trend to slightly greater heterogeneity.

Other studies have examined the use of more explicit cohesive devices such as connectives. Jin (2001), for example, examined the use of connectives in Chinese graduate students writings. He found that all students, regardless of proficiency, use cohesive devices but advanced writers use these devices more often than do intermediate writers. Similarly, Connor (1990) found that higher-proficiency L2 writers use more connectives. Past research, then, demonstrates that L2 writers judged to be advanced sometimes produce text which is less cohesive when measured by word overlap, but at other times their writing is more cohesive as measured by their use of connectives.

Overall, we have seen that part of second language acquisition research focuses on identifying linguistic indicators that are associated with writing proficiency. The main goal of these efforts is to obtain a better understanding of second language learners of English acquisition the language. Currently English as a Second Language research focuses on how to incorporate the relevant linguistic indices to create systems that automatically evaluate writing products of L2 learners.

2.2 Writing Proficiency Assessment

The Graduate Management Admissions Test (GMAT), the Test of English as a Foreign Language (TOEFL), the Graduate Record Examination (GRE) are three of several large scale assessment programs which evaluate the proficiency level of second language (L2) learners in English. Part of their evaluation process includes the assessment of writing proficiency by analyzing essays written by L2 learners under a controlled environment (i.e. number of words, time, topic). In general, raters for English as a Second Language (ESL) writing evaluate essays using a combination of criteria defined by research in ESL. Some of those evaluation markers include organization, content,

grammar, sentence structure, coherence, handwriting and editing skills (Vaughan, 1991). Many attempts have been made to incorporate the majority of those markers, so as to create systems which automatically evaluate the writing ability of second language learners.

For instance, Burstein et al. (2001) created *e*-rater; a system which automatically evaluates the writing ability of second language learners of English by identifying features related to writing proficiency in student essays so they can be used for scoring and feedback. *E*-rater uses ten broad feature types extracted from the text using Natural Language Processing techniques, eight represent writing quality and two content. These features correspond to high-level properties of a text, such as syntactic variety, organization of ideas, and vocabulary usage. Each of these high-level features is broken down into a set of ground features. For example, syntactic variety is subdivided into features which count the occurrence of syntactic constructions of various clauses, including infinitive (an example is given in (6a), *to plan parties*), complement (in (6b) *that he did it*), and subordinate clauses.

(6) (a) Rosie loves *to plan parties*.
(b) I heard the evidence *that he did it*.

The resulting counts for each feature are associated with cells of a vector which encodes all the syntactic features of a text. Similar vectors are constructed for the other high-level features. The syntactic structures such as complement clauses are used in combination with cue words and terms to create discourse annotations which denote the arguments made by the writer. For instance, in the essay text *e*-rater's discourse annotation indicates that a contrast relationship exists, based on discourse cue words, such as *however*. Discourse features have been shown to predict the holistic scores that human readers assign to essays, and can be associated with *organization of ideas* in an essay. Finally the essay arguments identified by the discourse annotations are analyzed in terms of vocabulary usage by examining the word usage between the boundaries of the argument. The word frequency in combination with the type of discourse plays a role in characterizing the level of proficiency of the L2 author.

Briscoe et al. (2010) on the other hand, identified lexical and grammatical properties which are highly discriminative for automatically assessing linguistic competence in learner writing. They focused on analyzing a text in terms of word ngrams (word unigrams and word bigrams), Part-of-Speech (PoS) ngrams (*unigrams*, *bigrams*, *trigrams*), error rate, grammatical structure and number of words per text. Lexical terms (e.g., *unigrams*) get extracted along with their frequency counts, as in a standard "bag-of-words" model (a text is represented only by the words it contains). These are supplemented by *bigrams* of adjacent lexical terms. *Unigrams*, *bigrams* and *trigrams* of adjacent sequences of PoS tags are extracted along with their frequency counts. All instances of these feature types are included with their counts in the vectors representing each essay.

Additionally, the grammatical structure is extracted using the parse tree generated by RASP parser (Briscoe et al., 2006). For each sentence every grammatical construction is found (for example, the grammatical structure 'S/pp-ap_s', indicates that a sentence (S) with preposition phrase (PP) with adjectival complement (ap_s), e.g., *for better or worse, he left*) and along with its frequency count is represented as a cell in the vector containing information about the rest of the feature types. The text length in words is used as a feature less for its intrinsic informativeness than for the need to balance the effect of text length on other features. For example, error rates, ngram frequencies, etc. will tend to rise with the amount of text, but the overall quality of a text must be assessed as a ratio of the opportunities afforded for the occurrence of some feature to its actual occurrence.

Automatic assessment of writing proficiency is a relatively new field of ESL research. There is still the need of experimentation as to which features to use and how to combine them so as to give the best results. For example, McNamara et al. (2010) tried to incorporate cohesive metrics as is with inconclusive results, whereas Burstein et al. (2001) used cohesion cues in combination with vocabulary usage which proved more successful in evaluating writing proficiency. There is still no agreement on methods or features to use in order to evaluate writing ability. Nevertheless, there exists fertile ground for experimentation in finding those linguistic indicators.

Chapter 3

Learner Corpora

3.1 Background

According to Sinclair (1996) Computer Learner Corpora (CLC) are defined as systematic electronic collections of spoken or written text produced by learners. The texts included in the corpus should be a representative and balanced selection based on a number of criteria such as learners' levels and the learners' first language (L1). They should not be intended merely for use in one particular learner study (or a limited number of studies) but for more general uses (Sinclair, 1996). For brevity the term computer learner corpora and learner corpora will be used interchangeably in the rest of this thesis.

The usefulness of a learner corpus is directly proportional to the care that has been exerted in collecting the data, as well as defining the design criteria. The process of collecting learner data differs from the common data collection because it involves some degree of control which means that learner corpora are rarely fully natural. Composition, for instance, represents free writing because learners can write what they want rather than having to produce items the investigator is interested it. But they are also controlled to some extent since some task variables, such as the topic or the time limit, are often imposed on the learner (Granger, 1999). To qualify as learner corpus data, the language sample must consist of continuous stretches of discourse which contain both erroneous and correct use of the language. Isolated sentences, words or only erroneous sentences cannot be considered as a legitimate learner corpus (James, 1998).

Additionally, the criteria under which learner corpora are constructed are very important for English as a Second Language (ESL) research. A random collection of heterogeneous learner data does not qualify as a learner corpus (Granger, 2002). Learner corpora should be compiled according to strict design criteria, relating to both the learner and the task (see Table 5)

Learner	Task Settings
Learning context	Time limit
Mother tongue	Use of reference tools
Other foreign languages	Exam
Level of proficiency	Audience/interlocutor
etc...	etc...

Table 5: Learner corpora design criteria according to Granger (2002)

Learner corpora are used either for Contrastive Interlanguage Analysis (CIA) or Computer-aided Error Analysis (Granger, 1999). The first one involves the quantitative and qualitative comparisons between native (NS) and non-native (NNS) data or between different varieties of non-native speaker data. The second focuses on identifying and analyzing errors in interlanguage (Granger, 1996).

Contrastive Interlanguage Analysis comprises either comparison between native and non-native (NS/NNS) or among non-native (NNS/NNS) speakers. NS/NNS comparisons aim to shed light on features that can distinguish second language learners from first language learners. It involves the linguistic analysis of written or spoken text, so as to isolate a range of features describing NNS such as under- and over- use of words, phrases and structures. The comparison of non native data with native ones is essential in second language teaching because it helps learners to improve their proficiency and brings it closer to some native speakers (Granger, 1998).

NNS/NNS comparisons involve examining in contrast learners' data from different population, mother tongue background, proficiency level, ... (Granger, 1999). For example, comparisons of learner data from different cultural backgrounds help identify which features are distinctive among the different national groups; thus are L1 dependent. Additionally, it can shed light on which features are shared by several learner populations and are therefore more likely to be developmental. This is important for natural language identification studies which try to highlight those linguistic indices that characterize learners of different cultural backgrounds (Ishikawa, 2010). In addition, NNS/NNS comparison provides useful insight to second language teachers. By contrasting learners data of different proficiency levels, they observe the language acquisition progress and create their course curriculum accordingly.

Computer-aided error analysis usually involves the method of selecting error-prone linguistic items (such as words, phrases, or syntactic structure) and then scrutinizing the learner corpus to identify instances of misuse. In another more time consuming approach the learners' data undergo error tagging in order to capture all the possible errors or at least all errors of a particular category (i.e. tense misuse). This is useful for learners to discover difficulties which they were not aware of (Granger, 1999). Computer aided error analysis is out of the scope of this thesis; thus we will not elaborate further on any research progress related to it.

3.2 Learner Corpora in Proficiency Assessment

There are three ways to use learner corpora to assess language proficiency, *corpus-informed*, *corpus--based* and *corpus-driven*. In *corpus-informed* approaches, learner corpora are used as a reference source to provide information regarding a learner's language use at certain levels of proficiency. The researcher examines the learner corpora so as to identify language features that can distinguish learners' levels of proficiency. These features are then given as a guideline to writing proficiency evaluators to guide them through the marking process or to validate their grading. For example, Hawkey and Barker (2004) performed a qualitative analysis of written scripts using both intuitive and computer-assisted approaches. They proposed key language features (such as organization, cohesion, range of structures, logical structure and vocabulary usage) which distinguish performance in writing at four pre-assessed proficiency levels. Additionally, they suggested how these features might be incorporated in a common scale for writing which would assist test users in interpreting levels of performance across exams and locating the level of one examination in relation to another.

In *corpus-based* approaches, learner data are examined so as to identify linguistic indicators that can refute or confirm a researcher's hypothesis. Occasionally, learners' use of language is compared to the language of native speakers. Hawkins and Filipović (2012) introduced the notion of criterial features, linguistic descriptors that are characteristic and indicative of L2 proficiency at each level. Driven by the hypothesis there exist specific linguistic features for each level of proficiency, they compared linguistic indicators found in L2 learners' text to those found in L1 learners of English. More specifically, they compared grammatical and lexical patterns extracted from learners' essays contained in Cambridge Learner Corpus[1] to those used by native speakers of English included in British National Corpus[2]. Their initial hypothesis is supported in the sense that there are linguistic properties that distinguish L2 learners (beginner, advanced) writing proficiency from L1 learners.

Finally in *corpus-driven* approaches, learner corpora are examined using statistical analysis techniques. The data processing is not influenced by any idea or claim like in corpus-based approaches and the involvement of the researcher is the minimum (in contrast with the corpus-informed approach). In this approach the data actually reveal the questions that should be asked by the researcher. Wulff and Gries (2011) propose a new way of measuring accuracy using conditional probabilities. By defining accuracy in L2 production as "the selection of a grammatical or lexical construction in its preferred context within a particular target variety and genre" and through probabilistic analysis of lexico-grammatical association patterns, they showcase constructions used by learners which are indicative for language assessment. For example, the verb *give* is used often in English. It can occur in ditransitive (7a) or in prepositional construction (7b).

[1] `http://www.cambridge.org/gb/cambridgeenglish/catalog?site_locale=en_GB`
[2] `http://www.natcorp.ox.ac.uk/`

While both are grammatically correct the first one is used more often by native speakers of English than the second one. Thus for ditransitive construction of the verb *give* has higher probability of distinguishing native speakers of English from non natives.

(7) (a) She gave the squirrel some bread.
(b) She gave some bread to the squirrel.

Similarly considering the infinitival, *to feed*, and gerundive, *feeding*, complementation constructions of verb *began* in (8a) and (8b) respectively. The latter one has the tendency to appear in L1 text more often than in L2. Thus this construction has a high probability distinguishing the two learner types than the first one.

(8) (a) She began *to feed* the squirrels.
(b) She began *feeding* the squirrels.

Crucially, the above is based on generalizations of verb/construction use across speakers and cases/contexts. Their approach provides language teachers with more concrete suggestions for the implementation of second language research into their teaching.

We introduced the three ways a learner corpora can be used in Second Language Acquisition research in order to perform a contrastive interlanguage or computer-aided error analysis. In this thesis we analyze our data using the corpus-based approach. Starting from our initial motivation, the role of tense in writing ability, we expand to other verb characteristics introduced in Chapter 4 to observe their relation with L2 proficiency.

3.3 Language Proficiency Levels

Analyzing students' writing at various proficiency levels can be done either using Contrastive Interlanguage Analysis (CIA) or Computer-aided error analysis. Whatever the approach, a major difficulty is that "proficiency level" has often been a fuzzy variable in learner corpus compilation and analysis.

Proficiency has mostly been operationalized and assessed globally by means of external criteria, typically learner-centered methods such as learners' institutional status (Callies et al., 2014). However, recent studies show that global proficiency measures based on external criteria alone are not reliable indicators of proficiency for corpus compilation. "Hidden" differences in proficiency often go undetected or tend to be disregarded in learner corpus analysis (Pendar and Chapelle, 2008). For example, all other things being equal, a person with two years of exposure to English at grade school and four years at the university is likely to be less proficient in English than someone with five years of exposure to English at grade school and ten years at the university, even if they are both considered advanced learners of English in terms of their year of university study in English. One

would also expect someone who has spent sixteen years in an English-speaking country to be more proficient than someone who had not spent any time in an Anglophone environment.

Analyses based on learner corpora compiled according to external criteria, may not offer a clear distinction among proficiency levels. Thus, there is a need of a corpus-based description of language proficiency to account for inter-learner variability and seek homogeneity in learner corpus compilation and L2 assessment (Lozano and Mendikoetxea, 2013). Currently, three major frameworks are used to describe language proficiency from three different countries, Europe, Canada and USA.

American Council for the Teaching of Foreign Languages (ACTFL) guidelines represent a hierarchy of global characterizations performance in speaking, listening, reading, and writing (American Council, 2012). Each description is a representative, not an exhaustive, sample of a particular range of ability, and each level subsumes all previous levels, moving from simple to complex in an "all-before-and-more" fashion. They were designed to distinguish language competence among university level students by defining nine levels of proficiency.

Common European Framework of Reference for Language (CEFR) was developed by the Council of Europe in 2001 (Council of Europe, 2001) and has become an important reference document for language testing in Europe. CEFR was developed in an attempt to overcome the barrier arising in the field of modern languages from the different educational systems in Europe. It has a high influence in foreign language teaching, learning and assessing because it sets clear guidelines on language learners regarding what they have to learn in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively.

Canadian Language Benchmarks: English as a Second Language for Adults were created for working purposes (Canadian Language Benchmarks, 2012). Many employers in Canada are turning to internationally-educated professionals to meet their demands for highly trained and skilled workers. The Canadian Language Benchmarks provide national language guidelines for assessing immigrants' English as a second language ability. The Benchmarks provide a set of descriptors of learners' English levels in listening speaking, reading, and writing, with the descriptors set in the context of 12 Benchmarks.

All of the above guidelines have been criticized in terms of validity. No real empirical basis can be claimed by any of the scales for the descriptors, despite the role played by statistical analysis and academic theory in the creation of the scales. For example, Fulcher (2004) argues CEFR's proficiency level definitions are mostly intuitive. Although CEFR has been highly influential in language testing and assessment, the way it defines proficiency levels using "can-do-statements" has been criticized, because they are often too impressionistic. For example, a learner at the C2

level is expected to maintain “consistent grammatical control of complex language”, whereas at C1 he/she should ”consistently maintain a high degree of grammatical accuracy“ (Council of Europe, 2001). Also the Council of Europe was unable to include certain aspects of language use, areas such as literary appreciation and several pragmatic and strategic aspects of language. These areas appeared to represent different factors or aspects of language use than language proficiency.

The ACTFL Guidelines have been criticized because no information has been released concerning how data were collected and questions about the universality of the scales often arise (Liskin-Gasparro, 2003). Moreover, Fulcher (2004) objects that the Guidelines describe performance in terms of the perceptions of a native speaker; thus the level definitions are lucky approximations and were not constructed based on empirical evidence. Finally the Canadian Language Benchmark scale has not yet been empirically validated and is heavily reliant on details of performance conditions that are provided, a fact that decreases the validity of the scale as it is applied in different contexts (Hudson, 2005). Also, these scales are broad and do not use a native speaker as the norm, although they do use such terms as ”native-like“.

Nonetheless, according to Callies et al. (2014) and Díez-Bedmar (2010) these frameworks have much to offer. Firstly, they have greatly simplified the language testing process. Though the varying levels of language learners have been discussed in explicit detail, they have been summarized into tables that easily fit onto one page. Secondly, the scales are designed for not only teachers, but also for learners, and are easily accessible to non-experts. Finally, while academic research focuses on the complexities as well as the social and political nature of language learning, the express goal of the scale makers is transparency and simplicity. Paradoxically, perhaps it is the simplicity and explicitness of the scales that make them a target for criticism.

