

Text Mining to Support Foreign Language Learning

フラナガン, ブレンダン ジョン

<https://doi.org/10.15017/1807065>

出版情報 : 九州大学, 2016, 博士 (情報科学), 課程博士
バージョン :
権利関係 : 全文ファイル公表済

Doctoral Dissertation

Text Mining to Support Foreign Language Learning

Brendan Flanagan

January 2017

Department of Advanced Information Technology
Graduate School of Information Science and Electrical Engineering
Kyushu University

Abstract

The ability to express intention and emotion are fundamental to human communication, and these skills are usually learnt during early childhood. This can be seen not only in a child using their first words, but also in an adult who is using their limited known vocabulary of a foreign language that they have just started to learn. A fundamental problem in language learning is the difficulty of expressing intention and emotion in a coherent sentence regardless of whether the language being learnt is a foreign or native language. Unintended syntactic errors can alter the intention of an expression, making the meaning of an expression different from the intended meaning, such as the word order error in the following sentence: “Here are *new my* temple pictures” where the error makes the intended meaning of the sentence unclear. Expressing emotion in a foreign language can be a daunting task for learners, and has parallels with other circumstances where a set of special emotion expressions that are not used in everyday communication, such as: the expression of complex aromas, flavors, and many other characteristics used in wine tasting notes. In this work, we investigate text mining techniques to support foreign language learners from the perspective of intention and emotion.

The first half of this thesis focuses on the expression of intention, starting with the automatic classification of 15 writing error categories by SVM analysis of data collected from a language learning SNS. This initial analysis serves as the basis for investigation into the characteristic differences between learner native language groups from the perspective of their writing errors. In particular, we examine the error co-occurrence profiles and the distance between 5 native languages, and how a learners’ native language can be predicted from automatic writing error analysis. We also examine the classification of a learners’ proficiency based on their usage of words. The contribution of these researches is the ability to automatically profile learner characteristics that affect the expression of intention, and predict other characteristics based on common trait association. Also, we address the problem of insufficient sample data, which is often faced when analyzing auto-

matic writing error detection as it is costly and time consuming to create learner writing corpora. To overcome this problem, we propose a method for extracting word order errors, which are a known problem for Japanese learners of English, and automatically compile a balanced corpus from the logs of a language learning SNS.

The second half of this thesis focuses on the expression of complex emotion, starting with the analysis of wine tasting notes by SVM models from 4 sensory viewpoints. These models are then used to extract characteristic words that can be used in emotive sensory expression, and to visualize the characteristics of collections of wine reviews from regions. We then examine the change in the use of emotional expression by analyzing differences in wine tasting notes over a period of 14 years. This research looks in particular at the changes of expression from the perspective of adjective antonym pairs that represent opposites of emotional meaning attributes by visualization. Finally, we examine in detail the complex adjective antonym meaning dimension of THICK and THIN, where the optimal quality is not found at the polar opposites of the dimension. The contribution of these researches is the ability to automatically analyze expressions of emotion, and provide methods to understand complex emotive expressions through visualization.

Acknowledgements

First and foremost I would like to sincerely thank my supervisor, Prof. Sachio Hirokawa, for all his support over the last five and a half years, from my time as a research student, master's student, and finally as a doctoral student. During this period, he has not only imparted knowledge on topics within my research area, but also taught me how to conduct and advance research, from conception to fruition. His wise guidance and warm encouragement have also enabled me to pursue challenges that I would have only dreamed of in years gone by.

I am also grateful to my dissertation committee members Prof. Hiroaki Ogata and Prof. Tsunenori Mine, and advisory committee members Prof. Kiyota Hashimoto and Prof. Kumiko Tanaka-ishii for their critical assessment and insightful comments, which challenged me as a researcher to achieve a higher standard of work.

I would also like to especially thank Prof. Yukio Tono (Tokyo University of Foreign Studies) for his expert advice on my research into learner writing errors, and the many opportunities he has afforded me over the years; Prof. Paula Buttery (Cambridge University) for her guidance and providing me with the opportunity of being a visiting researcher at Cambridge University for 6 weeks; Prof. Akiko Aizawa (National Institute of Informatics) and Prof. Mamoru Komachi (Tokyo Metropolitan University) for their guidance and the opportunities they have given me; Prof. Carita Paradis (Lund University) for her constructive advice on the analysis of emotive expressions in winespeak; Prof. Emi Izumi (Doshisha University) and Prof. Emiko Kaneko (University of Aizu) for their feedback and comments on the prediction of learner proficiency levels.

I would also like to express my appreciation to Prof. Eisuke Ito, Prof. Chengjiu Yin, Prof. Takahiko Suzuki, and Prof. Tetsuya Nakatoh for their help and comments. Thank you to my fellow students of the Hirokawa laboratory, especially to Toshihiko Sakai and Jun Zeng for their camaraderie, and Takanori Yamashita who has been a companion throughout our doctoral course and introduced me to the field of medical text mining.

Finally, I would like to express my deepest thanks to my family (especially mum and dad) and friends for their love, support and encouragement.

Contents

Abstract	i
1 Introduction	1
2 Preliminaries	5
2.1 Writing Error Categories	5
2.2 Sensory Viewpoint Expressions of Winespeak	9
3 Classification of English Language Learner Writing Errors	11
3.1 Related Work	13
3.2 Vectorization of Error Sentences for Categorization	15
3.3 Evaluation of error categorisation using SVM	17
3.4 Optimal Feature Selection	21
3.4.1 Detailed Analysis	25
3.4.2 Improvement of Prediction Performance by Optimal Feature Selection	25
3.5 Discussion	25
4 Clustering and Co-occurrence Analysis of English Writing Errors Based on Native Language	29
4.1 Clustering based on Error Category Prediction	30
4.1.1 Data Collection	30
4.1.2 Error Co-occurrence Analysis by Clustering	31
4.2 Co-occurrence Analysis By Native Language	33
4.2.1 Principal Component Analysis (PCA)	33
4.2.2 Error Co-occurrence Analysis of Native Languages by Clustering . .	33
4.2.3 Analysis by Tree Distance	40
4.3 Conclusion	41

5	Classification of Native Language by Writing Error Analysis and Characteristic Feature Extraction	45
5.1	Related Work	46
5.1.1	Native Language Identification	46
5.1.2	Native Language Prediction by Error Analysis	47
5.2	Data Collection	48
5.2.1	Writing Error Tagged Corpus	49
5.2.2	Native Language Tagged Corpus	50
5.3	Error Vector Analysis to Prediction of Learn Native Language	50
5.3.1	Biased Words	50
5.3.2	Method and Results	53
5.4	Discussion	56
6	Automatic Extraction and Prediction of Word Order Errors From Language Learning SNS	57
6.1	Automatic Word Order Error Sample Extraction by Edit Distance	58
6.2	Data Collection	58
6.3	Word Order Error Prediction by SVM and Feature Selection	61
6.3.1	Method	61
6.3.2	Baseline Prediction Performance Evaluation	61
6.3.3	The Effect of Feature Selection on Prediction Performance	62
6.4	Discussion	65
7	Classification of Speaking Proficiency Level by Machine Learning and Feature Selection	67
7.1	Proficiency Level Classification by SVM and Feature Selection	68
7.1.1	Feature Selection Measures	69
7.2	Proficiency Classification Performance	70
7.3	Characteristic Features of Level A1	71
7.4	Discussion	72
8	Predicting and Visualizing Wine Characteristics Through Analysis of Tasting Notes From Viewpoints	73
8.1	Related Work	74

8.2	Data Collection	74
8.3	Sensory Viewpoint Analysis and Prediction	75
8.4	Visualization of Sensory Sentiment as Radar Charts	78
8.4.1	Model Normalization For Characteristic Prediction and Visualization	78
8.4.2	Visualization of Sensory Sentiment by Region	79
8.5	Discussion	83
9	Analysis of The Diachronic Relations of Adjective Antonym Pairs in Wine Tasting Notes	85
9.1	Data Collection	86
9.2	Diachronic Analysis of Antonyms by Frequency	89
9.3	Diachronic Analysis of Antonyms by Mind Map	93
9.3.1	Mind Map Generation	93
9.4	Discussion	102
10	Analysis of Antonymic Adjective Meaning Dimensions in Winespeak	103
10.1	Introduction	103
10.1.1	Data Collection	104
10.1.2	Features of Thick and Thin	105
10.2	The feature space of the WIDTH dimension	106
10.3	Discussion	110
11	Conclusion	111
	Bibliography	115

List of Figures

1.1	An overview of concepts investigated in this thesis.	2
2.1	Error correlation of Lang-8 vs Weltig.	8
3.1	Error classification evaluation for each category (F-measure, 10-fold cross-validation).	19
3.2	Error classification evaluation for each category (10-fold cross-validation, Accuracy).	20
3.3	Correlation between the number of data samples and the F-measure of the evaluation.	21
3.4	Correlation between the number of data samples and the Accuracy of the evaluation.	22
3.5	Effect of Feature Selection for Error Category 2.	23
3.6	Effect of Feature Selection for Error Category 42.	24
3.7	Comparison of Prediction Performance	26
4.1	Number of sentences grouped by native language.	30
4.2	Clustering of all writing error data.	32
4.3	Principal Component Analysis of all writing error data.	34
4.4	Clustering of writing errors by Japanese natives.	35
4.5	Clustering of writing errors by Korean natives.	36
4.6	Clustering of writing errors by Chinese natives.	37
4.7	Clustering of writing errors by Taiwanese natives.	38
4.8	Clustering of writing errors by Spanish natives.	39
5.1	The process of creating error vector representations of each sentence.	51
5.2	Distribution of predicted errors for each language	51

5.3	The RSD distribution of word frequencies between five native languages.	54
5.4	Native language prediction evaluation for each vector (Accuracy, 10-fold cross validation).	55
6.1	Distribution of the edit distance of corrected word order error sentence samples.	60
6.2	The effect of optimal feature selection on prediction performance for the <i>Single</i> corpus with the sub-feature set of words, n-grams, and POS tags.	64
7.1	Performance accuracy of feature selection	71
8.1	An overview of the automatic prediction and visualization of wines from multiple viewpoints.	76
8.2	Prediction performance of feature selection using the top N absolute value score words for the touch sense.	77
8.3	Example radar chart of the region Pelješac which has the largest graph area.	80
8.4	Example radar chart of the region Primorska which has the smallest graph area.	80
8.5	Strongest taste sentiment: Sonoma County, Santa Barbara County.	81
8.6	Strongest vision sentiment: Alto Adige Valle Isarco.	81
8.7	Strongest smell sentiment: Ioannina.	82
8.8	Strongest touch sentiment: Barossa Valley, Clare Valley.	82
9.1	Distribution of wine tasting notes published from 1999 to 2014.	87
9.2	Top 10 frequently occurring terms by year.	88
9.3	Distribution of the term frequency ratio of the adjectives thick or thin by year.	89
9.4	Distribution of thick and thin by year from the viewpoints of smell, taste, touch, and vision.	90
9.5	Dissimilarity of the co-occurring adjectives of thick and thin by year.	91
9.6	Dissimilarity of the co-occurring adjectives of thick and thin by year from the viewpoints of smell, taste, touch, and vision.	92
9.7	A mind map of thin from 1999.	94
9.8	A mind map of thick from 1999.	95
9.9	Merging mind maps of thick and thin to show distinct and common nodes.	96

9.10	The merged mind map of thick and thin from 1999.	97
9.11	Mind map of thick and thin from 2002.	98
9.12	Mind map of thick and thin from 2004.	99
9.13	Mind map of thick and thin from 2013.	100
9.14	Jaccard dissimilarity of the mind maps of thick and thin by year.	101
10.1	Relation of thick and thin: BNC corpus (left), Winespeak (right).	104
10.2	Feature selection performance of classifying thick and thin by SVM.	107
10.3	Scatter plot of feature scores from the thick model and thin model.	108
10.4	Interpolated feature density heat map.	109

List of Tables

2.1	Merged error number list used for analysis.	7
2.2	Linear regression analysis results.	8
2.3	Outlier error categories and relation to Lang-8 error frequency.	9
2.4	17 Example words that describe sensory modalities as defined by Paradis and Eeg-Olofsson [51]	10
3.1	An example of an original and corrected sentence pair.	16
3.2	Indexed example sentence.	17
3.3	Evaluation of the classification of error categories.	18
3.4	Evaluation of the classification of errors into categories by 10-fold cross-validation.	20
3.5	The words and tags from the model created using SVM.	25
3.6	Prediction Performance Compared by F-measure	26
4.1	Error Category Numbers and Descriptions.	31
4.2	Characteristic co-occurring errors for each of the 5 main native languages.	40
4.3	Error category difference of native languages based on tree distance.	42
5.1	Error Category Numbers and Descriptions.	49
5.2	Distribution of sentences grouped by native language.	50
5.3	Top 10 positive and negative feature words by weight for native Japanese learners of English.	52
5.4	Biased words in the model for Japanese native language learners.	52
6.1	Number of corrected sentence samples extracted.	59
6.2	<i>Parallel</i> corpus baseline prediction performance.	61
6.3	<i>Single</i> corpus baseline prediction performance.	62