According to Hawkins and Filipović (2012) second language acquisition guidelines identify stages of proficiency, as opposed to achievement. They do not measure what individuals achieve through specific classroom instruction, but assess what individuals can and cannot do. These guidelines are not based on a particular linguistic theory or pedagogical method, and are intended for global assessment. Such global, vague and underspecified descriptions have limited practical value to distinguish between proficiency levels and also fail to give in-depth linguistic details regarding individual languages or learners’ skills in specific registers. These shortcomings have led to an increasing awareness among researchers of the need to identify more specific linguistic descriptors which can be quantified by learner data. To avoid these methodological limitations, a fruitful line of research combines the use of proficiency guidelines to establish students’ proficiency levels and the use of CIA to analyze their writing production.

3.4 Learner Corpora Used in CIA

Some Second Language Acquisition (SLA) researchers study how learners acquire a second language by collecting and analyzing learner data. For example, Kusher et al. (2001) used data coming from email exchanges written in English from Spanish university students. Through this activity a database of learner corpora was generated in order to report the developmental progress of L2 students over time. Lardiere (1998) collected data over the period of 18 years from an English learner, Patty. Patty was a native Chinese speaker who had lived in the United States for about 10 years by the time of the first recording and over 18 years by the time of the second and third recording. While this type of corpus allows for a detailed and longitudinal examination of interlanguage development, conclusions are limited as they cannot be extrapolated to other learners.

Much effort has been made to appropriately compile learner corpora that can aid in describing learner language (Granger, 1998). Because through the investigation of authentic natural language data, researchers can focus on theoretical and pedagogical issues while educators can concentrate on the needs of learners. Learner corpora that satisfy the design criteria discussed in Chapter 3.1 are predominantly found in Europe and Asia. We present a sample of existing learner corpora appropriate for Contrastive Interlanguage Analysis, as well as, details about each corpus including the size of the corpus, the purpose of the corpus, the proficiency level of the learners, and the availability of learner background information.

The International Corpus of Learner English, (ICLE) (Granger, 2003) can be taken as the starting point in the exploration of large-scale learner corpora and has inspired a growing interest in learner corpus research. The current version of ICLE (Granger, 2009) consists of 6,085 argumentative essays of maximum 700 words written by higher intermediate to advanced learners of English. ICLE is organized in different sub-corpora according to the first language of the writer (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana). This categorization allows Contrastive Interlanguage Analysis among interlanguage varieties e.g., L1 Spanish - L1 Italian which was the primary goal of ICLE. Additionally it provides fruitful ground to investigate aspects of non-nativeness in learner essays which are usually revealed by the overuse or underuse of words or structures with respect to the target language norm. This investigation is done by means of a comparison between individual L2 sub-corpora and native English corpora e.g.L1 English - L1 Turkish. For the latter one ICLE incorporated the Louvain Corpus of Native English Essays (LOCNESS) (Granger et al., 2002), containing approximately 235,000 words coming from argumentative essays written by British and North American students.

TOEFL11 (Blanchard et al., 2013) consists of 12,100 essays written by L2 English learners during the TOEFL[3] college-entrance test. The essays are sampled as evenly as possible, 1,100 essays per language (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish). TOEFL11 contains essays of various topics of an average length 600 words. The environment under which L2 learners wrote the essays was strictly controlled. They were not allowed to use any dictionary and had 30 minutes at their disposal. The TOEFL11 dataset is the largest publicly available corpus of English written by non-native writers that is well-balanced for topic across L1. It also is the first such corpus that is annotated for score level. It includes three proficiency levels (low/medium/high) determined by trained human raters. So far, CIA has been used to determine the native language of L2 learners (e.g., Chinese Learners of English - Italian Learners of English).

Both ICLE and TOEFL11 are the most representative and most used learner corpora in the field of CIA. Even though both are publicly available they are not free of charge which makes it difficult to obtain. There exist a plethora of similar learner corpora (such as PELCRA and JPU) but most of them require a fee. The second language research community has made attempts to create similar but smaller corpora which are available for academic purposes without any fees. We continue introducing appropriate learner corpora for Contrastive Interlanguage Analysis which are distributed free of charge which we have used in our analysis.

The International Corpus Network of Asian Learners of English (ICNALE) constructed by Ishikawa (2011), is the first corpus that concentrates on Asian learners of English. Although ICLE includes essays written by learners with as many as sixteen L1s including Chinese and Japanese, the coverage of Asian learners is rather limited. ICNALE contains 1,306,660 words from 5,600 argumentative essays written by university students from eight countries and areas in Asia (China, Indonesia, Japan, Korea, Taiwan, Thailand, Pakistan, Philippines and Singapore). ICNALE includes essays written by NS in addition to those by NNS, which enables to conduct a more robust NS/ NNS comparison. Additionally, we can find measures of proficiency level of the students according to the Common European Framework of Reference for Languages (Council of Europe, 2001), determined by a standardized placement test. The learners are classified into four writing proficiency levels which are based on CEFR levels, A2(Waystage), B1_1(Threshold, Lower), B1_2(Threshold, Upper) and $B2^+$(Vantage or Higher). The placement of each essay in the respective category is being done based on its score and following TOEIC[4] scoring system, A2($<$500), B1_1($>=$ 500), B1_2($>=$600) and $B2^+$($>=$ 700). Table 6 gives insight regarding the distribution of essays across the different proficiency levels. This proficiency-based subdivision makes it possible to compare NNS at different L2 proficiency levels as well as NNS with different L1 backgrounds. Because all

[3] http://www.ets.org/toefl

[4] http://www.etscanada.ca/toeic

writers are college students, the proficiency level starts from A2 and not A1 which is the CEFR official starting level.

	A2	**B1**	**B2$^+$**	**Native (N)**
No of essays	960	3776	465	402
No of words	326,665	735,882	127,185	116,928

Table 6: Size of proficiency categories in ICNALE

University students were asked to write argumentative essays on two topics: "It is important for college students to have a part time job" and "Smoking should be completely banned at all the restaurants in the country" (see sample essay in Figure 1). Limiting the number of topics makes the content of the corpus much more lexically homogeneous, which enables us to conduct a robust comparison among different writer-groups (Ishikawa, 2011). Students had twenty to forty minutes at their disposal to write essays of length no more than 300 words without the use of a dictionary. The main goal of ICNALE is to provide data gathered under the same writing conditions. According to Ishikawa (2011), the control of writing conditions is crucial when compiling learner corpora, because it provides a clean and unbiased comparison among learner's writing behavior. For example, comparing a short essay which Japanese learners write with the help of a dictionary about smoking, and a much longer one which Chinese learners write without the help of a dictionary about leisure and hobby, can give some interesting difference. However, no one can say which of the essay length, writers' L1, dictionary use, or topic causes the differences.

When students enter college, they often find that they have much more free time than they did in high school. As a result, many students decide to apply for part-time jobs to earn extra money. In my opinion, it is important to for college students to work part-time jobs. I have two reasons for this. First, most college students do not have much money to spare for their hobbies and extracurricular activities. Hobbies and extra-curricular activities are vital to enriching the life as a college student. Through these activities, students can make new friends or discover their talents. However, if students did not have the money to enjoy these activities, they would be missing out on these significant chances, and they would not be able to make the best of their time in college. So, in order to prevent this, it is important for college students to earn money to spend on their pastime. Also, getting work-experience can help students understand the value of money. Understanding the value of money is not only important in college, but it is an important sense throughout a person's life. Without this sense, students would not have any idea on how to spend money, and they could waste great amounts of money without realizing how much damage it could cause to them. In order to have students get a sense of how important money is, they should understand how difficult it is to earn money. This is why I think that it is important for students to get work-experience while they are still in college.

Figure 1: ICNALE: sample essay

The Corpus of English Essays Written by Asian University Students (CEEAUS) (Ishikawa, 2010) is a learner corpus which consists of 242,538 tokens from 1,100 argumentative essays written by native speakers of English and learners of English from China and Japan. CEEAUS is the first version of ICNALE; thus the topics and conditions under which students wrote the argumentative essays are exactly the same. CEEAUS was originally a first attempt of Ishikawa (2010) to create a small corpus for interlanguage comparisons and consists of five modules: CEEJUS (Japanese university students' essays), CEECUS (Chinese university students' essays), CEENAS (English native speakers essays) and CJEJUS (Japanese essays written by Japanese university students). Unlike ICNALE, CEEAUS provides L2 proficiency categorization only in essays written by English learners from Japan (CEEJUS); thus we used only this part of the corpus (see sample essay in Figure 2). The proficiency categorization followed the Common European Framework of Reference for Languages standards determined by the TOEIC test. The learners were classified in the same categories as in ICNALE: A2(<500), B1_1(>= 500), B1_2(>=600) and B2^{+}(>= 700). Table 7 provides information regarding the essay proficiency group distribution in CEEAUS corpus

	A2	**B1**	**B2^{+}**	**Native (N)**
No of essays	82	340	348	146
No of words	17,580	85,614	66,460	50,468

Table 7: Size of proficiency categories in CEEAUS

I think we shouldn't make it a rule to prohibit smoking in a restaurant. Everyone has right to do what he or she wants as far as they don't harm other people. Many says the smoke from the cigarette harm other people. I think so too, but is that only the smoke? We use a car and it makes terrible noise or exhaust gas. Those harm other people too, however we don't have the idea that we should prohibit driving a car. What is the difference between these cases? I think we can't take them right to do what they want easily whatever they do. In deed I really hate smoking. The smoke makes my clothes smell and it feels me sick. For people like me, I think there should be some rules to avoid the smoke. Nowadays many restaurants separate the area the one is for non-smoker and the other is for smoker. This is very good idea. I think there are many opinions to this problem and there are many ideas to solve this problem. However only I can say is that we shouldn't make rules easily. So I disagree that smoking should be burned at all restaurants in Japan.

Figure 2: CEEAUS: sample essay

Gachon Learner Corpus 2.1 (GLC) (Carlstrom and Price, 2013) consists of 3.5 million words from 17,110 individual texts produced by Korean university students. The writers were asked to construct argumentative essays on a variety of topics such as child abuse, eating disorders, aggressive driving and family values. Unlike the other two corpora (CEEAUS and ICNALE) in GLC the writing environment under which students wrote essays was not strictly controlled. Meaning, there

was no time restriction ("unlimited") and the length of an essay was not regulated. Overall students wrote an average of 200 words per essay and most of them were categorized based on their writing proficiency (see sample essay in Figure 3). The essays were graded following the TOEIC scoring system and distributed in L2 proficiency categories based the CEFR standards (the resulting categories are the same as with CEEAUS and ICNALE corpora). Table 8 provides some details regarding the distribution of students' essay in the three proficiency groups

	A2	**B1**	**B2$^+$**
No of essays	10853	4470	1787
No of words	2,170,600	894,000	357,400

Table 8: Size of proficiency categories in GLC

GLC, unlike the other two corpora, provides meta data regarding learner's background information (such as gender and age) and links this information to the scripts in the corpus. This information provides a researcher with the means to focus on texts that match some particular predefined attributes. In this way, the researcher can create, if desired, a customized sub-corpus for the purposes of investigation. For example, a wide range of comparisons can be performed on the data, such as female vs male learners (Granger, 1998).

A number of L2 corpora has been created over the past few years to meet the needs of ESL materials designers. We will briefly mention two other large learner corpora: the Longman Learner Corpus (Summers, 1993) and the Cambridge Learner Corpus (Nicholls, 2003), both containing data from compositions written by L2 English learners with different L1s. These corpora are large, about 10 million words each, and consist of the writings of a wide variety of students learning English around the world. The data in these corpora are analyzed by lexicographers in order to improve the usefulness of dictionaries and course books for language learners. None of them is available for research since their use is restricted to the commercial creation of pedagogical material for ESL learners.

The development of large learner corpora is the result of creating large English normative corpora. The latter ones are essential in contrastive interlanguage analysis because it provides material to compare writing behavior and ability between L1 and L2 learners. Some of the existing learner corpora (such as ICNALE) include a sub-corpus of native speakers to make the comparison easier. However, in cases where a sub-corpus is not provided SLA researchers turn to two well known and accessible native corpora: the *Louvain Corpus Of Native English Essays* and the *British Academic Written English Corpus*.

I think most drivers in seoul are bad drivers. as i know, seoul has the highest population density in the world. which means there are too many people in seoul compare to other cities that has about the same amount of land. because of that, there are also too many cars on the road of the seoul and each driver gets annoyed by heavy traffic. So drivers tend to drive their cars aggressively. drivers often cut off, tailgate even though the signal is becoming red, do not use turn signal lamp, stop suddenly, do not keep or follow signs such as stop line, signs that says 'stop' or 'do not cut off in tunnel or on bridge' and so on. such things make me to think that most drivers in seoul are bad drivers.

Figure 3: GLC: sample essay

Louvain Corpus Of Native English Essays (LOCNESS) (Granger et al., 2002) is a corpus of argumentative essays comprising 324,304 words from 322 essays produced by native speakers of English (between 18 to 21 years old). The type of essays are similar to the essays produced by the learners taking part in the ICLE corpus project, and thus LOCNESS is the comparable corpus to ICLE (such as homosexuality, nuclear power, and equality of the sexes). It contains three sub-corpora: a British school A-level essays sub-corpus of 60,209 words 114 essays, a British university essays sub-corpus of 95,695 words and an American university essays sub-corpus of 168,400 words in 176 essays. For the purpose of this study we used the essays from American and British university students (see an example essay in Figure 4).

I believe that the public has a right to be informed about anything and everything that they want to be informed about, and people want to be informed about the death penalty; therefore, media should have access to report on executions. However, there access should be restricted to exclude any and all devices which could endanger security or safety of the people involved. This argument is debated often around times when the death penalty is actually put into effect. One such case which receive national attention was in California. The defendant, and person to be executed, was Robert Alton Harris.

Figure 4: LOCNESS: sample essay

British Academic Written English Corpus (BAWE) (Gardner and Nesi, 2012) contains 3,000 essays (6,506,995 words) written by native speakers of English. Produced and assessed as part of university degree coursework, and fairly evenly distributed across 35 university disciplines and four levels of study (first year undergraduate to Masters level). Texts consist of 500 to 5,000 words and have been categorized into 13 broad genre families, including essays, critiques, case studies, explanations, methodology recounts, problem questions and proposals. We used only the argumentative essays of this corpus which gave us 989,600 words from 1,237 documents (see an example essay in Figure 5).

The effects of gender and class on the way we speak is a question that has engaged much time with linguists and has also caught the interest of the general public as the bestselling success of books such as "Men are from Mars, Women are from Venus" shows. If asked, any person off the street could no doubt offer you an opinion on the way people of different genders and social classes speak. This is possibly because the effects of social class and gender on language are so easily noticeable and so easily stigmatized. It is also because of the huge effect these two social variables have on language. It is worth noting here before the discussion proper starts, an important distinction, that is the one between gender and sex. Whereas 'sex' is a purely biological concept, 'gender' is a social construct - an example for instance is what one might consider. This is important here as differences in the speech of men and women can be both down to biological differences (pitch differences for instance) or gender differences, but only one - gender - is relevant here.

Figure 5: BAWE: sample essay

It is clear from this chapter that there does not exist a standard compilation, and organization of a learner corpus. In fact, each of the corpora has been designed and created for different purposes. Since each corpus seeks to describe learner language in a way that suits the needs of the corresponding researchers, and learners, decisions have been made on an individual basis regarding the purpose of the corpus, the size of the corpus, and the accessibility of the corpus to outside researchers. It is important to note that, while a corpus has been designed and used for an explicit task, other researchers can use the corpus differently by performing their own specific analysis on the data. In Table 9 we present the design criteria of all corpora we use in this thesis.

Corpus	**Type of Essays**	**Origin of Participant**	**# of Essays**	**A2**	**B1**	**B2$^+$**	**Native Speakers (N)**
		Chinese	800	100	500	200	
		Indonesian	400	75	300	25	
		Japanese	800	308	457	35	
		Taiwanese	400	58	296	46	
		Korean	600	90	444	66	
ICNALE	Argumentative	Thai	800	240	550	10	
		Pakistani	400	40	350	10	
		Singapore	400	0	280	120	
		Filipino	400	40	276	84	
		English	402				402
	Argumentative	Japanese	770	82	340	348	
CEEAUS		English	146				146
GLC	Argumentative	Korean	17110	10853	4470	1787	
LOCNESS	Argumentative	American	176				167
BAWE	Argumentative	British	1238				1238

Table 9: Design criteria of ICNALE, CEEAUS, GLC, LOCNESS and BAWE learner corpora

Chapter 4

Feature Sets

Second Language Acquisition (SLA) researchers have identified grammatical, syntactic and lexical components which are related to second language (L2) writing development. Ellis (2008), Hinkel (2002) and Grant and Ginther (2000) claimed that the use of *past* and *present* tense by L2 writers increased across proficiency levels. Researchers also observed that skilled L2 writers make use of *passive* voice more than less skilled learners (Ellis, 2008, Kameen, 1980). Hinkel (2002) also discovered that advanced learners used more *progressive* and *perfective* aspect than beginners. Additionally, Grant and Ginther (2000) observed that as proficiency levels increased L2 writers incorporated more subordination. Finally Crossley et al. (2011) report that lexical indices can be used to predict the language proficiency levels of second language learners based on their writing samples.