6.4	Optimal feature selection prediction performance for the <i>Parallel</i> corpus. . .	62
6.5	Optimal feature selection prediction performance for the <i>Single</i> corpus. . .	63
6.6	Single corpus top 10 positive and negative features.	63
6.7	Trigram word order error characteristic feature examples.	65
7.1	Equivalent levels of CEFR, CEFR-J, and SST	68
7.2	Measures used for Feature Selection.	69
7.3	Classification performance when using all feature words.	70
7.4	Characteristic features of A1 and other comparative levels.	72
8.1	Top 10 positive and negative score words for the smell sense.	75
8.2	Feature selection: Optimal N and evaluation for each of the sensory view- points.	78
9.1	Occurrence frequency of top 10 terms.	87
10.1	Top 10 positive and negative features of thick and thin.	105

Chapter 1

Introduction

A fundamental problem in language learning is the difficulty of expressing intention and emotion in a coherent sentence regardless of whether the language being learnt is a foreign or native language. In its most elementary form, an expression can be thought of as a simple combination of words that are just enough to communicate needs or desires. This can be seen in a child using their first words, or an adult who is using their limited known vocabulary of a language that they have just started to learn. According to Tomasello et al. [62], the human ability of expression is divided into two categories: (1) goal-oriented intention, (2) emotion, experience, and activity, both of which are acquired in early childhood. Unintended errors can alter the intention of an expression, making the meaning of an expression different from the intended meaning, such as the word order error in the following sentence: “Here are *new my* temple pictures”. The aim of this thesis is to investigate the support of language learners, and in particular foreign language learners, from the perspective of intention and emotion, which will be introduced in the following sections of this chapter. An overview of the problems examined in this thesis is shown in Figure 1.1 with concentric circles representing the hierarchy of concepts.

The rest of this thesis is organized as follows. In chapter 2, the categories of foreign language learner error that are the basis of intention analysis are introduced. We also detail the four sensory viewpoints that are the foundation of emotion analysis. The chapters following chapter 2 are related to two main topics: intention in chapters 3 to 7, and emotion in chapters 8 to 10.

In their analytic framework for intentions, Xue et al [72] look at an individual's connotative intentions. They breakdown an act of correcting an error into two intentions: the *syntactic intention* that is the act of correcting a syntactic error that changes sentence

Language Learner Expression

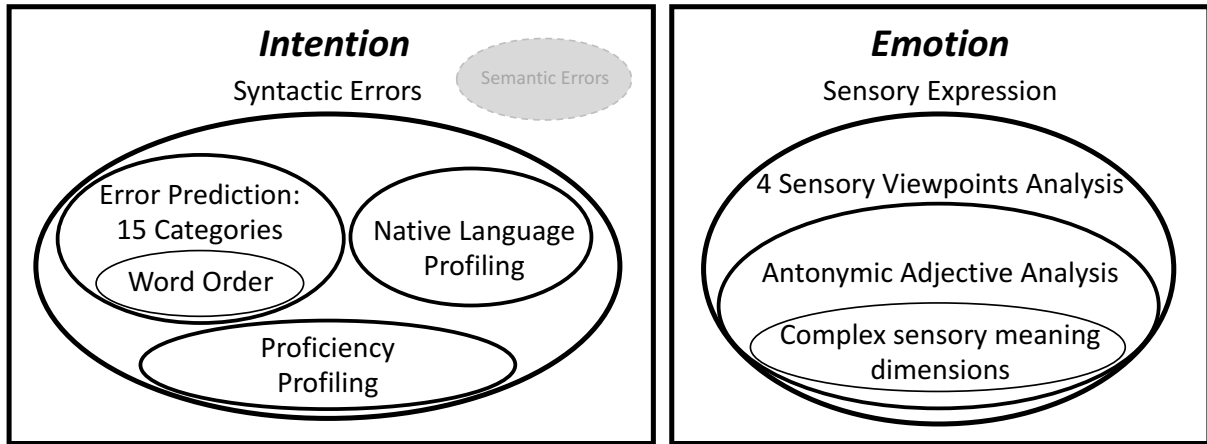


Figure 1.1: An overview of concepts investigated in this thesis.

A to B, which then invokes a change in *semantic intention* that changes the meaning of sentence A to the intended meaning contained in sentence B. Language learners are often prone to making errors during the learning process which changes the *syntactic intention* of their expressions, which also changes their externally perceived *semantic intention*.

One of the aims of this thesis is to investigate the support of foreign language learners in the expression of *syntactic intention* by focusing on the characteristics of learners and its influence on writing errors. It should be noted that the investigation of errors in *semantic intention*, such as: where the sentence is syntactically correct but the intended meaning is incorrect, are beyond the scope of this thesis. We start by examining fundamental causes of errors in *syntactic intention* by investigating a method of classifying 15 categories of learner writing errors by Support Vector Machine (SVM) in chapter 3. The classification performance is further improved by applying a technique of optimal feature selection, which is also used to extract and rank the characteristic features of writing errors. The data for the experiments were collected from a leading language learning social network service (SNS), Lang-8¹. This has significant importance to this area of research as previous work has focused on the analysis of traditional learner corpora that are costly and time consuming to create [42]. As foreign language learners who are studying the same target language come from a range of different native language backgrounds, it is important to investigate the possible effects that their native language might have on *syntactic intention* expression errors in the language they are learning. To realize this investigation,

¹<http://www.lang-8.com>

in chapters 4 and 5, the predicted score from the 15 error category classifiers is analyzed as a predicted error vector representation of learner writing. Chapter 4 investigates the co-occurrence of errors and the difference between five native language groups. We propose a scoring method that represents the distance between native language groups based on the characteristics of co-occurring errors. In chapter 5, we propose and evaluate a method of classifying the native language of a learner based on the predicted error vector representation of their writing. A problem that is often faced in the analysis of automatic writing error detection is the shortage of data available for training, testing and evaluating systems. In chapter 6, to overcome this problem we introduced a method for extracting word order errors and automatically compiling a balanced corpus from the logs of a language learning SNS. SVM classifiers were trained on the corpus and tested. In chapter 7, a method of automatically predicting the proficiency of a language learner based on the transcripts of speaking exams is proposed.