We use these insights and study grammatical structures involving morphological and syntactic characteristics of verbs. In the matter of verb morphology the learner's ability to use *tense*, *aspect* and *voice* in finite and non finite verbs groups is examined. In terms of verb's syntactic function, the use of verbs in subordinate clause and their position in the syntax tree is captured. We identify each feature using Stanford Parser's (Klein and Manning, 2003) *syntactic trees*.

Syntax Trees: are based on the Phrase Structure formalism in which a set of rewrite rules are used to describe a given language's syntax and are closely associated with the early stage of transformational grammar (Chomsky, 1969). A sample set of grammar rules are given in Table 10. The symbols on the right side of the arrows are combined into the ones on the left. The resultant parse tree is a data structure originating from terminal nodes (the leaves) and concluding in the root node.

S	→	NP	VP
NP	→		PRP
NP	→	CD	NNS
VP	→	VBP	VP
VP	→	VBN	VP
VP	→	VBG	PP
PP	→	IN	NP

Table 10: Sample set of grammar rules

Consider the sentence *I have been smoking for ten years* to obtain a syntax tree we need to obtain the part of speech of each word. For this purpose we use the Penn Treebank II (Marcus et al., 1993) annotation which assigns meaningful tags to words (Part of Speech (POS) Tags) according to their role in the sentence. Mapping these symbols to the set of grammar rules of Table 10 will give the syntax tree of Figure 6

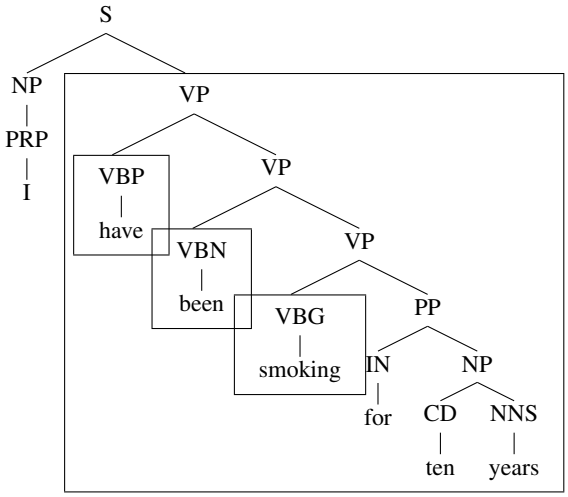

Figure 6: Sample syntax tree of : *I have been smoking for ten years.*

Additionally, this annotation system assigns tags to group of words (Constituents) most common constituents are the noun phrase (*NP*) and the verb phrase (*VP*). For example, in Figure 6 the constituent *VP* indicates that *have been smoking for years*, is the verb phrase of the sentence. In this study we focus only on the verb(s) of a verb phrase and we discard any complement object; thus we examine the verb group only. A verb group is part of a verb phrase and consists of optional auxiliaries, optional adverbs and a verb. For instance, in Figure 6 *have been smoking for ten years.* is the verb phrase of the sentence while, *have been smoking* is the verb group identified by the POS tags *VBP/have VBN/been VBG/smoking* (see Table 11 for POS tag definitions). We continue our feature presentation by providing their definition following examples regarding their implementation. Later in this chapter we will refer to some POS tags and constituents; thus in Table 11 we provide definitions and examples of those we will use.

Tag	Description	Example
VB	verb, base form	think
VBZ	verb, 3rd person singular present	she thinks
VBP	verb, non-3rd person singular present	I think
VBD	verb, past tense	they thought
VBN	verb, past participle	a sunken ship
VBG	verb, gerund or present participle	thinking is fun
SBAR	subordinating conjunction	
VP	verb phrase	

Table 11: Set of Penn Treebank POS tags and constituents analyzed

4.1 Verb Morphology

A sentence must have at least one finite verb and zero or more non-finite verbs. A finite verb must have *tense*, *aspect*, *voice* , where a non-finite verb has only *aspect*, *voice* and can take the form of *infinitive* (to visit), *present participle* (visiting) or *past participle* (visited).

(9) (a) It **is** harder for women **to lose** weight than men.
(b) Overprotective parents **raise** fearful children.

Tense commonly serves in natural language to anchor the situation described by the sentence to the time axis. There are three types of tense *present*, *past* and *future* (Comrie, 1985, Quirk et al., 1985). The finite verbs *is*, *raise* in (9a) and (9b) respectively are of *present* tense. But *to lose* does not have tense because it is nonfinite.

Grammatical Aspect refers to how an event or action is to be viewed with respect to time, rather than tense which positions an event to its actual location in time (Comrie, 1976, Quirk et al., 1985). Aspect is not inherently deictic, and it does not anchor the situation to the time axis. Aspect may however affect temporal structure (Comrie, 1976, Quirk et al., 1985). Quirk et al. (1985) defines three types of aspect *indefinite, perfective* and *progressive*. *Indefinite* aspect does not indicate whether the action is a complete action or a habitual action (*hate* in 10a and the nonfinite verb *to lose* in 9a are of indefinite aspect). *Perfective* indicates an action view as complete (*have met* in 10b) and *progressive* action view as incomplete (in 10c *was sleeping*). They syntactically define a fourth aspect *perfective progressive* (*have been talking* in 10d) which occurs when the perfective and progressive aspects are combined in the same verb group, the meaning associated with each of them is also combined .

(10) (a) I **hate** bad service.
(b) I think I **have met** him once before.
(c) I **was sleeping** all day yesterday.
(d) They **have been talking** for the last hour.

Voice is defined by Encyclopedia Britannica (2002) as a grammatical category that "indicates the relationship between the participants in a narrated event (subject, object) and the event itself". When the subject is the agent or doer of the action, the verb is in the *active* voice (Quirk et al., 1985). When the subject is the patient, target or undergoes the action, the verb is said to be in the *passive* voice (Quirk et al., 1985). In (11a), (11b) both *can be observed* and the nonfinite *being elected* are of passive voice. Where *was* (11a) has active voice.

(11) (a) **Being elected** by my peers **was** a great thrill.
(b) The aurora Borealis **can be observed** in the early morning hours.

Modal auxiliaries: A verb occasionally is preceded by modal auxiliary verbs: *can, could, may, might, shall, should, will, would* and *must*. These "helping" verbs do carry grammatical tense (modals are different from the other auxiliary verbs in that they shift the test into irrealis, thus "past" is not really past) but they do not appear in nonfinite form (Quirk et al., 1985). For example, *should* which is modal auxiliary does not have a participle form, *shoulding*. As a result a verb group that contains one of those modal auxiliaries is considered to be in finite form. According to Quirk et al. (1985) the modals *could, should, might* and *would* are the past tense forms of *can, shall, may* and *will* respectively (Quirk et al., 1985). We adapt this distinction and we associate these modal auxiliaries (with the exception of *will*) with their pseudo-tense form. Thus *could, should, might* and *would* are modal auxiliaries in *past tense form* (*modal past*) and *can, shall, may* in *present tense form* (*modal present*). For example in (11b) *can be observed* has a modal (*can*) in *present tense form*. If a modal or modal auxiliary does not fall into those two categories then it does not have any tense form.

We extract *tense*, *aspect* and *voice* using a tool created by Doandes (2003) for the CLaC laboratory. Doandes (2003) encoded the grammar rules from Quirk et al. (1985) and developed a set of grammatical patterns for both finite (Table 13) and nonfinite (Table 14) verb groups using the Penn Treebank annotation (Marcus et al., 1993). Additionally, she created two POS tag sets to include different forms of the auxiliaries *have* and *be*, useful to determine perfective aspect and passive voice. In Table 12 we report those sets along with the values they represent.

Have POS Tag	Value	Be POS Tag	Value
HAVE_VB	have	BE_VB	be
HAVE_VBD	had	BE_VBD	was/were
HAVE_VBG	having	BE_VBG	being
HAVE_VBZ	have/has	BE_VBN	been
HAVE_VBP	have/has	BE_VBP	am/are/is
		BE_VBZ	am/are/is

Table 12: POS tags for auxiliaries *have* and *be*

To illustrate how Doandes (2003)s' tool works, consider the sentence *I was asked to leave*

the restaurant in (12 a). The POS Tag sequence *VBD/was VBN/asked* corresponds to Doandes (2003)'s grammatical pattern *BE_VBD/was VBN/asked*. Matching this pattern in Table 13 (framed grammatical pattern) indicates that the finite verb group *was asked* has *past* tense, *indefinite* aspect and *passive* voice.

(12) (a) I was asked to leave the restaurant.
(b) PRP/I VBD/was VBN/asked TO/to VB/leave DET/the NN/restaurant
(c) PRP/I BE_VBD/was VBN/asked TO/to VB/leave DET/the NN/restaurant

Similarly the POS Tag sequence *TO/to VB/leave* is found in Table 14 (framed grammatical pattern) which attributes to the nonfinte verb group *to leave* the aspect *indefinite* and voice *active*.

Modal	Aux1	Aux2	Aux3	Verb	Voice	Tense	Aspect	Grammatical Pattern	ABCD Type
WILL_MD				VB	active	future	indefinite	will+VB	A
WILL_MD	HAVE_VB			VBN	active	future	perfect	will + have + VBN	AB
WILL_MD	HAVE_VB	BE_VBN		VBG	active	future	perfect progressive	will + have + been + VBG	ABC
WILL_MD	HAVE_VB	BE_VBN	BE_VBG	VBN	passive	future	perfect progressive	will+have+been+being+VB	ABCD
WILL_MD	HAVE_VB		BE_VBN	VBN	passive	future	perfect	will+have+been+VBN	ABD
WILL_MD		BE_VB		VBG	active	future	progressive	will+be+VBG	AC
WILL_MD		BE_VB	BE_VBG	VBN	passive	future	progressive	will+be+being+ VBN	ACD
WILL_MD			BE_VB	VBN	passive	future	indefinite	will+be+VBN	AD
	HAVE_VBD			VBN	active	past	perfect	had+VBN	B
	HAVE_VBP			VBN	active	present	perfect	havehas+VBN	B
	HAVE_VBZ			VBN	active	present	perfect	havehas+VBN	B
	HAVE_VBD	BE_VBN		VBG	active	past	perfect progressive	had+been+VBG	BC
	HAVE_VBP	BE_VBN		VBG	active	present	perfect progressive	have/has + been+ VBG	BC
	HAVE_VBZ	BE_VBN		VBG	active	present	perfect progressive	have/has + been+ VBG	BC
	HAVE_VBD	BE_VBN	BE_VBG	VBN	passive	past	perfect progressive	had+been+being+VBN	BCD
	HAVE_VBP	BE_VBN	BE_VBG	VBN	passive	present	perfect progressive	havehas + been + being + VBN	BCD
	HAVE_VBZ	BE_VBN	BE_VBG	VBN	passive	present	perfect progressive	havehas + been + being + VBN	BCD
	HAVE_VBD		BE_VBN	VBN	passive	past	perfect	had + been + VBN	BD
	HAVE_VBP		BE_VBN	VBN	passive	present	perfect	havehas + been + VBN	BD
	HAVE_VBZ		BE_VBN	VBN	passive	present	perfect	havehas + been + VBN	BD
		BE_VBD		VBG	active	past	progressive	was/were + VBG	C
		BE_VBP		VBG	active	present	progressive	am/are/is + VBG	C
		BE_VBZ		VBG	active	present	progressive	am/are/is + VBG	C
		BE_VBD	BE_VBG	VBN	passive	past	progressive	were/was + being + VBN	CD
		BE_VBP	BE_VBG	VBN	passive	present	progressive	am/are/is + being + VBN	CD
		BE_VBZ	BE_VBG	VBN	passive	present	progressive	am/are/is + being + VBN	CD
			BE_VBD	VBN	passive	past	indefinite	was/were + VBN	D
			BE_VBP	VBN	passive	present	indefinite	am/are/is + VBN	D
			BE_VBZ	VBN	passive	present	indefinite	am/are/is + VBN	D
				VBD	active	past	indefinite	VBD	
				VBP	active	present	indefinite	VBP	
				VBZ	active	present	indefinite	VBZ	

Table 13: Grammatical patterns for finite verb groups according to Doandes (2003). *A: modal. B: perfective, C: progressive, D:passive*

Aux1	Aux2	Aux3	Verb	Aspect	Voice	Grammatical Pattern	BCD Type
			VB	indefinite	active	(TO) + VB	simple
HAVE_VB			VBN	perfect	active	(TO) + have + VBN	B
HAVE_VB	BE_VBN		VBG	perfect progressive	active	(TO) + have + been + VBG	BC
HAVE_VB	BE_VBN	BE_VBG	VBN	perfect progressive	passive	(TO) + have + been + being +VB	BCD
HAVE_VB		BE_VBN	VBN	perfect	passive	(TO) + have + been + VBN	BD
		BE_VB	VBG	progressive	active	(TO) + be + VBG	C
	BE_VB	BE_VBG	VBN	progressive	passive	(TO) + be + being + VBN	CD
		BE_VB	VBN	indefinite	passive	(TO) + be + VBN	D
			VBG	indefinte	active	VBG	simple
HAVE_VBG			VBN	perfect	active	having + VBN	B
HAVE_VBG	BE_VBN		VBG	perfect progressive	active	having + been + VBG	BC
HAVE_VBG	BE_VBN	BE_VBG	VBN	perfect progressive	passive	having + been + being + VBN	BCD
HAVE_VBG		BE_VBN	VBN	perfect	passive	having + been + VBN	BD
	[____]	BE_VBG	VBN	progressive	passive	____+ being + VBN	CD
		[____]	VBN	indefinite	passive	____+ VBN	D

Table 14: Grammatical patterns for nonfinite verb groups, according to Doandes (2003). *B: perfective, C: progressive, D:passive*

Finally, when a sentence has a verb group with modal auxiliary, it is identified by the POS tag *MD*. For instance, the sentence of Figure 7 contains a verb group with a modal auxiliary *MD/can*. We assign the tense form of the modal to the verb group including it. In this case *can* is a modal auxiliary of present tense form (*modal present*).

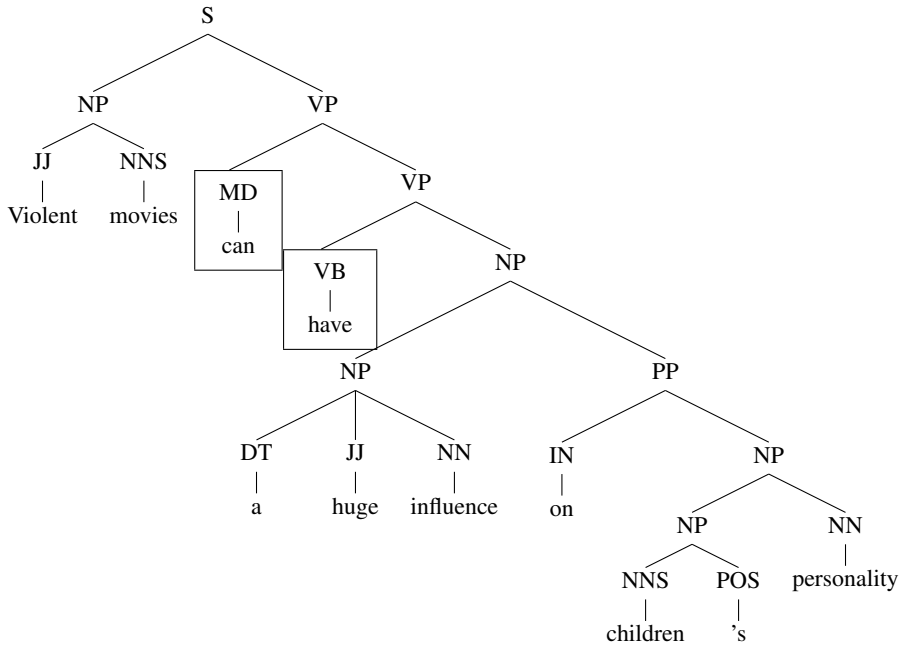

Figure 7: Verb group with modal auxiliary

4.2 Position of the Verb in the Parse Tree

Type of embedding describes whether a verb is part of an independent or subordinate (dependent) clause. A subordinate clause is considered to be an index of structural complexity in English. Syntactically, the clearest cases of subordination are those signaled by subordinating conjunctions (such as *after*, *although*, *because*, *that* and *when*). They serve not only to mark syntactic boundaries, but also to signal the functional relationship of the combined clauses to each other (Quirk et al., 1985). There are two ways to classify subordinate clauses structurally (finite, non-finite) or functionally (nominal, adverbial, comparative and comment). We follow the first approach and define two verb categories for dependent clauses: finite subordinate clause (*subC_fin*), nonfinite subordinate clause (*subC_nofin*) and one for independent clauses (*ind*)). For example, in (13) *want* is an independent clause verb (*ind*), *to eat* belongs to a nonfinite subordinate clause (*subC_nofin*) and *am stressed* to a finite dependent clause (*subC_fin*).

(13) When I **am stressed** I *want* **to eat**.

There are two main syntactic patterns that define subordinate clauses in a Penn Treebank-style parse tree (14a) and (14b). The framed subtrees of Figure 8 show these pattern distinctions. We use these patterns to identify the two subordinate categories *subC_fin* and *subC_nofin*.