Expressing emotion in a foreign language can be a daunting task for learners, and has parallels with other circumstances where a set of special emotion expressions that are not used in everyday communication. A good example of this can be seen in the description of food and beverages that consist of complex aromas, flavors, and many other characteristics as they usually are expressed using specialist terminology used in a subjective manner. Within this area, the descriptions of wine are notorious for the use of specialist terminology and the expression of commonly used words in an uncommon manner. We propose that the process of a person learning these specialist terminology emotion expressions is similar to that of a foreign language learner learning to express emotion in the language they are learning.

The investigation of emotion expression is an ongoing research topic in text mining commonly referred to as sentiment analysis, where the emotional content of writing is predicted. Common subjects ranging from movie review analysis to the analysis of political views on SNS. In a recent survey of the field, Medhat et al. [45] identified that sentiment analysis research often falls into two categories of polarity emotion relations: positive/negative, or generalized analysis of text. While these techniques are applicable to a range of sentiment analysis problems, complex emotion expression relations, such as those found in wine tasting reviews have received little attention. The aim of the emotion expression analysis in this thesis is to identify methods of extracting characteristic features of not only positive/negative polarity relations, but also complex relations, such

as the double negative antonymic adjective pair of thick and thin found in wine tasting notes, where the optimal property is found in the middle of the extremes of the poles. The proposed methods could be applied to the extraction of emotion expression characteristic features to support foreign language learners in the expression of emotion. In chapter 8, we investigate the emotional sensory expressions in wine tasting notes. Firstly, we present a method of classifying by SVM the four types sensory expressions that are commonly used in tasting notes. The models are then applied to a large corpus of wine tasting notes to predict the characteristics of wine regions and provide visualization of the degree of sensory expression used in the review of wines in the region. In chapter 9, we examine changes in emotional sensory expressions in wine over the span of 14 years from the perspectives adjective antonym pairs that represent opposites of expressing attributes. A visualization system is proposed for analyzing differing properties, and examples where the change in use are examined in detail. Chapter 10 presents a detailed analysis of the adjective antonym meaning dimension of THICK and THIN.

Chapter 2

Preliminaries

This chapter provides a preliminary explanation of: the writing error categories that are used in the analysis of the expression of intention in chapters 3 to 7, and the sensory expression viewpoints of wine tasting notes that are used in the analysis of the expression of emotion in chapters 8 to 10.

2.1 Writing Error Categories

Foreign language writing experiments are often conducted in controlled environments as outside influences can have an impact on the performance of works produced by learners. Most of the previous research in this field has aimed to control these factors by undertaking experiments in academic settings. This has enabled researchers to control the subject of the works produced by the learners, and other factors such as time limits and environment.

Kroll [41] investigated the difference between writings that were produced in the highly controlled environment of a classroom and those that were produced at home where time was not limited and the learner could have more time to think about their composition. Kroll hypothesised that students might be able to produce better writing in an environment in which they have less pressure and more time to think about the task at hand. An experiment was conducted and the essays of foreign language writings were graded by the frequencies of errors categories that occurred. These frequencies were then compared and it was found that there was not a statistically significant difference between the writings produced in the different environments.

Polio and Fleck [54] examined whether additional revisions of essays influenced the linguistic accuracy of the content as it is theoretically interesting to researchers in the areas of second language acquisition and second language writing pedagogy. Polio and

Fleck [54] built on the error categories used in previous research reported by Kroll [41]. However Polio and Fleck [54] concluded that the practical implications in the context of writing assessment might be too small.

Weltig [71] investigated the influence of writing error categories on the scores of essays by foreign language writing learners. This research built on the error categories that were used in the two previously introduced works by Kroll [41] and Polio and Fleck [54]. A combination of their defined error categories was used. Weltig [71] introduced additional error categories as it was thought they could have an influence on the scoring of writings, such as spelling errors and punctuation errors. The results of the investigation revealed that certain errors do have a greater influence on the overall score attributed to the writings. As the error categories of these researches utilize a different set of error number lists for their analysis, a merged error number list was created as seen in Table 2.1. This merged list was used as a basis for the error categories analyzed in this thesis.

A main source of data that is analyzed to investigate the expression of Intention in this thesis is a set of writings that were collected from a leading SNS-based mutual correction website, Lang-8¹. For initial analysis, we collected 500 corrected sentences at random that were written in English from the diaries of language learners on Lang-8. These were then filtered for inconsistencies, such as: containing multiple languages, or large corrections and rewrites which are not effective for machine learning, leaving 399 candidate sentences. The sentences were then manually checked for errors and tagged in the corpus accordingly. The feedback provided by native speakers often contained several different error pattern corrections within a single response. Taking this into consideration, sentences that contain more than one error type were categorized as having multiple error patterns accordingly. Some feedback contained comments about the correction and/or multiple suggestions for a single word or phrase that were mostly to do with lexical or phrase choices and categorized as such.

Linear regression analysis was used to establish whether a correlation exists between the frequency of errors in the common categories of previous studies [41, 71] and that of the manually tagged Lang-8 corpus. As shown in Figure 2.1 and Table 2.2, the results of the analysis show that there is a significant correlation, with a critical alpha level of $p < 0.05$, and $t = 4.3509$, 4.4179 , and 3.8011 for Kroll Class, Kroll Home, and Weltig, respectively.

¹<http://www.lang-8.com>

Table 2.1: Merged error number list used for analysis.

	Error #	Kroll #	Weltig #	Error Category
Sentence structure errors	1	1	1	Whole sentence or clause aberrant
	2	2	2	Subject formation
	3	3	3	Verb missing
	4	4	4	Verb complement/object complement
	5	5		Prep. Phrase/infinitive mixup
	6	6	5	Dangling modifier
	7	7	6	Sentence fragment
	8	8	7	Run-on sentence
	9	9	8	Parallel structure
	10	10	9	Relative clause formation
	11	11	10	Word order
	12	12	11	Gapping error
	13	13	12	Extraneous words
	14	14		Awkward phrasing
	15			13
Verb-centered errors	16		14	Wrong modal
	17	15	15	Tense
	18	16	16	Voice
	19	17	17	Verb formation
	20	18	18	Subject-verb agreement
	21	19	19	Two-word verb
Reference errors	22	20	20	Noun-pronoun agreement
	23	21	21	Quantifier-noun agreement
	24	22	22	Epenthetic pronoun
	25	23	23	Ambiguous/unlocatable referent
	26	24		Voice shift
	27		24	Wrong case
Word-level choice	28	25	25	Lexical/phrase choice
	29	26	26	Idiom
	30	27	27	Word form
	31		28	Noun phrase morph
	32		29	Comparative formation
	33	28	30	Singular for plural (except verbs)
	34	29	31	Plural for singular (except verbs)
	35	30	32	Quantity words
	36	31	33	Preposition
	37		34	Genitive
	38	32	35	Article errors
	39		36	Deixis problem
	40	33	37	Punctuation
	41		38	Negation
	42		39	Spelling
	43		40	Possessive

Table 2.2: Linear regression analysis results.