(14) (a) SBAR → S → VP
(b) VP → S → VP

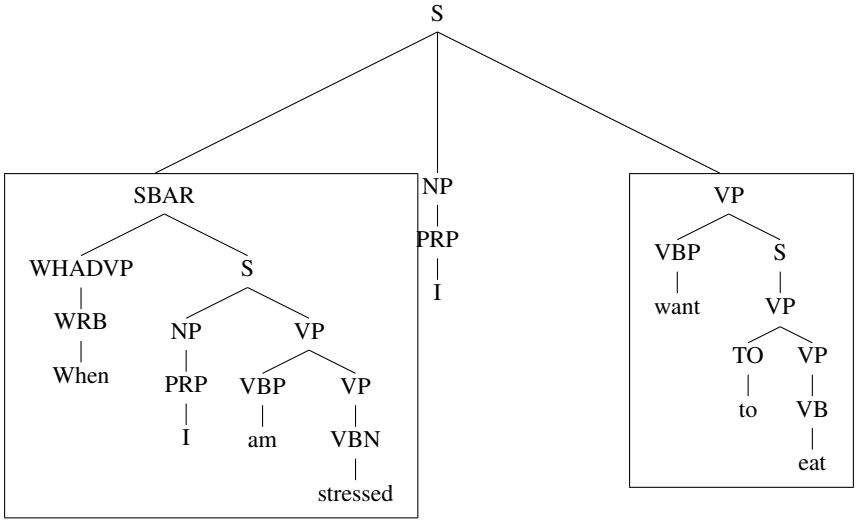

Figure 8: Parse tree illustrating finite and nonfinite subordinate clauses

Finite subordinate clause (subC_fin): A verb is of type *subC_fin* when it is finite and belongs to a subordinate clause identified by the pattern (14a).

Nonfintie subordinate clause (subC_nonf): A verb is of type *subC_nofin* when it is nonfintie and belongs to a subordinate clause identified by either pattern (14a) or (14b).

Independent clause (ind): A verb is of type **ind** when it does not belong to any subordinate clause and is finite.

For example, consider the sentence *When I am stressed I want to eat*. The syntax tree given by Stanford Parser (Klein and Manning, 2003) is depicted in in Figure 8. Doandes (2003)'s tool identifies two verb groups *am stressed* and *to eat*. The first one has tense *present*, aspect *indefinite*, voice *passive* and the second has no tense, aspect *indefinite* voice *active*. Starting with the first verb group *am stressed* we examine the parse tree (Figure 8) and we see it falls into the subordinate clause pattern SBAR $\rightarrow$ S $\rightarrow$ VP (see 14a). Additionally, Doandes (2003)'s tool assigns to this verb group *present* tense, indicating that it is a finite verb group. We assign then *am stressed* to the group of finite subordinate clauses (*subC_fin*).

The verb group *to eat* in the parse tree (Figure 8) falls under the subordinate clause pattern VP $\rightarrow$ S $\rightarrow$ VP (see 14b) and does not have a tense value (given by Doandes (2003)'s tool) which means that it is nonfinite. We assign *to eat* to the group of nonfinte subordinate clauses (*subC_nonfin*). Finally *want* does not belong to any subordinate clause pattern and it is a finite verb group since it has tense *present*. We assign *want* to the independent clause category (*ind*).

There are cases where the nonfinite subordinate clause is found under the pattern *SBAR* $\rightarrow$ *S* $\rightarrow$ *VP* (see 14b). For example, the verb group *accepted* in Figure 9 is a nonfinite verb group since it has no tense value (given by Doandes (2003)'s tool). Additionally is found under the constituent pattern *SBAR* $\rightarrow$ *S* $\rightarrow$ *VP* which indicates subordination. Thus we assign *accepted* to the category *subC_nonfin* which represents that a verb group is part of a nonfinite subordinate clause. Finally the second verb group of the sentence *will mean* belongs to the category *ind* (independent subordinate clause) because it is not included in any subordinate clause pattern.

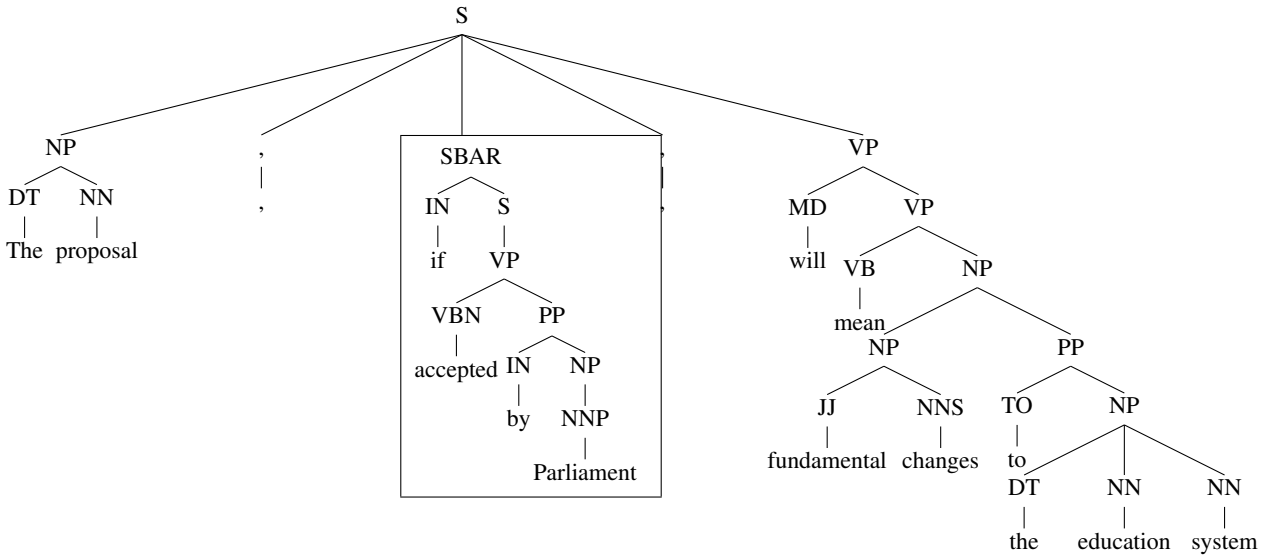

Figure 9: Parse tree illustrating verb group in nonfinite subordinate clause

Syntactic position (degree of embedding) of a verb group is determined by the dependent or independent clause containing it. Overall the depth of a syntax tree is considered to be a measure of sentence complexity. Generally, longer sentences are more complex syntactically but when sentences are of same length, the depth of their parse trees (syntax trees) can be indicative of increased complexity (Nenkova et al., 2010). We identify the syntactic position of a verb group by identifying its depth in the syntax tree including it. For instance, in Figure 10, the verb group *was working* has degree of embedding *2* because the verb phrase (VP) preceding has depth value *2*. Similarly the verb group *was* has degree of embedding *4*.

(15) (a) When I *was* student I *was working*.

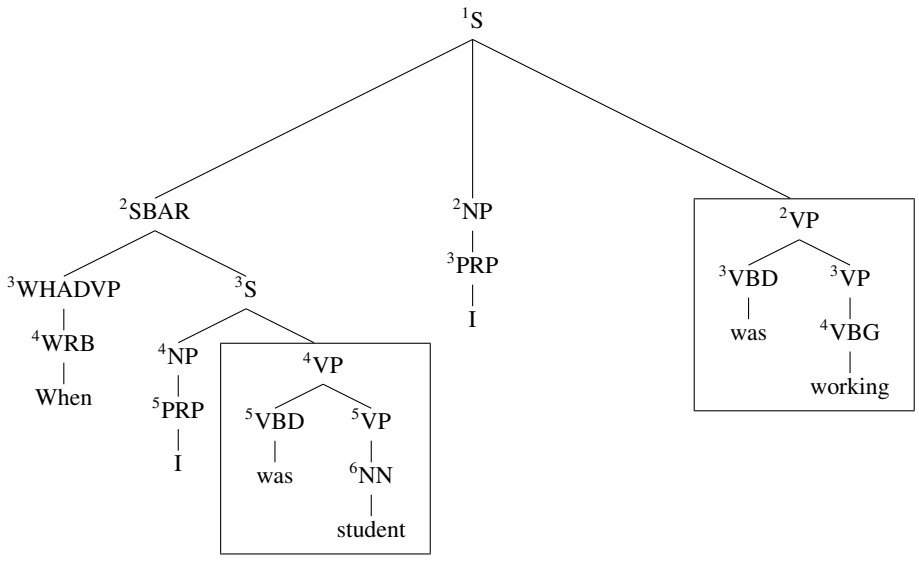

Figure 10: Parse tree illustrating degree of embedding

4.3 Word Level N-grams

Whether in first or second language, writers make use of specific lexical choices which reflect in part their proficiency (Hinkel, 2002, Meara et al., 2002). Especially, word n-grams have been frequently used as lexical features in the previous second language acquisition research (Heilman et al., 2006, Petersen and Ostendorf, 2009).

Word n-grams refer to one word or to the group of two or more continuous words that appear in text (Cavnar and Trenkle, 1994). For example, the word unigrams of (16) are: *Having, a, part-time, job, is,...*

(16) *Having a part-time job is beneficial for university students.*

For every learner's essay we extract word n-grams on the sentence level by removing any punctuation marks and create sequences of n words. As shown in the following example (17) we extract up to trigrams because of sentence length limitations (We found sentences with only 3 words e.g. *I hate smoking*). For every essay we obtain the most important word n-grams, in terms of distinguishing language proficiency, using Term Frequency-Inverse Document Frequency ($TFIDF$) (Salton and Buckley, 1988).

(17) (a) **unigrams=** {*Having, a, part-time, job, is,...*}
(b) **bigrams=** {*Having a, a part-time, part-time job,...*}
(c) **trigrams=** {*Having a part-time, a part-time job,...*}

Term Frequency (TF) refers to the number of times a particular word n-gram, wn_i, appears in an essay, e_j. The intuition is that an n-gram that occurs more frequently represents the essay better than an n-gram that occurs less frequently. However not all n-grams that occur more frequently in an essay are equally important. The effective importance of an n-gram also depends on how infrequent the term is in other essays and this is handled by Inverse Document Frequency (Salton and Buckley, 1988).

Inverse Document Frequency (IDF) represents the fact that a term which occurs in many essays is not a good discriminator, and should be given less weight than one which occurs in fewer essays. In mathematical terms, IDF is the log of the inverse probability of a term being found in any essay

$$IDF = \log \frac{N}{n_i} \tag{1}$$

where N is the number of essays in the corpus and n_i is the number of essays in which the word n-gram wn_i occurred.

TFIDF combines the weights of TF and IDF by multiplying them. TF gives more weight to a frequent n-gram in an essay and IDF downscales the weight if the n-gram occurs in many essays.

$$TFIDF = (1 + \log tf_{ij} * \log \frac{N}{n_i}) \tag{2}$$

For example assume that we are given four essays that correspond to the proficiency groups of English: *Beginners*, *Intermediate*, *Advanced*, *Native speakers*. We present in Table 15, the unigram frequencies of each essay, where each row corresponds to a word, each column corresponds to a proficiency group and the numbers represent the frequency of the corresponding words in each document. We observe that the word *consequently* occurs frequently only in *Intermediate* proficiency group. Thus our intuition is that this word can be indicative for distinguishing the four proficiency groups.

Term	Beginners	Intermediate	Advanced	Native Speakers
a	20	30	22	23
smoking	0	10	15	0
however	30	22	20	15
moreover	0	0	19	15
student	0	0	23	20
consequently	0	35	0	0

Table 15: Example of unigram frequency

Transforming the above term frequency matrix into a $TFIDF$ weight matrix we get the results of Table 16. These indicate that the word *consequently* is a good discriminator for *Intermediate* learners of English (TFIDF value 1.0). As a result we apply $TFIDF$ to each proficiency group to identify which word level n-grams can distinguish each level.

Term	**Beginners**	**Intermediate**	**Advanced**	**Native Speakers**
a	0	0	0	0
smoking	0	0.61	0.65	0
however	0	0	0	0
moreover	0	0	0.68	0.67
student	0	0	0.71	0.69
consequently	0	1	0	0

Table 16: Example of $TFIDF$ values for unigrams

4.4 Summary

In this chapter we addressed the features we use to determine the writing proficiency of second language learners of English. Focusing on the verb phrase we identify certain morphological and syntactic aspects. We analyze verb morphology in terms of *tense*, *aspect*, *voice* and its position in a parse tree by means of *type* and *degree of embedding*. We enhance our feature set by extracting word level n-grams using the information retrieval technique Term Frequency-Inverse Document Frequency (TFIDF) (Salton and Buckley, 1988). We present in the following Chapter 5 the necessary experimental foundations that will lead us to the results regarding the predictive potential of our verb group characteristics and word level n-grams.

Chapter 5

Experimental Setup

In order to determine the predictive potential of verb phrase characteristics (*tense*, *aspect*, *voice*, *type of embedding*, *degree of embedding*) and word level n-grams (unigrams, bigrams, trigrams) (discussed in Chapter 4) in writing proficiency we designed four experiments. We firstly, present the data sets used to perform our experiments and we continue by describing how we extracted our features from Second Language (L2) learners text. We used two approaches to represent learners' essays before given as input to the machine learning algorithms of our choice binary-based and frequency-based. We additionally, present two feature selection techniques we applied to determine the most relevant features in our experiments. Finally, we introduce the machine learning algorithm and evaluation metrics used.

5.1 Experiment Description

In order to gain an overview of our feature sets behavior in predicting writing proficiency we performed four experiments. The learner corpora used for each experiment are described bellow:

Asian L2 Learners-Same topics (ICNALE): For this experiment we used argumentative essays from the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011). ICNALE contains essays written by Asian learners on two topics. The essays are distributed across four language proficiency groups: A2 (beginners) 960 essays, B1 (intermediate) 3,776 essays, B2^{+} (advanced) 465 essays and Native Speakers (N) 402 essays. This is our first attempt to report how our feature sets behave when applied on limited topic essays (two topics) written by Asian non native speakers of English. We randomly selected 1,300 essays (324 from each class) for our training model and 320 essays (80 from each class) for testing.

Japanese L2 Learners-Same topics (Japanese): The reasoning behind this experiment is that we want to explore how our feature sets behave when we examine only Japanese L2 learners. We used essays written by Japanese second language learners of English collected from ICNALE (Ishikawa, 2011) and the Corpus of English Essays Written by Asian University Students (CEEAUS) (Ishikawa, 2010). Both ICNALE and CEEAUS contain argumentative essays written by Japanese learners of English on two topics. The essays are distributed across four language proficiency groups as follows: 390 essays for A2 (beginners), 491 for B1 (intermediate), 383 for $B2^+$ (advanced) and 508 for Native speakers of English (N). We randomly selected 1,227 essays (307 from each class) for our training model and 312 essays (78 from each class) for testing.

Korean L2 Learners-Different topics (GLC): Gachon Learner Corpus 2.1 (GLC) (Carlstrom and Price, 2013) is a learner corpus consisting of argumentative essays written by Korean university students on multiple topics. The essays are distributed in three proficiency groups A2 (beginners), 10,853 essays, B1 (intermediate), 4,470 essays and $B2^+$ (advanced) 1,787 essays. Since GLC does not have written samples of native speakers of English (as ICNALE and CEEAUS have) we gathered argumentative essays from Louvain Corpus Of Native English Essays (LOCNESS)(Granger et al., 2002) and British Academic Written English Corpus (BAWE)(Gardner and Nesi, 2012) corpora resulting in total of 1,414 essays. We randomly selected 4,484 essays (1121 from each class) for our training model and 1,171 essays (293 from each class) for testing.

Asian L2 Learners-Different topics (ALL): Our final experiment consists of essays from all five learner corpora (ICNALE, GLC, BAWE, LOCNESS and Japanese part of CEEAUS) in an attempt to report how our feature sets behave when we examine Asian second language learners on multiple topics essays. The essays distribution across the four proficiency groups is as follows: A2 (beginners) 11,895 essays, B1 (intermediate) 8,586 essays, $B2^+$ (advanced) 2,600 essays and Native Speakers (N) 1,962 essays. We randomly selected 6,280 essays (1,570 from each class) for our training model and 1,569 essays (392 from each class) for testing.

Name	**A2**	**B1**	**$B2^+$**	**N**
ICNALE	960	3,776	465	402
Japanese	390	491	383	508
GLC	10,853	4,470	1,787	1414
ALL	11,895	8,586	2,600	1,962

Table 17: Original essay distribution across corpora

Considering all four experiments we came across the issue of highly imbalanced data sets (see Table 17). In this study we are approaching the class-imbalance problem by adjusting second language learner essays to the minority class using random under-sampling. Random under-sampling

is a simple approach to re-sampling (Yap et al., 2014). Documents of the majority class are randomly eliminated until the ratio between the minority and majority class is at the desired level (see Table 18). In all four experiments we randomly selected essays for testing and training our classification model. In particular, in ALL we selected at random equal amounts of essays from each learner coporus to avoid any bias.