	Kroll (Class)	Kroll (Home)	Weltig
r^2	0.6351	0.6409	0.5834
t	4.3509	4.4179	3.8011
p	0.0002	0.0001	0.0007
y	$2.9376 + 4.2918x$	$4.9722 + 3.6384x$	$7.2613 + 21.1171x$

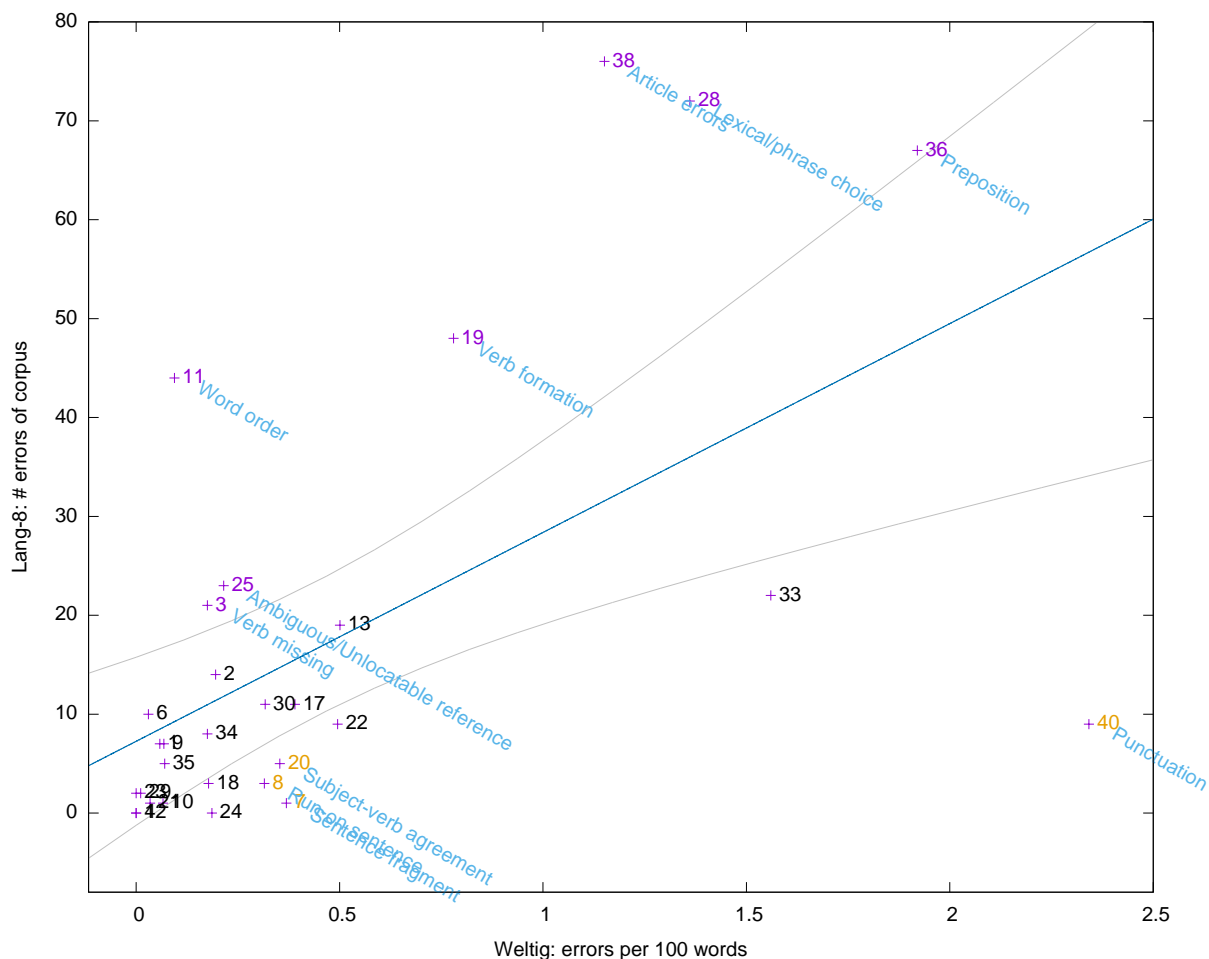


Figure 2.1: Error correlation of Lang-8 vs Weltig.

Table 2.3: Outlier error categories and relation to Lang-8 error frequency.

More frequent in Lang-8		Less frequent in Lang-8	
#	Error Cat.	#	Error Cat.
3	Verb missing	7	Sentence fragment
11	Word order	8	Run-on sentence
19	Verb formation	20	Subject-verb agreement
25	Ambiguous/Unlocatable reference	40	Punctuation
28	Lexical/phrase choice		
36	Preposition		
38	Article errors		

These correlations were then used to identify possible outlier errors not residing within the 95% confidence interval. A total of 22 different error categories were found outside the 95% confidence interval, with 11 of these errors being common across all three regression analyses. These common outlier errors suggest a characteristic difference in the error frequency of writings and corrections on Lang-8 when compared to those from an academic setting, such as: Kroll and Weltig. This may be a result of the differences in influencing factors, such as: motivation, the subject of the writing, and personal factors (age, socioeconomic background, etc).

As seen in Table 2.3, seven error categories occur more frequently on Lang-8 when compared to results from previous work by Kroll and Weltig. Of these, the error categories “Word order”, “Verb formation”, “Preposition” and “Article errors” are considerably outside the 95% coincidence interval, and therefore could be seen as a characteristic of the types of errors that occur in writings on Lang-8. In Chapter 3, we examine the classification of 15 frequently occurring error categories, of which the seven characteristic Lang-8 error categories are also included.

2.2 Sensory Viewpoint Expressions of Winespeak

Wine tasting notes contain expressions that are not used in common English, and therefore represent a special purpose use of language. There is a growing body of cognitive linguistic research into expressions that are used to describe the intricate emotional sensory responses to wines. From this body of research, we selected the four sensory modalities that were defined by Paradis and Eeg-Olofsson [51] as the basis of the sentimental analysis presented in chapters 8 to 10 of this thesis that investigate the complex emotional

Table 2.4: 17 Example words that describe sensory modalities as defined by Paradis and Eeg-Olofsson [51]

Modality	Example keywords words
Vision	purple, ruby, straw, gold, light, dark
Smell	fruity, floral, spicy, smoky, weak
Taste & Touch	flabby, soft, heavy, thin, long, crisp

expression. The four sensory modalities and example keyword words that were defined are shown in Table 2.4, and consists of 17 different sensory modalities descriptors that are divided into four main viewpoints. TASTE and TOUCH are reportedly difficult to separate, and therefore are assigned the same descriptor keywords. In the research presented in this thesis, all four sensory viewpoints are analyzed independently, with both TASTE and TOUCH qualities being assigned in the case where the differentiation between the two viewpoints is ambiguous.

Chapter 3

Classification of English Language Learner Writing Errors

In order to overcome mistakes, learners need feedback to prompt reflection on their errors [44]. This is a particularly important issue in education systems, as the system effectiveness in finding errors or mistakes could have an impact on learning. Finding errors is essential to providing appropriate guidance in order for learners to overcome their flaws. Traditional classroom-based language study has offered interaction with other learners and feedback from teachers and peers.

In the last decade or so with the global spread of the Internet, the number of people studying languages on the web has increased. Of particular interest are sites that offer a social or collaborative approach to study languages, and are often based on a SNS (Social Networking Service) platform. To some extent these SNS-based websites offer feedback and interaction that might otherwise be absent in autonomous learners studies. Language learning SNS sites work on the language exchange function, where native speakers of the target language offer corrections and feedback to the language learners. In principal, these learners would then correct the writings of a learner studying their native language. For example, person A is a native Japanese speaker who is learning English as a foreign language and posts an English sentence on the website. Person B who is a native English speaker corrects the sentence. Person B is also learning Japanese as a foreign language and posts a sentence on the website in Japanese which is then corrected by person A. This mutually beneficial environment helps learners to achieve their respective goals of learning a foreign language, which in turn is another foreign language learner's mother tongue.

These websites contain numerous foreign language writings that have been created by

learners and corrected by speakers of the target language. It could be thought of as a crude crowd-sourced foreign language writing parallel corpus. Taking advantage of this data can help to further enhance the effectiveness of language learning through providing automated feedback and guidance. In this chapter we use machine learning to analyze the writings collected from a leading SNS-based mutual correction website, Lang-8¹.

As there have been remarkable advances in machine learning research recently, we propose that machine learning techniques could be used to automatically detect and classify the errors in foreign language writing sentences. A machine learning classifier model for error detection could be created and used to determine the characteristics of the learners' errors. To prepare data for machine learning, 500 corrected sentence pairs by learners of English foreign writing on the Lang-8 website were chosen at random. The corrected sentence pairs were then manually classified into error categories. These error categories were based on the previous research investigated by Kroll [41], Polio and Fleck [54], and Weltig [71] to examine the characteristics of foreign language writings. Using these error categories, we manually detected and classified the sample sentence pairs into error categories.

The raw sentence pairs from the Lang-8 website were marked up with tags that are supposed to represent the changes that have been made by English speakers providing feedback. These tags are applied by users and are not methodically implemented to indicate the inserted, deleted, or edited text. On further investigation we found that the tags did not accurately indicate the changes and therefore could not be used for the purpose of our intended research. To overcome this problem we processed the sentence pairs using an alignment algorithm [43] to extract the actual edits provided in the feedback by the English speaker. The results of this process were then used to re-tag the edits accurately. This data in conjunction with the words of the sentence pairs was analyzed for machine learning. The purpose of this chapter is to evaluate the prediction performance of using an SVM classifier to detect errors in English foreign language writing. The technique of searching for an optimal feature selection from Sakai and Hirokawa [55] is applied to the machine learning problem to enhance the classification performance of predicting errors that occur in language learner writings.

¹<http://www.lang-8.com>

3.1 Related Work

Various different methods have been proposed for the automatic analysis of foreign learner writing errors. Previous research into this problem has focused on the use of the following methods: analyzing the output of tools such as search engine hits and word processor grammar/spelling checkers, and the application of statistical and machine learning techniques.

Early research into the analysis of the output of tools was conducted by Koppel et al. [40], who applied the MS Word spell checker with a sentence tagger and an n-gram corpus to detect errors. The native language of ESL (English as a second language) learners was determined by stylistic text feature (function word selection, errors and syntax) analysis of their writings. Hirano et al. [26] used the frequency of results from a web search engine to check if a sentence from a technical paper contains an article error. It was stipulated that as the language used in technical papers is more complex than simple phrases, it is difficult to use a search engine to determine if there is an error or not as the number of search results is often too small to have any significance. It was proposed that using queries built based on the results of POS (parts-of-speech) tagging would better serve as a determiner if the sentence contains an error. More recently, Tanimoto and Ohta [61] examined using the number of search results as an indicator in an attempt to identify erroneous words in English sentences. NICE (Nagoya Interlanguage Corpus of English) was used in tri-grams and 4-grams as training data for SVM machine learning to create a model that can determine if an English sentence contains an error. There are some notable disadvantages of analyzing the output of tools that are not created specifically for the analysis of foreign learner writing errors: word processing tools that were created for native writing fail to detect errors unique to foreign language learners, and the results of search engines can be affected by indexed data from automated and low quality translations. In consideration of these limitations, the work presented in this thesis focuses on analyzing learner writing errors without relying on the output of the above mentioned tools.

Research into the application of statistical and machine learning techniques to learner writing error analysis has gained a lot of attention in recent years. Bailey and Meurers [2] examined the use of machine learning methods to augment feedback from computer-aided language learning systems by using the shallow matching features to detect meaning errors. They focused on the analysis of short answers to reading comprehension problems.

They achieved an accuracy of almost 90% for learner response content error detection on a learner corpus collected from real-life ESL learners completing assigned exercises. Brockett et al. [4] approached the problem by using techniques that are usually synonymous with phrasal statistical machine translation. They used a parallel corpus of texts that were made up of ESL learner writings with both pre and post-editing correction similar to that found on Lang-8. Some previous researches [24, 25, 8] have used maximum entropy classifiers to detect article errors (incorrect use of: A, an, the...). Parts of speech tags and local context words were used to determine the probability of noun phrases. This technique was found to be superior than past techniques, however it was noted that the classifier lacked the ability to determine the context of previously mentioned entities. Tetreault and Chodorow [63] also used a maximum entropy classifier augmented with combination features and a series of thresholds to detect preposition errors (incorrect use of a word expressing the relation between a noun/pronoun and another word or element in the phrase). It was found that the system could detect up to 84% of preposition errors. A disadvantage of using this approach is that it cannot automatically model the interactions among features. Tsur and Rappoport [70] applied machine learning techniques to study the effect of language transfer, which is a major topic in second language acquisition (SLA). Language transfer studies the effect that a learner's native language has on foreign language study. They hypothesized that language transfer affects the level of basic sounds and short sound sequences, manifested by the words that people choose when writing in a second language. Thus, foreign language words are strongly influenced by native language sounds and sound patterns. They applied SVM machine learning to train a classifier using the International Corpus of Learner English (ICLE) in an effort to realize the hypothesis. The use of n-grams for the classification of texts has featured numerous times in the previous research for both the general classification of texts and also detection of errors. Schwarm and Ostendorf [57] and Petersen and Ostendorf [53] used n-grams combined with support vector machine classifiers to find appropriate reading material for students according to their reading level. Gamon et al. [23] used decision trees to perform error detection and correction for prepositions and definite/indefinite determiners on a reduced feature set using an n-gram corpus. Overall evaluation of the system was positive in providing error detection and also suggesting a correction. It was noted that the biggest challenge was solving false positives as it can confuse non-native speakers. Other research has focused on the development of corpora to overcome the limitations of past resources.