Name	Train Size	Test Size
ICNALE	1,300	320
Japanese	1,227	312
GLC	4,484	1,171
ALL	6,280	1,569

Table 18: Balanced essay distribution across corpora

5.2 Feature Extraction

For our verb phrase feature extraction we use an open source software for developing resources that process text, General Architecture for Text Engineering (GATE) [1] (Cunningham, 2002). GATE provides processing resources that we use for preprocessing our text. These are part of the ANNIE information extraction system (Cunningham et al., 2011), that has been developed using GATE. The preprocessing resources we use are listed bellow:

1. Tokenization: Break the text into individual tokens.
2. Sentence Splitting: Divide the text in a document into individual sentence units.
3. Part of Speech Tagging: Annotate each token with its corresponding lexical category (i.e verb, noun, punctuation. . .).

Verb Grouper: After preprocessing we apply Doandes (2003)'s processing resource (*Verb Grouper*) which is a GATE plug-in, to extract *tense*, *aspect* and *voice* of a verb group for each sentence of the text. This resource takes as an input the Part of Speech tags of every token and then iterates over every tag in the given sentence. When it finds a POS tag sequence that indicates the appearance of a verb group in the sentence, it matches a set of predefined grammatical patterns to identify the *tense*, *aspect* and *voice* of that verb group (defined in Chapter 4). A separate annotation set is created for each verb group, named *VC*. The features for any verb group annotation are its *tense*, *aspect*, *voice*.

[1] `https://gate.ac.uk/`

Verb Grouper Output:

(18) (a) *I [$_{present,indefinite,active}$ believe] that [$_{notense,indefinite,active}$ eating] healthy[$_{present,indefinite,active}$ is] important.*

(b) *Dole [$_{past,indefinite,passive}$ was defeated] by Clinton.*

Verb Subordinator: After identifying the verb groups we run Stanford Parser (Klein and Manning, 2003) to obtain the syntax tree for the sentence we examine. GATE (Cunningham et al., 2011) has a Stanford Parser plug-in that provides the parse tree of a sentence in a separate annotation set (*SyntaxTreeNode*). We use both Verb Grouper's Stanford Parser's (Klein and Manning, 2003) annotations to develop our GATE plug-in, *verb subordinator*, that indicates the type of embedding of a verb group. This plug-in goes over the Verb Grouper annotations then locates each verb group in the syntax tree. The output is three annotation sets that illustrate each verb group's type of embedding: finite subordinate clause, *subC_fin*, nonfinite subordinate clause, *subC_nonfin* and independent clause, *ind*. The feature for each annotation set is its type of subordination.

Verb Subordinator Output:

(19) (a) *I [$_{ind}$ believe] that [$_{subC\_nonfin}$ eating] healthy[$_{subC\_fin}$ is] important.*

(b) *Dole [$_{ind}$ was defeated] by Clinton.*

Verb Tree Depth: The last GATE plug-in we developed was to annotate a verb group's degree of embedding (level of embedding). Using Stanford Parser's (Klein and Manning, 2003) syntax tree we iterate though its nodes assigning each one of them their respective tree depth. Then using Verb Groupers' annotations we assign every verb group its corresponding tree depth. A separate annotation set is created for each verb group named VC_depth. The feature for each annotation is its degree of embedding.

Verb Tree Depth Output:

(20) (a) *I [$_2$ believe] that [$_6$ eating] healthy[$_5$ is] important.*

(b) *Dole [$_2$ was defeated] by Clinton.*

N-grams: In a final step before of data processing we create word level n-grams for each sentence of the document. We create three sets of n-grams, using Python[2] programming language, *unigram*, *bigram*, *trigram*. Each set represents one word and groups of two and three continuous words respectively. We treat each proficiency group as a single document and calculate the Term

[2]http://www.python.org/

Frequency-Inverse Document Frequency (TFIDF) (Salton and Buckley, 1988) for each word. With this approach we obtain knowledge regarding which n-grams can distinguish the four proficiency groups. The word level n-grams that are sent to WEKA as input are those that have TFIDF values above *0.0*. A *0.0* TFIDF value indicates that the specific n-gram appears frequently in all proficiency classes.

5.3 Classification Task

We performed text classification using the Java API of Weka[3] (Hall et al., 2009). To transform our documents into a representation suitable for Weka we firstly extracted from each text our features and presented them in two ways, frequency-based and binary-based. Additionally, we applied two feature selection techniques separately, Forward feature selection (Kohavi and John, 1997) and Information Gain (Mitchell, 1997). The machine learning algorithm of our choice was Support Vector Machine (Cortes and Vapnik, 1995).

5.3.1 Essay Representation

After extracting the values for all eight feature sets we represent an essay either by considering the frequency of each feature value (Frequency-Based representation) or by reporting the presence or absence of a feature value (Binary-Based representation). Other classification tasks not related to writing proficiency prediction have reported discrepancies between these two representations. Wu et al. (2013) experimented on both representations in an attempt to determine the native language of authors based on an essay written in a second language. They reported that binary-based representation was more successful than frequently-based. Similar results were obtained by Manevitz and Yousef (2002) where they classified newspaper articles based on their topics (finance, lifestyle ...). We believe that in order to have a complete overview on how our feature sets behave in writing proficiency prediction we need to consider both representations and report any discrepancies.

5.3.1.1 Frequency-Based Text Representation

In frequency-based text representation for every essay we count the number of times a feature value of a specific feature set appears in that essay (e.g number of times *present* tense occurs in text) divided by the number of times all feature values of that set appear in that essay (e.g. number of times *present*, *past*, *future*, *notense*, *modal present* and *modal past* tense appears in text)

[3] `http://www.cs.waikato.ac.nz/ml/weka/`

Example (21a) is sample essay that belongs to Class A (advanced learners), (21a) is sample essay that belongs to Class B (Beginner learners).

(21) (a) I think we shouldn't make it a rule to prohibit smoking in a restaurant. Everyone has the right to do whatever he or she wants as far as they don't harm other people. (Class A)
   i. tense frequencies: *present* 0.57, *notense* 0.28, *modal past* 0.14, *future* 0.0, *past* 0.0, *modal present* 0.0
   ii. feature vector: 0.57, 0.28, 0.14 , 0.0, 0.0, 0.0, A

(b) I think most drivers in Seoul are bad drivers. As far as I know, Seoul has the highest population density in the world. (Class B)
   i. tense frequencies: *present* 1.0, *notense* 0.0, *modal past* 0.0 , *future* 0.0, *past* 0.0, *modal present* 0.0
   ii. feature vector: 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, B

The corresponding file given to WEKA is:

@Relation
@Attribute present numeric
@Attribute notense numeric
@Attribute modal past numeric
@Attribute future numeric
@Attribute past numeric
@Attribute modal present numeric
@Attribute class{A,B}

@DATA
0.57, 0.28, 0.14 ,0.0, 0.0, 0.0, A
1.0, 0.0, 0.0, 0.0, 0.0, 0.0,B

Features are defined with "@ATTRIBUTE", their name and their type, and feature vectors are placed under the "@DATA", each row presenting a feature vector.

5.3.1.2 Binary-Based Text Representation

In binary representation, 1 indicates this feature value exists in the essay (i.e., *present* tense), 0 indicates this feature value is not in this essay.

Example (22a) is sample essay that belongs to Class A (advanced learners), (22a) is sample essay that belongs to Class B (Beginner learners).

(22) (a) I think we shouldn't make it a rule to prohibit smoking in a restaurant. Everyone has the right to do whatever he or she wants as far as they don't harm other people. (Class A)
   i. tense occurrence: *present* 1, *notense* 1, *modal past* 1, *future* 0, *past* 0, *modal present* 0
   ii. feature vector: 1, 1, 1, 0, 0, 0, A

(b) I think most drivers in Seoul are bad drivers. As far as I know, Seoul has the highest population density in the world. (Class B)
   i. tense occurrence: *present* 1, *notense* 0, *modal past* 0 , *future* 0, *past* 0, *modal present* 0
   ii. feature vector: 1, 0, 0, 0, 0, 0, 0, B

The corresponding file given to WEKA is:

```
@Relation
@Attribute present nominal
@Attribute notense nominal
@Attribute modal past nominal
@Attribute future nominal
@Attribute past nominal
@Attribute modal present nominal
@Attribute class{A,B}

@DATA
1, 1, 1, 0, 0, 0, A
1, 0, 0, 0, 0, 0, 0, B
```

Features are defined with "@ATTRIBUTE", their name and their type, and feature vectors are placed under the "@DATA", each row presenting a feature vector.

5.3.2 Feature Selection

Machine learning provides tools by which high dimensional data can be automatically analyzed. Feature selection is a technique that helps this process by identifying the most relevant features and removing irrelevant, redundant or noisy data (Guyon and Elisseeff, 2003). Irrelevant features are those that provide no useful information and redundant are those that do not add more information than the existing selected features on a particular dataset. The existing feature selection methods can be grouped into two categories: wrappers and filters. Wrappers evaluate subsets of features using the learning algorithm that is going to be applied to the data (Kohavi and John, 1997). Where filters select features by using heuristics based on general characteristics of the data (Kononenko, 1995). In this study, using WEKA, we apply two feature selection techniques forward feature selection and information gain, which correspond to wrappers and filter categories respectively.

5.3.2.1 Forward Feature Selection

Forward feature selection is a greedy method which searches through the space of feature subsets (Kohavi and John, 1997). It starts with a base set of (potential no) features and continues adding one feature at a time to a set of already selected features and checks how good that feature is by training and testing the classifiers on k cross-validation splits. The next best performing feature is then added to the set of selected features, and then the next iteration begins. It stops when the addition of any remaining features results in a decrease in evaluation.

Algorithm 1 Forward Feature Selection

Given: feature set $\{X_i,\ldots, X_n\}$, training set D, learning method SVM, G the set of best features
F $\leftarrow \{\ \}$
G $\leftarrow \{\ \}$

while score of F is improving **do**
  for i$\leftarrow$ 1 to n **do**
    if $X_i \ni$ F **then**
      $G_i \leftarrow$ F $\cup \{X_i\}$
      $Score_i$ = Evaluate(G_i, SVM, D)
    end if
    F $\leftarrow G_b$ with best $Score_b$
  end for
end while
return feature set F

5.3.2.2 Information Gain

Information Gain is a method that measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy (Mitchell, 1997). It can generalize to any number of classes (Mitchell, 1997). The entropy of a discrete random variable X with a probability distribution $p(x)$ is defined as:

$$H(p) = -\sum_{x \in X} p(x) \log p(x) \tag{3}$$

where $p(x)$ is the probability of a training example in the set x to be of the writing proficiency class. We want to determine which feature in our set of training feature vectors is most useful for discriminating among writing proficiency classes. Information gain tells us how important a given feature value of the feature vector is. We apply information gain using WEKA to decide the ordering of our features. For each feature we obtain a value between 0 and 1. 0 indicates that the feature value is not relevant in distinguishing writing proficiency.

5.3.3 Machine Learning Algorithm

The machine learning algorithm we choose is Support Vector Machine (SVM) (Cortes and Vapnik, 1995) which can be found in the WEKA API. SVM is a supervised learning model used for classification tasks. In a supervised machine learning algorithm the computer is presented with example inputs and their desired outputs and the goal is to learn a general rule that maps inputs to outputs (Mitchell, 1997). SVM is based on the concept of decision planes that define decision boundaries.

A decision plane is one that separates a set of objects having different class memberships. The goal of SVM is to design a hyper-plane that is as wide as possible that classifies correctly all the instances into their categories. A binary (i.e. two-class) classification problem is called linearly separable, if a hyper-plane can be positioned in such a way, that all instances of one class fall on one side and all instances of the other class fall on the other side. In case of two features, the hyper-plane corresponds simply to a line, in case of three features, the hyper-plane corresponds to an actual 2-D plane.

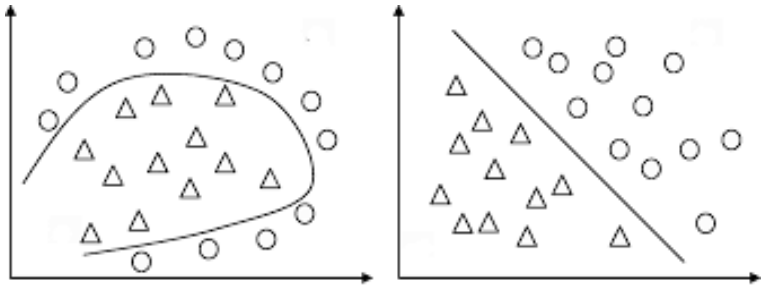

Figure 11: Non-linear vs linear problem

In classification problem where dimensional space is high (many features) it is not possible to find a hyper-plane that classifies the instances instantly; thus these problems are called non-linear. When the classification is non-linear, SVM non-linear solutions can be efficiently found by using the "kernel trick" (Aizerman et al., 1964): The data is mapped into a high-dimensional space in which the problem becomes linearly separable. The "trick" is that this is only "virtually" done by calculating kernel functions. This is the main advantage of SVM that can be independent of dimensionality in the feature space (Cortes and Vapnik, 1995).

5.4 Evaluation Metric

In this thesis we use ROC Area Under the Curve (AUC) as evaluation metric and we denote it as AUC. A ROC curve explicitly shows the tradeoff between the true positive rate and the false positive rate of a binary/multi classification system on different operating points, putting the class distribution and the misclassification costs out of the evaluation of classifiers' performance.

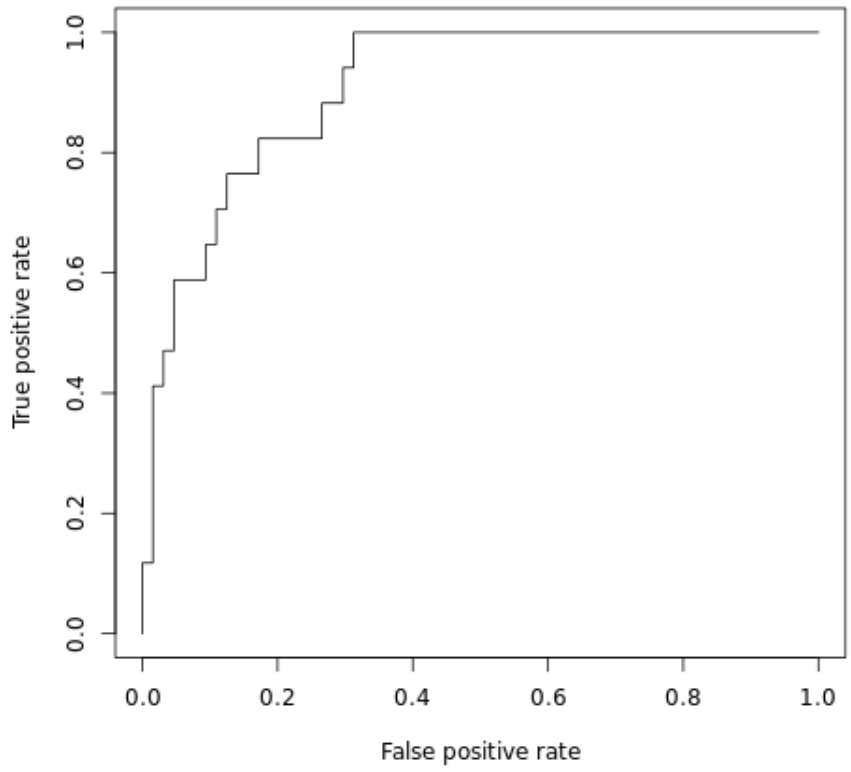

Figure 12: ROC Under Area Under the Curve

The ROC curve is constructed by depicting each essay as a single point considering its true positive and false positive rate (see Figure 12). The area under the ROC curve represents the AUC evaluation metric and it is calculated by splitting it into recognizable shapes such as trapezoid and rectangular and calculating their area. The sum of all the areas represent the AUC and corresponds to a single value between 0 and 1. Joachims (2005) supports that AUC is the optimal metric when support vector machines are used. We obtain AUC values using the build in function of WEKA. Table 19 illustrates the AUC output of Figure 12 as produced by WEKA given a classification task with classes A, B:

Class	True Positive Rate	False Positive Rate	AUC
A	0.98	0	0.99
B	0.94	0.03	0.952

Table 19: Weka AUC output for two classes, *A*, *B*

We described the foundations of our text classification tasks in order to determine the predictive potential of our verb phrase characteristics (*tense*, *aspect*, *voice*, *type of embedding*, *degree of embedding*) and word level n-grams (unigrams, bigrams, trigrams) in writing proficiency. We continue by presenting the results of each experiment and providing a detailed analysis of their performance.

Chapter 6

Results

Second language acquisition research has devoted resources so as to find linguistic features that are related to writing proficiency of non native speakers of English. The ultimate goal is to apply those indices in systems that can automatically evaluate writing proficiency. We present the relation of word level n-grams (*unigrams*, *bigrams*, *trigrams*) and verb group characteristics (*tense*, *voice*, *aspect*, *type* and *degree of embedding*) to writing proficiency in terms of performance contribution. Having already reported the relation of some of those features (discussed in Chapter 2) to writing proficiency we perform our own text classification tasks aiming to record how this basic linguistic features behave when considered in automatic writing proficiency prediction on the four data sets introduced in Section 5.1.