Sugiura et al. [58], discuss corpus design and reviewed the International Corpus of Learner English (ICLE). Based on the corpus weaknesses identified, they set about compiling a new English learner corpus and a parallel corpus of native English speakers, called NICE (Nagoya Interlanguage Corpus of English). They performed analysis on the NICE corpus using mechanical text features, such as: type, token, number of sentences, and average word length to compare the language learners performance with native speakers. Miki [46] looks at the use of a parallel corpus that is constructed using the essay writings of foreign language learners and exact forms of the sentences that are provided by native English language speakers. NICE was used as a dataset to examine how Japanese English language learner’s use “I think” in comparison with native speakers. Unlike other studies on the over usage of expressions which focus on quantifying the errors, by using a parallel corpus they were able to determine how the expression was being inappropriately used to augment the language learners writing. Miyake et al. [47] also used the same method and NICE parallel corpus to examined the use of “there” with the long-term intention of identifying the “Japaneseness” and “nativeness” relating to the use of constructions.

Previous work has mainly focused on the analysis of professionally curated corpora that were created in a controlled academic setting. In this chapter, we analyze foreign learner writing errors by SVM analysis of a parallel corpus that was created with corrected writing collected from a language learning SNS. The main contribution of this work is the use of feature selection to not only improve the classification performance, but also identify a smaller optimal subset of characteristic error features that can be extracted from SNS learner writing. The work presented in this chapter also serves as the basis for subsequent contributions presented in Chapter 4 where we examine the relation of native language and error co-occurrence by clustering analysis, and Chapter 5 in which we propose a unique technique of predicting a learners native language by analyzing the output of learner writing error detection.

3.2 Vectorization of Error Sentences for Categorization

In order to evaluate the classification of errors in English sentences, the following process was undertaken to construct basic data. Firstly, 500 corrected sentences written in English were chosen at random from diaries written by language learners on the Lang-8 website.

Table 3.1: An example of an original and corrected sentence pair.

Original Sentence	I woke up alone, with lose memory, lying on the white beach, not knowing where I was.
Corrected Sentence	I woke up alone, with no memory, lying on a white beach, not knowing where I was.

However, in some cases large portions of the sample sentences were rewritten or contained comments that would reduce the effectiveness of machine learning and were removed, leaving 399 candidate sentences.

Analysis was performed not on just the sentences, but on pairs of sentences: the original sentence that contains errors, and the corrected sentence that contains tagged edited words. These sentence pairs are a result of mutual corrections that have occurred on the Lang-8 website. In this chapter, the GETA search engine² was used to index the original and corrected sentence pairs. Word is usually stemmed when building an index, however it was decided that the indexed words should not be stemmed as analysis was performed at the word level. In Lang-8, the edits made by English speakers on the sentences are marked up using span tags, such as ``. The class attribute of these span tags changes depending on action of the English speaker. If a word is removed then the `sline` class is applied. Classes that describe the font colour and weight are also used, such as `f_bold`, `f_red`, and `f_blue`. However the intention with which these classes are assigned is unregulated and not uniformly applied across the all sentences. In this chapter, it was decided that because of the inconsistency of tag use that better results would be achieved by using an alignment algorithm to programmatically detect and tag changes in sentence pairs. Table 3.1 shows an example untagged sentence.

As seen in this example, “lose” and “the” are corrected with “no” and “a”. These corrections are identified using the alignment algorithm and the results are tagged as: `delete:lose`, `delete:the`, `insert:no`, and `insert:a`. In the search engine that was used in this chapter the corrections are expressed as `d:lose`, `d:the`, `i:no`, and `i:a` along with the other words in the sentence. The corrections were also added without distinguishing whether the edit is an insertion or deletion, and were indexed as: `e:lose`, `e:the`, `e:no`, and `e:a`.

These sentences were classified into 42 error categories by the author of this thesis whose language is English. It was determined that the above example contains errors of two categories: Error number 38, which is an article error, and error number 41,

²<http://geta.ex.nii.ac.jp>

Table 3.2: Indexed example sentence.

c:38/ c:41
d:lose/ d:the i:no/ i:a
e:the /e:lose/ e:a /e:no
the/ a/ woke/ no/ not/ on/ white/ memory/ with/ lying/
beach/ up/ i/ knowing/ where/ alone/ was/ lose/

which is a negation error. These errors are indexed in the search engine as c:38 and c:41 respectively. The three indexes for error category, edited words and non-edited word are then vectorised. Using this it is then possible to determine if a sentence has an article error by examining if it contains “i:a, d:the, e:a, and e:the”. It also makes it possible to determine if the sentence contains a negation error by checking if it contains “i:no, and e:no”. Simple classification would analyze just the words of the sentence. However we analyse the information about the corrections along with the words of the sentence to determine the error categories with the sentence.

A special use search engine was built using indexes as shown in Table 3.2. The information about the error categories, c:38, c:41, was not used in the classification of error categories.

3.3 Evaluation of error categorisation using SVM

An evaluation of error categorisation using SVM to classify the errors of 399 sentences with all the data as training data is shown below in Table 3.3. It should be noted that the columns in this table are sorted by F-measure in descending order. The prediction performance of the classification of errors 36 (preposition), 42 (spelling), 2 (subject formation) and 28 (lexical/phrase choice) is more than 90%. However, as this evaluation analyses all the data as training data it cannot be used as a general evaluation of the prediction performance.

We then used ten-fold cross-validation to evaluate the prediction performance of the classifier. All 399 sentences were then randomly divided into 10 even groups. In each group 90% of the data was used for SVM training to generate a model. The prediction performance of the classifier was then tested using the remaining 10% of the data from the same group. The average of ten test results for each error category is used as a measure of the prediction performance of each classifier respectively. These results are displayed

Table 3.3: Evaluation of the classification of error categories.