6.1 General Remarks

As described in Section 5.3.1, we represented the argumentative essays for our text classification tasks using two approaches binary-based and frequency-based. The individual feature results show that either considering binary or frequency document representation, the performance differences among all experiments are small. This illustrates a uniform distribution of our feature sets across the learner corpora. We continue our investigation further aiming to report the predictive potential of their combinations, as well as, how they behave regarding corpora with different design criteria.

Feature Set	Approach	ICNALE	Japanese	GLC	ALL
	Binary	0.563	0.537	0.607	0.589
Tense	Frequency	0.474	0.402	0.579	0.535
	Binary	0.547	0.562	0.604	0.592
Aspect	Frequency	0.490	0.467	0.575	0.493
	Binary	0.517	0.549	0.568	0.550
Voice	Frequency	0.456	0.484	0.495	0.487
	Binary	0.512	0.500	0.532	0.516
Type of Embedding	Frequency	0.455	0.449	0.460	0.493
	Binary	0.551	0.556	0.639	0.612
Degree of Embedding	Frequency	0.468	0.451	0.590	0.590
	Binary	0.560	0.556	0.504	0.501
Unigrams	Frequency	0.455	0.493	0.462	0.476
	Binary	0.538	0.531	0.516	0.506
Bigrams	Frequency	0.472	0.453	0.451	0.472
	Binary	0.504	0.531	0.518	0.513
Trigrams	Frequency	0.449	0.452	0.445	0.441

Table 20: Frequency-based vs. Binary-based performance across all experiments

Our data characteristics are captured better with the binary format probably because of document brevity. Our learner corpora are too homogeneous in terms of vocabulary and grammar usage as a result frequency values are very close to each other (in other terms look "too continuous"). The support vector machine has difficulty classifying those values; thus a binary approach with clear *0*, *1* entries is easier for the classifier to leverage.

Additionally, we compared two different feature selection techniques Information Gain (Mitchell, 1997) and Forward Feature selection (Kohavi and John, 1997), which were presented in Section 5.3.2. Both approaches provided similar results across the four experiments which indicates that they capture the same redundant or irrelevant features. The only difference regards the processing time of our data, information gain was much faster than forward selection. Given this observation we continue our analysis by presenting results on our different experiments considering the output from information gain.

6.2 Detailed Analysis

We report the predictive potential of our feature sets in writing proficiency when applied on data sets with different design criteria. In this section we analyze four experiments (as described in Section

5.1) categorized based on author's cultural background and topic variety. Our goal is to compare the behavior of our features individually or in combination (total of 256 feature combinations) across these four experiments.

An overview of our results across all experiments (see Table 21, where *ngrams* indicate *unigram bigram trigram*, *uni*, *bi*, *tri* correspond to *unigram*, *bigram* and *trigram* respectively) indicate two main observations: Firstly, our features sets behave differently when applied to our corpora. Even when applying our feature sets to data sets that share the same design criteria, like ICNALE and Japanese, we obtained different results. However, we noticed a consistency in all four attempts which includes the combination of *tense* and *aspect* being present in most best performing feature combinations.

ICNALE	**AUC**	**Japanese**	**AUC**	**GLC**	**AUC**	**ALL**	**AUC**
tense voice aspect const uni bi	0.596	voice aspect	0.593	tense aspect const tri	0.647	de	0.613
tense voice aspect const uni tri	0.594	voice aspect const	0.593	tense aspect const uni	0.646	tense aspect const	0.612
tense voice aspect ngrams	0.594	voice aspect const de tris	0.580	tense aspect uni tri	0.645	const de	0.611
tense voice aspect uni bi	0.594	voice aspect de tris	0.580	tense aspect ngrams	0.645	const de uni	0.611
tense voice aspect uni tri	0.592	tense de	0.574	tense aspect tri	0.645	tense de uni tri	0.610
tense voice aspect uni	0.588	voice de tri	0.568	tense aspect bi tri	0.645	tense de tri	0.610
tense voice aspect const uni	0.588	voice const de	0.568	tense aspect const uni bi	0.644	tense de bi tri	0.610
tense aspect ngrams	0.584	voice const de tri	0.568	de uni	0.644	de uni	0.610
tense aspect const ngrams	0.584	voice de	0.568	tense aspect const bi	0.644	const de tri	0.608
tense aspect const uni bi	0.584	aspect de tri	0.562	tense aspect uni	0.643	const de uni tri	0.608
tense aspect uni bi	0.584	aspect const de tri	0.562	tense aspect uni bi	0.641	voice de uni	0.607
tense aspect uni	0.584	tense voice aspect const de tri	0.562	tense aspect bi	0.641	tense voice aspect const	0.607
tense aspect const uni	0.584	tense voice aspect de tri	0.562	tense voice aspect tri	0.640	tense de ngrams	0.607
tense aspect const bi tri	0.583	tense voice aspect de uni	0.562	tense voice aspect ngrams	0.639	tense voice aspect	0.607
tense aspect bi tri	0.581	tense voice aspect	0.562	de	0.639	const de ngrams	0.606

Table 21: Best performing feature combinations across four corpora

Obtaining the performance of each feature combination is not enough because just a numerical value is not indicative of the potential of our features. We want to know exactly how our model classifies our data across the four proficiency groups. For this purpose, we analyze each confusion matrix by reporting the proficiency groups in which we have a higher rate of correctly essay classifications. From our experience as readers of second language learners' essays, we believe that it is easier to distinguish the two extreme proficiency classes beginners (A2) and native speakers of English (N). This intuition is validated by looking at the confusion matrices of all experiments (see Table 22). Thus we use the confusion matrices to validate our intuitions and obtain a better insight regarding how our features behave when relating them to each proficiency group.

	ICNALE tense voice aspect const unigram bigram				Japanese voice aspect				GLC tense aspect constituent trigram				ALL de			
Proficiency class	**A2**	**B1**	**B2$^+$**	**N**	**A2**	**B1**	**B2$^+$**	**N**	**A2**	**B1**	**B2$^+$**	**N**	**A2**	**B1**	**B2$^+$**	**N**
A2	**65**	0	10	5	**60**	5	1	14	**126**	40	102	24	**128**	97	67	99
B1	6	0	32	38	50	5	8	17	93	73	111	16	136	92	67	96
B2+	20	0	**50**	10	28	9	25	19	96	40	**132**	25	60	75	**100**	113
N	1	0	4	**75**	0	2	3	**75**	17	10	16	**250**	8	18	17	**350**

Table 22: Confusion matrices of our experiments

6.2.1 Asian L2 Learners-Same Topics (ICNALE)

We present how verb characteristics and word level n-grams behave in writing proficiency detection using argumentative essays from Asian learners of English on two topics. As described in Section 5.1 the essays were distributed across four language proficiency groups: A2 (beginners) 960 essays, B1 (intermediate) 3,776 essays, B2$^+$ (advanced) 465 essays and Native Speakers 402 essays. Their results when run individually show that *tense* is the outperforming feature followed by *unigram*. Their AUC performance is not high enough so as to distinguish writing proficiency just by themselves. Nevertheless, it is an indication of their relation to second language proficiency.

Feature Set	**AUC**
tense	0.563
unigrams	0.560
degree of embedding	0.551
aspect	0.547
bigrams	0.538
voice	0.517
type of embedding (constituent)	0.512
trigrams	0.504

Table 23: Individual feature performance in ICNALE

In general, we obtain better performance when we combine all our feature sets (yielding 256 combinations). We get the highest result when we combine all of our features excluding *trigrams* and *degree of embedding* (AUC 0.596). Features such as *voice* that individually perform low (AUC 0.517) when considered with the rest of our indices produce better results. A fact that illustrates we gain greater knowledge about our characteristics and their relation to writing proficiency by combining them rather than considering them alone. Moreover, we observe that the occurrence of *tense aspect* and *unigrams* are present in most feature combinations. Especially the prevalence of *tense* coincides with Bardovi-Harlig and Reynolds (1995)'s argument that it can be a distinguishing factor in second language writing proficiency.

Feature Set	AUC
tense voice aspect constituent unigram bigram	0.596
tense voice aspect constituent unigram trigram	0.594
tense voice aspect unigram bigram trigram	0.594
tense voice aspect unigram bigram	0.594
tense voice aspect unigram trigram	0.592
tense voice aspect unigram	0.588
tense voice aspect constituent unigram	0.588
tense aspect unigram bigram trigram	0.584
tense aspect constituent unigram bigram trigram	0.584
tense aspect constituent unigram bigram	0.584
tense aspect unigram bigram	0.584
tense aspect unigram	0.584
tense aspect constituent unigram	0.584
tense aspect constituent bigram trigram	0.583
tense aspect bigram trigram	0.583
tense unigram bigram	0.581
tense aspect constituent bigram	0.581
tense aspect bigram	0.581

Table 24: Best performing feature combinations in ICNALE

Considering the confusion matrix of the outperforming feature combination *tense voice aspect constituent unigram bigram* we notice that essays of native speakers (N), advanced ($B2^+$) and beginner (A2) proficiency groups are categorized better than intermediate (B1). The distinguishing power of this feature combination across the different proficiency groups is also illustrated through AUC performance of each class: A2 0.604, B1 0.497, $B2^+$ 0.585 and N 0.701. Although our results are not high the fact that we obtain proper classification in three proficiency groups indicates that this combo is a promising baseline for predicting writing proficiency.

	Classification output			
Proficiency class	**A2**	**B1**	**$B2^+$**	**N**
A2	**65**	0	10	5
B1	6	0	32	38
B2+	20	0	**50**	10
N	1	0	4	**75**

Table 25: Confusion matrix for feature combination *tense voice aspect constituent unigram bigram*

Applying our features to essays written by Asian second language learners of English gave as the insight that *tense*, *aspect* and *unigrams* can play a role in predicting writing proficiency. We examine further how *tense* and in general all of our feature sets behave when we apply them in essays written only by Japanese learners of English.

6.2.2 Japanese L2 Learners-Same Topics (Japanese)

In this experiment we present the behavior of our feature sets, individually and in combination, in order to predict the writing proficiency of Japanese learners of English on the same topic essays. As described in Section 5.1 the essays were distributed across four language proficiency groups: 390 essays for A2 (beginners), 491 for B1 (intermediate), 383 for $B2^{+}$ (advanced) and 508 for Native speakers of English (N). The individual performance of our feature sets indicates that *aspect* is the best performing feature followed by *degree of embedding (de)* and *word level unigrams*. *Tense* on the other hand ranks much lower in this experiment. Comparing these results with the ones obtained in Section 6.2.1 we notice that *tense* when examining Asian L2 learners is the best performing individual feature (AUC 0.563) but this is not the case with Japanese L2 learners. Another difference involves the feature *aspect*. For Asian L2 learners it ranks fourth (AUC 0.547) among the seven feature sets, but in this experiment it ranks first (AUC 0.562)

Feature Set	**AUC**
aspect	0.562
degree of embedding (de)	0.556
unigrams	0.556
voice	0.549
tense	0.537
trigrams	0.531
bigrams	0.531
type of embedding (constituent)	0.500

Table 26: Individual feature performance for Japanese L2 learners

We combine our feature sets (resulting in 256 combinations) and we notice *voice aspect* to be the outperforming combo (AUC 0.593). In general our verb group characteristics are more prevalent than word level n-grams in this experiment. Scrutinizing our data set we realized that Japanese learners have the tendency to use the same n-grams across all proficiency levels (such as *Japanese should*, *as in Japan*, *my opinion is*). As a result, our lexical indices are outperformed by the verb group characteristics. Additionally, we notice combinations such as *voice aspect*, *tense de*, *tense voice aspect* or even the individual feature *aspect* to consume most of the best performing positions. This behavior is different than the one described in Section 6.2.1, where verb group characteristics appear always in combination with word level n-grams. The only difference between this experiment and the one in Section 6.2.1 is the language of origin of L2 learners. Here we focused only in Japanese learners of English where in 6.2.1 we have Asian students from eight different countries including Japan. Thus we could partially attribute the low performance of *tense* and the absence of *n-grams* on the country of origin of L2 authors.

Feature Set	AUC
voice aspect	0.593
voice aspect constituent	0.593
voice aspect constituent de trigrams	0.580
voice aspect de trigrams	0.580
tense de	0.574
voice de trigram	0.568
voice constituent de	0.568
voice constituent de trigram	0.568
voice de	0.568
aspect de trigram	0.562
aspect constituent de trigram	0.562
tense voice aspect constituent de trigram	0.562
tense voice aspect de trigram	0.562
tense voice aspect de unigram	0.562
tense voice aspect	0.562
tense voice aspect constituent de unigram	0.562
aspect	0.562
aspect unigram	0.562
constituent de	0.556

Table 27: Best performing results for Japanese L2 learners

When we examine the confusion matrix of *voice aspect* (the top feature combo in this experiment) we notice that the best proficiency group distinction occurs between beginners (A2) and native speakers of English (N). Where intermediate (B1) and advanced ($B2^+$) are completely misclassified. This is also depicted in the AUC performance of each proficiency group: A2 0.611, B1 0.495, $B2^+$ 0.525 and N 0.673. This trend is similar in every feature combination, which indicates that our features when it comes to examining Japanese learners of English (on essays in two topics) cannot act as a distinguishing factor between intermediate and advanced learners. This is different from the results we obtain in Section 6.2.1, where the top performing combo *tense voice aspect constituent unigram bigram* was classifying appropriately most of essays in A1, $B2^+$ and N classes. Although their AUC performance is not that different (*voice aspect* AUC 0.593 and *tense voice aspect constituent unigram bigram* AUC 0.596) we notice how different our feature sets behave in those two experiments.

	Classification Output			
Proficiency Class	**A2**	**B1**	**$B2^+$**	**N**
A2	**60**	5	1	14
B1	50	5	8	17
B2+	28	9	25	19
N	0	2	3	**75**

Table 28: Confusion matrix of feature combination *voice aspect*

Overall in this experiment we notice the occurrence of *voice* and *aspect* and the general prevalence of verb group characteristics in most feature combinations. However *tense* by itself is not as predictive as it was in Section 6.2.1 which we partially attribute to the origin of L2 writers. In this study we focus only on Japanese where in Section 6.2.1 we examined Asian L2 learners in general. We continue our exploration by shifting our focus from these data sets that overlap to a completely different learner corpora so at to observe how our feature sets behave under different data criteria.

6.2.3 Korean L2 Learners-Different Topics (GLC)

We present how our verb group features and word level n-grams behave when we examine essays written by Korean learners of English in multiple topics. As described in Section 5.1 the essays are distributed across four language proficiency groups: A2 (beginners) 10,853 essays, B1 (intermediate) 4,470 essays, $B2^{+}$ (advanced) 1,787 essays and Native Speakers (N) 1,414 essays. *Degree of embedding*, *tense* and *aspect* give the best performance when examining all feature sets individually. Additionally, we notice that all word level n-grams acquire the lowest rankings among the seven feature sets.

Feature	**AUC**
degree of embedding (de)	0.639
tense	0.607
aspect	0.604
voice	0.568
type of embedding (constituent)	0.532
trigram	0.518
bigram	0.516
unigram	0.504

Table 29: Individual feature performance in GLC

When it comes to combine our feature types we notice that *tense* and *aspect* are part of most feature combinations. But we do not notice a consistency regarding which word level n-grams participate more. For once more we highlight the importance of observing how our features behave in consolidation. Since it is evident we obtain better results when we combine them.

Feature	AUC
tense aspect constituent trigram	0.647
tense aspect constituent unigram	0.646
tense aspect unigram trigram	0.645
tense aspect unigram bigram trigram	0.645
tense aspect trigram	0.645
tense aspect bigram trigram	0.645
tense aspect constituent unigram bigram	0.644
de unigram	0.644
tense aspect constituent bigram	0.644
tense aspect unigram	0.643
tense aspect unigram bigram	0.641
tense aspect bigram	0.641
tense voice aspect trigram	0.640
tense voice aspect unigram bigram trigram	0.639
de	0.639
tense voice aspect bigram trigram	0.638
voice de	0.638
tense bigram trigram	0.638
tense trigram	0.637
tense voice aspect unigram trigram	0.636

Table 30: Best performing feature combinations in GLC

Additionally, we note that degree of embedding (*de*) is the only individual feature among the seven that performs well by itself. By eyeballing the data we notice that in Korean corpus writers construct sentences by overusing the embedding structure. However, this is not the case in Japanese learners' essays where they express themselves with simpler structures. For example, in (23) we present an essay from Korean corpus and in (24) an essay from the Japanese experiment. Both sentences were taken from the same proficiency group, B2$^+$. The numbers attached to each verb group indicate its degree of embedding. We notice that Korean second language learners use the embedding structure more than Japanese. From our perspective this may be related to either the design criteria of the corpora or their language of origin. As described in Section 3.4 the writing conditions under which Korean second language learners wrote essays were not strictly controlled. Meaning there was a time restriction (one hour) but the length of an essay was not regulated. However, Japanese students required to write essays of maximum 300 words in less than forty minutes.

(23) *But when adults are asked*^{4} *to do*^{7} *the same thing, they typically get*^{2} *nervous and refuse*^{2} *to even try*^{5}*, claiming*^{8} *that they have*^{11} *no talent.*

(24) *I think*^{2} *smoking should be completely banned*^{5} *at all restaurants in Japan.*

Additionally, a combination that catches our attention and involves *degree of embedding (de)* is *voice de*. This combo appears also in Section 6.2.2 where we analyzed Japanese L2 learners' essays. Coming across *voice de* we observe relation between those two verb group features that we would

not have come across if we did not combine all our features. To make our reasoning clearer we note that *tense*, *aspect* appear together is in the top in all three experiments. However this combination is not surprising to us since those features are introduced together in every grammar book. However, when *voice*, a grammatical feature, and *de*, a feature expressing complexity appear together it gives us a feature combination that was not anticipated.

From the confusion matrix of best performing feature combo *tense aspect constituent trigram*, we get better classification of native speakers of English (N) followed by advanced ($B2^+$) and beginners (A2) L2 learners. An addition to this observation is the AUC performance of each proficiency group: A2 0.581, B1 0.573, $B2^+$ 0.589 and N 0.847. As in Section 6.2.1 in this experiment also our performance is not high enough to consider this combination alone for writing proficiency detection. However the class distribution indicates that it is a promising baseline.

	Classification Output			
Proficiency Class	**A2**	**B1**	**$B2^+$**	**N**
A2	**126**	40	102	24
B1	93	73	111	16
B2+	96	40	**132**	25
N	17	10	16	**250**

Table 31: Confusion matrix of feature combination *tense aspect constituent trigram*

Overall considering essays written by Korean second language learners of English we observe that *tense* by itself is not predictive of writing proficiency. We also observe confirmation of the intuitive assumption that *tense* and *aspect* are partially predictive of learner proficiency in this corpus. We continue our analysis by performing a final experiment when we combine all three data sets together so as to have a complete overview of how our features behave without considering country of origin or topic homogeneity.

6.2.4 Asian L2 Learners-Different Topics (ALL)

The final experiment of this study includes combining all essays from ICNALE (Ishikawa, 2011), CEEAUS (Ishikawa, 2010) and GLC (Carlstrom and Price, 2013) corpora into one data set. Thus we present how our feature sets behave without considering L2 learners' cultural background or topic homogeneity. As described in 5.1 the essay distribution across the four proficiency groups is as follows : A2 (beginners) 11,895 essays, B1 (intermediate) 8,586 essays, $B2^+$ (advanced) 2,600 essays and Native Speakers (N) 1,962 essays. The results of examining our feature sets individually indicate that our verb group features outperform our lexical indices. *Degree of embedding (de)*, surpasses in performance the individual feature sets with *aspect* and *tense* following with very small differences in value.

Feature	AUC
degree of embedding (de)	0.613
aspect	0.592
tense	0.589
voice	0.550
type of embedding (constituent)	0.516
trigram	0.513
bigram	0.506
unigram	0.501

Table 32: Individual feature performance in ALL

Overall in the other three experiments when examining features individually *degree of embedding* was always ranking first to third position. However, it is the first time that we come across *degree of embedding* as the top feature among 256 combinations. The confusion matrix of *degree of embedding* shows that essays written by native speakers (N) and beginners (A2) are classified better than intermediate (B1) and advanced ($B2^+$). Especially, intermediate learners' essays are mainly misconceived as beginners and advanced as native speakers. Investigating further essays from all proficiency groups we notice a similarity in writing style between A2 and B1, as well as, $B2^+$ and N proficiency group, which results in categorizing them wrongly. The AUC values for each class verify the above observation: A2, 0.663, B1 0.523, $B2^+$ 0.599, N 0.818

	Classification Output			
Proficiency Class	**A2**	**B1**	**$B2^+$**	**N**
A2	**128**	97	67	99
B1	136	92	67	96
B2+	60	75	100	113
N	8	18	17	**350**

Table 33: Confusion matrix of feature *de*

In addition, the individual results show that *tense* by itself is not a predictor of writing proficiency, nevertheless when combined with the rest of our features is part of most best performing combos. For example *tense* by itself has a low performance (0.589) but when combined with *aspect* and *type of embedding (constituent)* which also have low individual performances (0.592 and 0.516 respectively) reach an AUC of 0.612. *Tense* is a shallow but rather complex feature. Its predictive potential depends on many factors such as cultural background and genre; thus giving a crisp answer (yes or no) whether or not is related to writing proficiency is not straight forward.

Feature	AUC
de	0.613
tense aspect constituent	0.612
constituent de	0.611
constituent de unigram	0.611
tense de unigram trigram	0.610
tense de trigram	0.610
tense de bigram trigram	0.610
de unigram	0.610
constituent de trigram	0.608
constituent de unigram trigram	0.608
voice de unigram	0.607
tense voice aspect constituent	0.607
tense de unigram bigram trigram	0.607
tense voice aspect	0.607
constituent de unigram bigram trigram	0.606
voice de unigram bigram	0.606
constituent de bigram trigram	0.606
constituent de bigram	0.606
constituent de unigram bigram	0.606
voice de bigram	0.605

Table 34: Best performing feature combinations in ALL

Another verb phrase characteristic that catches our attention is the type of subordination (*constituent*). In the literature is noted that subordination can define writing proficiency (Grant and Ginther, 2000). However our results showed that subordination (*constituent*) by itself is not related to language proficiency. This is partially explained by the way we implemented and categorized subordinate clauses. Grant and Ginther (2000) categorized subordinate clauses based on how they function in a sentence (adverbial and relative clauses) and captured the frequency of these type in the overall text. We focused on the structural nature of subordinate clause by differentiate among finite and nonfinite subordinate clauses. This way of capturing subordination is not enough to determine language proficiency. Nevertheless, *constituent* is found in most feature combinations which indicates that it works as a predictor of L2 proficiency but not to that extend to give satisfactory results as an individual feature.

6.3 Discussion

Considering the confusion matrices in 6.2 we noticed that all our experiments have in common the proper classification between beginners (A2) and native speakers of English (N). This lead us to isolating these two proficiency groups and applying our feature combinations in a classification task of considering only two classes A2 and N. The results showed that *tense* and *aspect* was the top performing feature combination. However, even with two classes our top performance (AUC 0.771)

is not that different from the one given by ALL (AUC 0.613). These results lead us to analyze further the nature of our corpora and observe if our feature value variation is directly related to our output.

Feature	AUC
tense aspect	0.771
tense aspect constituent	0.771
tense aspect constituent bigram	0.768
tense aspect bigram	0.766
tense voice aspect bigram	0.766
tense voice aspect bigram trigram	0.765
tense aspect constituent trigram	0.765
de	0.765
tense voice aspect	0.764
tense aspect constituent bigram trigram	0.763
tense aspect trigram	0.763
tense voice aspect constituent bigram	0.762
tense aspect bigram trigram	0.761
tense voice aspect constituent bigram trigram	0.761
tense voice aspect constituent	0.760
tense aspect unigram bigram	0.758
voice aspect constituent	0.757
voice aspect	0.757
aspect constituent	0.756

Table 35: Beginners vs. native speakers of English

Our analysis depends mainly on the consistency of our corpora. While we have confidence in our findings, there is a need for further discussion on how any data limitations affect our feature's performance. Although the GLC, CEEAUS and ICNALE corpora were designed to be comparable across proficiency levels their homogeneity regarding the type of essays they contain influenced our results. The fact that all of them consist of argumentative essays limits the variety of our verb group characteristics *tense*, *aspect* and *voice*. The high occurrence of *present* tense *indefinite* aspect and *active* voice are mostly related to the construction of argumentative essays. Writers of argumentative essays typically support their arguments by describing specific events and by providing generalizations and generalizable statements or by describing events that are considered general truths (Baker et al., 2013). According to Beason and Lester (2010), *present tense* should be used to make statements of facts or generalizations and *past tense* should be used to narrate a story or an event that happened in the past. Hinkel (2004) supports the previous statement by reporting that students tend to use present tense and indefinite aspect more when they write argumentative essays.

The above is evident across all proficiency levels in each corpus. The values, *present*, *indefinite* and *voice* consume a large portion of each feature's distribution. The side effect when analyzing learner data of the same text type (in this thesis we examine argumentative essays), is the dominance of a specific value which results in suppressing the effectiveness of the feature. In our case *present* tense, *indefinite* aspect, *active* voice characterize most of the essays and do not act as discriminators

across proficiency groups. A fact that also affects the predictive potential of *tense*, *aspect* and *voice* in general. We believe that this genre homogeneity can be considered a limitation only when seeking to derive generalizable results. Thus finding a feature or a feature combination which can predict writing proficiency, independent of any factor related to the second one (such as L1 language of learner, genre, . . .).

Tense	**A2**	**B1**	**B2$^+$**	**N**	**Aspect**	**A2**	**B1**	**B2$^+$**	**N**	**Voice**	**A2**	**B1**	**B2**	**N**
present	53	52	49	39	indefinite	72	69	66	57	active	70	68	65	57
past	7	6	8	16	progressive	1	2	2	4	passive	3	4	5	8
modal present	6	7	5	4	perfect	1	1	2	4	novoice	27	28	30	35
modal past	3	4	5	4	noaspect	27	28	30	35					
future	3	3	2	1										
notense	27	28	30	35										

Table 36: Percent occurrence of tense, aspect, voice features in GLC essays

Tense	**A2**	**B1**	**B2**	**N**	**Aspect**	**A2**	**B1**	**B2**	**N**	**Voice**	**A2**	**B1**	**B2**	**N**
present	51	48	44	44	indefinite	68	66	61	58	active	67	64	58	60
past	3	4	4	5	progressive	2	2	2	3	passive	4	4	5	6
modal_present	8	8	7	4	perfective	1	1	1	3	novoice	29	31	36	36
modal_past	5	6	7	7	noaspect	29	31	36	36					
future	3	3	3	4										
notense	29	31	36	36										

Table 37: Percent occurrence of tense, aspect, voice features in ICNALE

Tense	**A2**	**B1**	**B2$^+$**	**N**	**Aspect**	**A2**	**B1**	**B2$^+$**	**N**	**Voice**	**A2**	**B1**	**B2$^+$**	**N**
present	60	56	54	4	indefinite	73	71	72	59	active	71	69	67	57
past	4	5	4	9	progressive	1	2	1	3	passive	4	4	5	7
modal_present	7	7	8	5	perfective	1	1	1	3	novoice	25	27	28	36
modal_past	4	4	5	8	noaspect	25	26	26	35					
future	2	2	2	3										
notense	22	23	25	31										

Table 38: Percent occurrence of tense, aspect, voice features in CEEAUS

We already showed in the previous section that different corpora can give different results; thus we did not find a "unique" feature set that works perfectly with all corpora. Overall assessing the predictive potential of our verb phrase characteristics and word level n-grams in writing proficiency task gave the following interesting point. First, features when examined individually can form a promising baseline in writing proficiency prediction. However better performance comes from looking into the feature combinations and not at those features individually. Finally, we addressed how the corpus homogeneity can affect the performance of our features. The essays we examined

in their majority contain *present* tense, *indefinite* aspect and *passive* voice. This uniformity affected the performance of our features

6.4 Future Work

In this thesis, we demonstrated the predictive potential of verb phrase characteristics and word level n-grams in writing proficiency. However the obtained results indicate some ground for further investigation improvement of the existing features. For example, in this work we identify a verb's type of embedding by considering the structure of the subordinate clause containing it. A subordinate clause can also be categorized based on the functional purpose in the sentence containing. On the basis of its function in a sentence, subordinate clause can be divided in to following types: noun, adjective, adverb clause. Second language acquisition researchers have already presented a relation between adverbial clauses and second language writing proficiency (Grant and Ginther, 2000). Considering both functional and structural purpose of the subordinate clause which entails expanding a verb's type of embedding will provide a more complete view of the role of subordination in analyzing the second and first language learners writing style.

Another verb phrase characteristic that can be investigated further is the functional use of modal verbs. In our work we analyze an auxiliary modal verb in terms of tense. Our results showed that modal present and modal past are two tense values that can distinguish learner types and writing proficiency. Adding the functional use of auxiliaries can improve our systems especially in distinguishing better the different proficiency groups. A modal auxiliary verb gives much information about the function of the main verb that it governs. Modals have a wide variety of communicative functions, but these functions can generally be related to a scale ranging from possibility ("may") to necessity ("must"). Research performed by Chen (2010) showed that epistemic modality including modal auxiliaries (e.g.,might, may) and epistemic lexical verbs (e.g., think, indicate) can act as a distinguishing factor between native and non native speakers of English. Expanding this to writing proficiency detection would improve our results.

An interesting component that requires further research involves expanding our verb phrase characteristics by capturing the textual and semantic function of a verb. Verbs are often divided into semantic classes according to their meanings and textual functions. Quirk et al. (1985) for example, classified some factual verbs as public, private, and suasive. Semantic classes of verbs in English are numerous but only a few are common in students' essays. In general terms, their frequency rates in texts provide evidence of the extents of the writers' vocabulary ranges. Examining those semantic classes and their predictive potential in characterizing both L2 and L1 learners' writing style will provide a complete overview of verb phrase's role in second language acquisition research.

An area of interest involves acquiring further knowledge regarding noun phrase characteristics

that can improve our systems. Nouns traditionally have been divided into classes based on their semantic features and textual functions. Relating their enumerative, advance/retrospective, language activity, illocutionary, interpretive, and resultative functions, as well as, those that convey meanings of textual vagueness and indeterminacy to writing proficiency and learner type prediction is an area that still needs more exploration. Additionally, observing the use of determiners in L2 writing is another factor that should be examined. The use of determiner to precede a noun or noun phrase is usually not a problem for writers who have grown up speaking English, nor is it a serious problem for non-native writers whose first language is a romance language such as Spanish. For other writers, though, this can be a considerable obstacle on the way to their mastery of English. In fact, some students from eastern European countries - where their native language has either no articles or an altogether different system of choosing articles and determiners - find difficulty in using them.

Finally, another area that warrants further investigation involves examining textual cohesion devices and their relation to L2 writing style. For example, personal pronouns play an important role in textual cohesion because they are deictic and specifically referential . Their use in written discourse is pervasive, and they unify the information flow by representing the discourse roles of the participants. Personal pronouns in written text are treated as lexical entities and, thus, they have the function of lexical cohesive links. Because pronouns function as referential markers in the text flow, their appropriate use is deemed important in evaluations of both L1 and L2 writing skills. Another cohesive device worth investigating is the use of linking adverbials. Linking adverbials explicitly indicate the semantic relationship between textual segments and play a crucial role in making a text logically cohesive. Therefore, it is vital for English learners, whose writing or speech is often said to be lacking in logical lucidity, to use them qualitatively and quantitatively in an appropriate way.

Chapter 7

Conclusion

Second language proficiency research gives contradicting results regarding the relation of *tense* to writing proficiency. Our study shows *tense* to be rather stable across corpora and learner category on the corpora investigated, yet together with *aspect*, clear trends emerge that are weakly predictive of learner category. This feature combination combines well with other grammatical features such as *voice* and *degree of embedding* and shows even greater promise for more natural and varied language samples. However, the occurrence of *tense* and *aspect* in most best performing feature combination show their potential for a more natural and varied dataset. Our investigation suggests that small, designed corpora have very idiosyncratic patterns and that linguistic features have to be tested for their interoperability with each other and for their effectiveness on the corpora used.

More specifically, in this thesis, we presented the potential of word level n-grams *unigram*, *bigram*, *trigram* and verb phrase characteristics *tense*, *aspect*, *voice*, *type* and *degree of embedding* as predictors of writing proficiency. Our approach resulted in base line systems which address the dependent relation of our linguistic indices to learner corpora. We used three L2 learner corpora ICNALE (Asian authors), CEEAUS (Japanese L2 learners) and GLC (Korean L2 learners) and two native speaker of English learner corpora BAWE and LOC. All these data sets contain one type of text, argumentative essays, but they differ on topic variety and author's cultural background. We created four experiments, using the above corpora, to monitor the behavior of our feature sets: Asian L2 learners-same topics (ICNALE), Japanese L2 learners-same topics, Korean L2 learners-different topics and Asian learners-different topics (ALL). We used two approaches to represent learner's essays binary-based and frequency-based. Results indicated that the second method performed better across all experiments. Additionally, we applied two feature selection techniques Forward Features Selection and Information Gain but their results were similar in both approaches.

Initially, we examined the predictive power of our features individually on our learner corpora which gave rather uncertain results regarding the relation of each feature to writing proficiency.

Considering this unsurprising outcome, we exhaustively tested the combinations of features across all corpora. The results indicated the emerging of certain common trends. *Tense* and *aspect* were present in most combinations when examining L2 learners' essays. Additionally, *degree of embedding* captured our attention since its individual performance ranked fourteenth out of 256 feature combinations in GLC (Korean learners) experiment. Interestingly the appearance of this feature to the other two experiments (ICNALE, Japanese) was limited. This was due to the fact that the Korean second language learners constructed more complex sentences than writers of ICNALE and CEEAUS corpora.

We also considered the outperforming feature combination for each experiment and examined which proficiency group was predicted the best. We notice that we obtain the best classification when our algorithm classifies essays of beginners (A2) and native speakers of English (N). We always obtain a misclassification when an essay belongs to advanced proficiency level ($B2^+$) and sometimes a confusion when we deal with essays written by intermediate L2 learners (B1). However, $B2^+$ essays are usually wrongly categorized as N and B1 as A2. The same classification patterns appear in all four experiments; thus using verb phrase characteristics and word level n-grams succeed in distinguishing beginner L2 learners from native speakers of English. However, they give a low rate in predicting proficiency groups intermediate and advanced. Even then, the misclassification patterns follow the common reasoning which indicates the difficulty of separating a B1 learner from an A2 and a $B2^+$ from a native speaker of English.

Finally, we examined the occurrence of our verb features for each corpus used. We noticed the high occurrence of *present* tense, *indefinite* aspect and *passive* voice in across all corpora which justified the low performance of these three feature sets. This homogeneity is directly linked to the fact that all of them consist of argumentative essays. The above lead us to the conclusion that engineered corpora artificially limit the occurrence of certain features. Overall, we believe that statements regarding the relation of a grammatical features to writing proficiency have to be reported together with the the feature value distribution of the reference corpus used.

Bibliography

Canadian Language Benchmarks (2012). *Canadian Language Benchmarks: English as a Second Language for Adults*. Center of Canadian Language Benchmarks.

Aizerman, M. A., Braverman, E. A., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control*, number 25 in Automation and Remote Control, pages 821–837.

American Council (2012). *ACTFL Proficiency Guidelines*. ACTFL, INC, Alexandria, VA, USA.

Baker, J., Brizee, A., and Angeli, E. (2013). *Essay Writing: The Argumentative Essay*. Purdue University, West Lafayette, IN, USA.

Bardovi-Harlig, K. and Reynolds, D. W. (1995). The role of lexical aspect in the acquisition of tense and aspect. *TESOL Quarterly*, 29(1):107–131.

Beason, L. and Lester, M. (2010). *A Commonsense Guide to Grammar and Usage with 2009 MLA Update*. Bedford/St. Martin's, Boston, MA, USA.

Begi, N., Kader, M. I., and Vaseghi, R. (2013). A corpus-based study of Malaysian ESL learners' use of modals in argumentative compositions. *English Language Teaching*, 6(9):146–157.

Bell, J. and Burnaby, B. (1984). *A handbook for ESL literacy*. Pippin Publishing, London, UK.

Biesenbach-Lucas, S., Meloni, C., and Weasenforth, D. (2000). Use of cohesive features in ESL students' e-mail and word-processed texts: A comparative study. *Computer Assisted Language Learning*, 13(3):221–237.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):1–15.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, pages 77–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Briscoe, T., Medlock, B., and Andersen, O. (2010). Automated assessment of ESOL free text examinations. Technical report, University of Cambridge, Computer Laboratory, Cambridge, UK.

Burstein, J., Leacock, C., and Swartz, R. (2001). Automated evaluation of essays and short answers. In *Proceedings of the 5th CAA Conference*, pages 1–4, Loughborough. Loughborough University.

Callies, M., Diesz-Bedmar, M. B., and Zayteseva, E. (2014). *Using learner corpora for testing and assessing L2 proficiency*. Second Language Acquisition series. Multilingual Matters, Clevedon.

Carlstrom, B. and Price, N. (2013). Data-driven learning made easy. In *Proceedings of the 21st Annual KOTESOL International Conference*, Seoul, Korea. KOREA TESOL.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Nevada, USA. Univesity of Nevada.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations Newsletter*, 6(1):1–6.

Chen, H. I. (2010). Contrastive learner corpus analysis of epistemic modality and interlanguage pragmatic competence in L2 writing. *Arizona working papers in SLA and teaching*, (17):27–51.

Chomsky, N. (1969). *Aspects of the Theory of Syntax*. Research Laboratory of Electronics Cambridge, Mass: Special technical report. MIT Press.

Comrie, B. (1976). *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK.

Comrie, B. (1985). *Tense*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK.

Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, 24:67–87.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Applied Linguistics Non Series. Cambridge University Press, Cambridge, UK.

Crossley, S. A., Defore, C., Kyle, K., and Dai, Jianmin. & McNamara, D. S. (2013). Paragraph specific n-gram approaches to automatically assessing essay quality. In *Proceedings of the 6th Educational Data Mining (EDM,) Conference*, pages 216–220, Heidelberg, Germany. Springer.

Crossley, S. A., Salsbury, T., McNamara, D. S., and Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28(4):561–580.

Cunningham, H. (2002). GATE - a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (April 21, 2011). *Developing Language Processing Components with GATE Version 6*. GATE.

Díez-Bedmar, M. B. (2010). *From secondary school to university: The use of the English article system by Spanish learners*. Castelló de la Plana: Publicacions de la Universitat Jaume I.

Doandes, M. (2003). Profiling for belief acquisition from reported speech. Master's thesis, Concordia University, Montreal, QC, CA.

Ellis, N. C. (2008). *Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure*, pages 63–103. Blackwell Publishing Ltd, Hoboke, NJ, USA.

Encyclopedia Britannica (2002). *Encyclopedia britannica*. Encyclopedia Britannica.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2):139 – 155.

Ferris, D. R. (1994). Lexical and syntactic features of ESL, writing by students at different levels of L2 proficiency. *TESOL Quarterly*, 28(2):pp. 414–420.

Frase, L. T., Faletti, J., Ginther, A., and Grant, L. (1999). Computer analysis of the TOEFL test of written English. Technical Report 64, Educational Testing Service, Princeton, NJ, USA.

Fulcher, G. (2004). Deluded by artifices? the Common European Framework and harmonization. *Language Assessment Quarterly*, 1(4):253–266.

Gardner, S. and Nesi, H. (2012). A classification of genre families in university student writing. *Applied Linguistics*, 34(1):1–29.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In *Second Language Research Forum: Building Bridges between Disciplines*, pages 240–254, Somerville, MA, USA. Cascadilla Proceedings Project.

Granger, S. (1996). *From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora*. Lund Studies in English. Lund University Press, Lund, Sweden.

Granger, S. (1998). *Learner English on computer*. Studies in language and linguistics. Longman, London, UK.

Granger, S. (1999). *Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus*, pages 191–202. Rodopi, Amsterdam, Netherlands.

Granger, S. (2002). *A bird's eye view of learner corpus is research*. Language Learning & Language Teaching. Benjamins, Amsterdam & Philadelphia.

Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3):538–546.

Granger, S. (2009). *The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation*. John Benjamins Publishing Company, Amsterdam & Philadelphia.

Granger, S., Dagneaux, E., and Meunier, F. (2002). *International Corpus of Learner English v1*. Presses universitaires de Louvain, Louvain-la-Neuve.

Grant, L. and Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2):123 – 145.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Haiyang, A. and Xiaofei, L. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Díaz-Negrillo, A., Ballier, N., and Thompson, P., editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 249–264, Amsterdam, Philadelphia. John Benjamins.

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

Hawkey, R. and Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, 9(2):122–159.

Hawkins, J. and Filipović, L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. English Profile studies. Cambridge University Press, Cambridge, UK.

Hawkins, J. A. and Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1:1–23.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2006). Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Interspeech*. International Speech Communication Association.

Higgs, T. V. (1984). *Teaching for proficiency: the organizing principle*. ACTFL foreign language education series. National Textbook Company.

Hinkel, E. (2002). *Second Language Writers' Text: Linguistic and Rhetorical Features*. ESL and applied linguistics professional series. Lawrence Erlbaum Associates, Mahwah, N J, USA.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic text. *TESOL Quarterly*, 37(2):275–301.

Hinkel, E. (2004). Tense, aspect and the passive voice in L1 and L2 academic texts. *Language Teaching Research*, 8:5–29.

Hudson, T. (2005). Trends in assessment scakes and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, 25:205–227.

Hulstijn, J. H. (2011). *Explanations of associations between L1 and L2 literacy skills*, pages 85–112. John Benjamins Publishing Company.

Hunston, S. (2006). *Corpora in applied linguistics*. Cambridge University Press, Cambridge, UK.

Ishikawa, S. (2010). A corpus-based study on Asian learners' use of English linking adverbials. *Themes in Science and Technology Education. Special Issue on ICT in language learning*, 3(1-2):139–157.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In Weir, G. R. S., Ishikawa, S., and Poonpon, K., editors, *Corpora and Language Technologies in Teaching, Learning and Research*, pages 3–11, Glasgow, UK. University of Strathclyde Press.

James, C. (1998). *Errors in language learning and use: exploring error analysis*. Applied linguistics and language study. Longman, London, UK.

Jiang, X., Guo, Y., Geertzen, J., Alexopoulou, D., Sun, L., and Korhonen, A. (2014). Native language identification using large, longitudinal data. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jin, W. (2001). A quantitative study of cohesion in Chinese graduate students' writing: Variations acros genres and proficiency levels. Technical report, Purdue University, West Lafayette, IN, USA.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 377–384. ACM Press.

Kameen, P. (1980). *Syntactic skill and ESL writing quality*. Teachers of English to Speakers of Other Languages, Inc. (TESOL).

Kitao, K. S. and Saeki, N. (1992). Process and social aspects of writing: Theory and classroom application. *Annual Reports of Studies*, 33:86–102.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.

Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *IJCAI95*, pages 1034–1040. Morgan Kaufmann.

Korhonen, A., Krymolowski, Y., and Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genova, Italy.

Kubota, R. (1998). An investigation of L1L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of Second Language Writing*, 7(1):69 – 100.

Kusher, J., Lantolf, J., Thorne, S. L., Jiménez, A., Ross, B., and Salaberri, S. (2001). Spanish acquisition: Analysis of learner corpora generated through inter-cultural telecollaboration. *Procesamiento del Lenguaje Natural*, 37:301–310.

Lardiere, D. (1998). Dissociating syntax from morphology in a divergent L2 end-state grammar. *Second Language Research*, 14(4):359–375.

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals*, 36(4):483–490.

Lozano, C. and Mendikoetxea, A. (2013). *Learner corpora and second language acquisition: The design and collection of CEDEL2*. Studies in Corpus Linguistics. John Benjamins Publishing Company.

Manevitz, L. M. and Yousef, M. (2002). One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154.

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.

Mattar, H. (2002,2003). Is avoidance ruled out by similarity? the case of subordinating/conjunctions adverbs in English and Arabic. *Poznań Studies in Contemporary Linguistics*, 38:103–115.

McDoual, A. (2010). A corpus based investigation into the use of English modal auxiliaries by adult korean L2-learners. *Korea University Working Papers in Linguistics*, 4:38–51.

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

Meara, P., Jacobs, G., and Rodgers, C. (2002). Lexical signatures in foreign language free-form texts. *ITL Review of Applied Linguistics*, 135 - 136:85 – 96.

Min, K. E. (2013). *How grammar matters in NNS academic writing: The relationship between verb tense and aspect usage patterns and L2 writing proficiency in academic discourse*. PhD thesis, University of Illinois at Urbana-Champaign.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

Nenkova, A., Chae, J., Louis, A., and Pitler, E. (2010). Structural features for predicting the linguistic quality of text. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*, volume 5790 of *Lecture Notes in Computer Science*, pages 222–241. Springer Berlin Heidelberg.

Nicholls, D. (2003). The Cambridge learner corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics*, pages 572–581.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of collegelevel L2 writing. *Applied Linguistics*, 24(4):492–518.

Paltridge, B. (1996). Genre, text type, and the language learning classroom. *ELT Journal*, 50(3):237–243.

Panagiotopoulos, A. and Bergler, S. (2014). How predictive is tense for language profiency? a cautionary tale. In Gelbukh, A., Espinoza, F., and Galicia-Haro, S., editors, *Human-Inspired Computing and Its Applications*, volume 8856 of *Lecture Notes in Computer Science*, pages 139–150. Springer International Publishing.

Patanasorn, C. (2013). Association between the simple past and emergence of the present perfect in EFL learners' writing. In *The Asian Conference on Language Learning*, pages 587–600, Osaka, Japan. The International Academic Forum.

Pendar, N. and Chapelle, C. (2008). Investigating the promise of learner corpora: Methodological issues. *CALICO Journal*, 25(2):189–206.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(1):89–106.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik (1985). *A Comprehensive Grammar of the English Language*. Longman, London, UK.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523.

Seeger, M. (2001). Learning with labeled and unlabeled data. Technical report, University of Edinburgh, Edinburgh, UK.

Sinclair, J. M. (1996). Preliminary recommendations on corpus typology. Technical report, University of Birmingham, Birmingham, UK.

Summers, D. (1993). Longman/Lancaster English language corpus criteria and design. *International Journal of Lexicography*, 6(3):181–208.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44(2):307–336.

Trosborg, A. (1987). Apology strategies in natives/non-natives. *Journal of Pragmatics*, 11(2):147 – 167.

VanPatten, B. and Benati, A. G. (2010). *Key Terms in Second Language Acquisition*. Key Terms. Bloomsbury Academic, London, UK.

Vaughan, C. (1991). Hollistic assesment: What goes on in the rater's mind. In Hamp-Lyons, L., editor, *Assessing second language writing in academic contexts*, Writing research, pages 111 – 125. Ablex Publishing Corporation, Norwood, NJ.

Vethamani, M. E., Manaf, U. K. A., and Akbari, O. (208). Students' use of modals in narrative compositions: Forms and functions. *English Language Teaching*, 1(1):61–74.

White, E. M. (1994). *Teaching and assessing writing: recent advances in understanding, evaluating, and improving student performance*. Jossey-Bass higher and adult education series. Jossey-Bass.

Wu, C.-Y., Lai, P.-H., Liu, Y., and Ng, V. (2013). Simple yet powerful native language identification on TOEFL11. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 152–156.

Wulff, S. and Gries, S. T. (2011). *Corpus-driven methods for assessing accuracy in learner production*. Task-based language teaching : issues, research and practice. John Benjamins Publishing Company, Amsterdam & Philadelphia.

Yap, B., Rani, K., Rahman, H., Fong, S., Khairudin, Z., and Abdullah, N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In Herawan, T., Deris, M. M., and Abawajy, J., editors, *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, volume 285 of *Lecture Notes in Electrical Engineering*, pages 13–22. Springer Singapore.

Zareva, A. (2007). Structure of the second language mental lexicon: how does it compare to native speakers' lexical organization? *Second Language Research*, 23(2):123–153.

Appendix A

Imbalanced Data

A major difficulty in text classification tasks using supervised techniques (Support Vector Machine,...) is that they commonly require high-quality training data to construct an accurate classifier. Unfortunately, in many real-world applications the training sets are extremely small and sometimes they present imbalanced class distributions (i.e., the number of examples in some classes are significantly greater that the number of examples in others).

In order to overcome these problems, recently many researchers have been working on different solutions to the class-imbalance problem. It has been showed that by augmenting the training set with additional, unlabeled, information it is possible to improve the classification accuracy using different learning algorithms such as Support Vector Machines. It has also been reported that by adjusting the number of examples in the majority or minority classes it is possible to tackle the suboptimal classification performance caused by the class-imbalance (Seeger, 2001). In particular, there is evidence that under-sampling, a method in which examples of the majority classes are removed, leads to better results than over-sampling, a method in which examples from the minority classes are duplicated (Chawla et al., 2004).

In this study we are approaching the class-imbalance problem by adjusting second language learners' essays to the minority class using random under-sampling. Random under-sampling is a simple approach to re-sampling. Documents in majority class are randomly eliminated until the ratio between the minority and majority class is at the desired level. Theoretically, one of the problems with random under-sampling is that one cannot control what information about the majority class is thrown away. In particular, very important information about the decision boundary between the minority and majority class may be eliminated. Despite its simplicity, random under-sampling has empirically been shown to be one of the most effective re-sampling methods (Yap et al., 2014).

In order to make sure that our results using random under-sampling are not biased we performed

a set of preliminary experiments on learner corpora with high class-imbalance problems. For example, consider the experiment Korean L2 Learners-Different topics (GLC) described in Section 5.1. The essays are distributed as follows: A2 (beginners), 10853 essays, B1 (intermediate), 4470 essays, $B2^{+}$ (advanced) 1787 essays and Native Speakers of English (N) 1414 essays. Classes A2 and B1 have more written text than $B2^{+}$ and N. Thus we separate the majority classes into subsets of 1414 essays (equivalent to the minority class) resulting into six subsets for class A2 and three for $B2^{+}$. Then we perform six classification tasks (as many the subjects of the majority class). Results from all tasks were similar in terms of performance and predictive feature combinations. These showed a uniform behavior of our feature sets across our corpus. The same outcome we obtained when we applied this technique to the rest of our corpora.