Error Category	Precision	Recall	F	Accuracy
36	0.9310	0.9643	0.9474	0.9850
42	0.9773	0.8958	0.9348	0.9850
2	1.0000	0.8571	0.9231	0.9950
28	0.8696	0.9677	0.9160	0.9724
38	0.2698	1.0000	0.4250	0.5388
19	0.1845	1.0000	0.3116	0.5238
11	0.1201	1.0000	0.2145	0.3208
33	0.0955	1.0000	0.1743	0.5013
25	0.0806	1.0000	0.1493	0.4286
3	0.0599	1.0000	0.1131	0.2531
17	0.0521	1.0000	0.0990	0.5439
13	0.0492	1.0000	0.0939	0.3709
6	0.0488	1.0000	0.0930	0.5113
37	0.0478	1.0000	0.0913	0.5013
30	0.0461	1.0000	0.0881	0.4812

in Table 3.4, Figure 3.1, and Figure 3.2.

Table 3.4 shows the overall results of the tests along with the number of sentence samples for each error category. The table is sorted by the F-measure of each of the models in descending order.

The F-measure performance of each model is displayed in Figure 3.1. As you see, the F-measure of all the models is less than 40%, with error category 42 (spelling), 28 (lexical/phrase choice) and 38 (article errors) being the more effective models with an F-measure of only 38.07%, 36.72%, and 36.52%, respectively. On the lower end of the scale the model for error category 6 (dangling/misplaced modifier) has an F-measure of 1.05%.

The accuracy of the generated models also varies for each error category. As shown in Figure 3.2, the model for error category 42 (spelling) has the greatest in all the models with an accuracy of 77.80%. Error category 3 (verb missing) has the lowest accuracy in all the models at 26.47%.

Overall, the prediction performance of the classifier as seen above cannot be considered effective enough for practical use. Figures 3.3 and 3.4 are plots of correlations between the number of samples, F-measure, and accuracy for each of the error category models. A positive correlation can be seen in both plots, indicating that as the number of samples increases so does the F-measure and accuracy of the evaluation. This suggests that if the

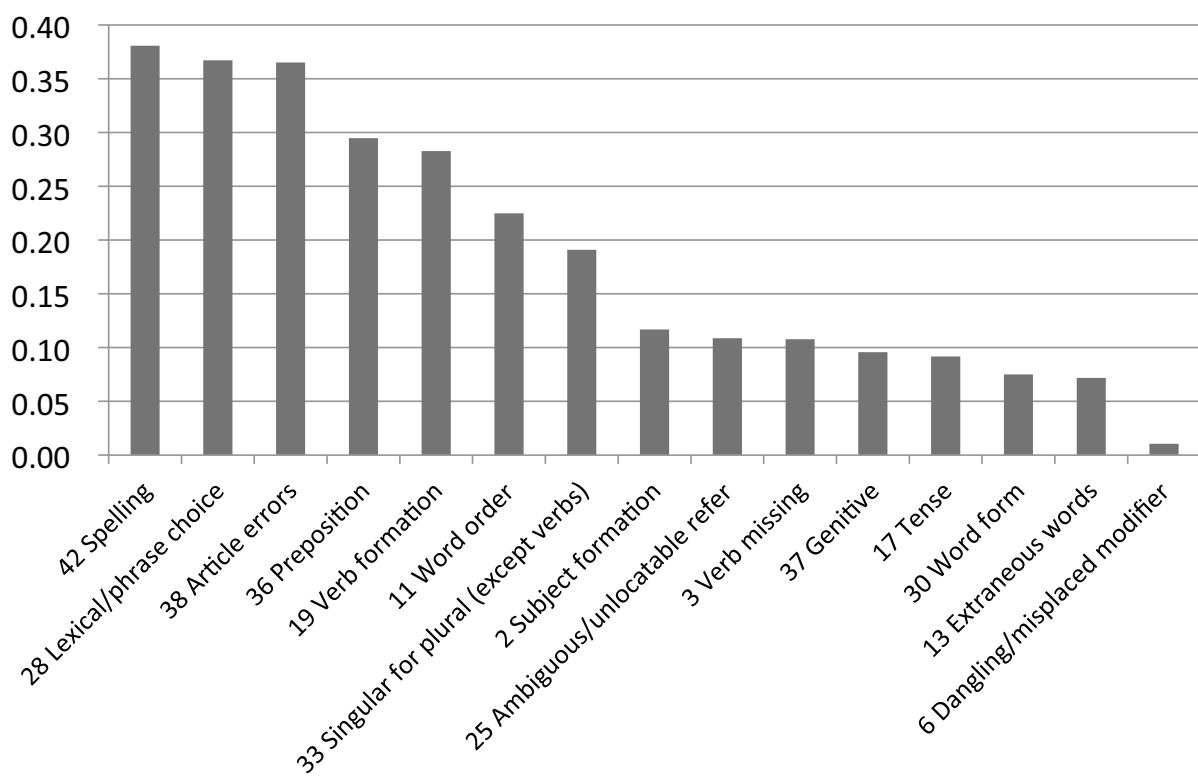


Figure 3.1: Error classification evaluation for each category (F-measure, 10-fold cross-validation).

Table 3.4: Evaluation of the classification of errors into categories by 10-fold cross-validation.

Error Type	Number of Samples	Precession	Recall	F	Accuracy
42 Spelling	48	0.4153	0.3906	0.3807	0.7780
28 Lexical/phrase choice	62	0.3109	0.5206	0.3672	0.7218
38 Article errors	68	0.2265	0.9857	0.3652	0.4023
36 Preposition	56	0.2049	0.5742	0.2948	0.6288
19 Verb formation	43	0.1865	0.6881	0.2828	0.6547
11 Word order	37	0.1472	0.6514	0.2248	0.5999
33 Singular for plural	21	0.1129	0.8000	0.1910	0.5796
2 Subject formation	14	0.0758	0.3333	0.1169	0.5217
25 Ambiguous/unlocatable refer	20	0.0687	0.2833	0.1087	0.4843
3 Verb missing	19	0.0585	0.8250	0.1077	0.2647
37 Genitive	10	0.0539	0.4667	0.0957	0.4941
17 Tense	10	0.0588	0.4167	0.0917	0.3633
30 Word form	10	0.0418	0.3833	0.0750	0.4491
13 Extraneous words	12	0.0385	0.6500	0.0718	0.4516
6 Dangling/misplaced modifier	10	0.0063	0.0333	0.0105	0.5078

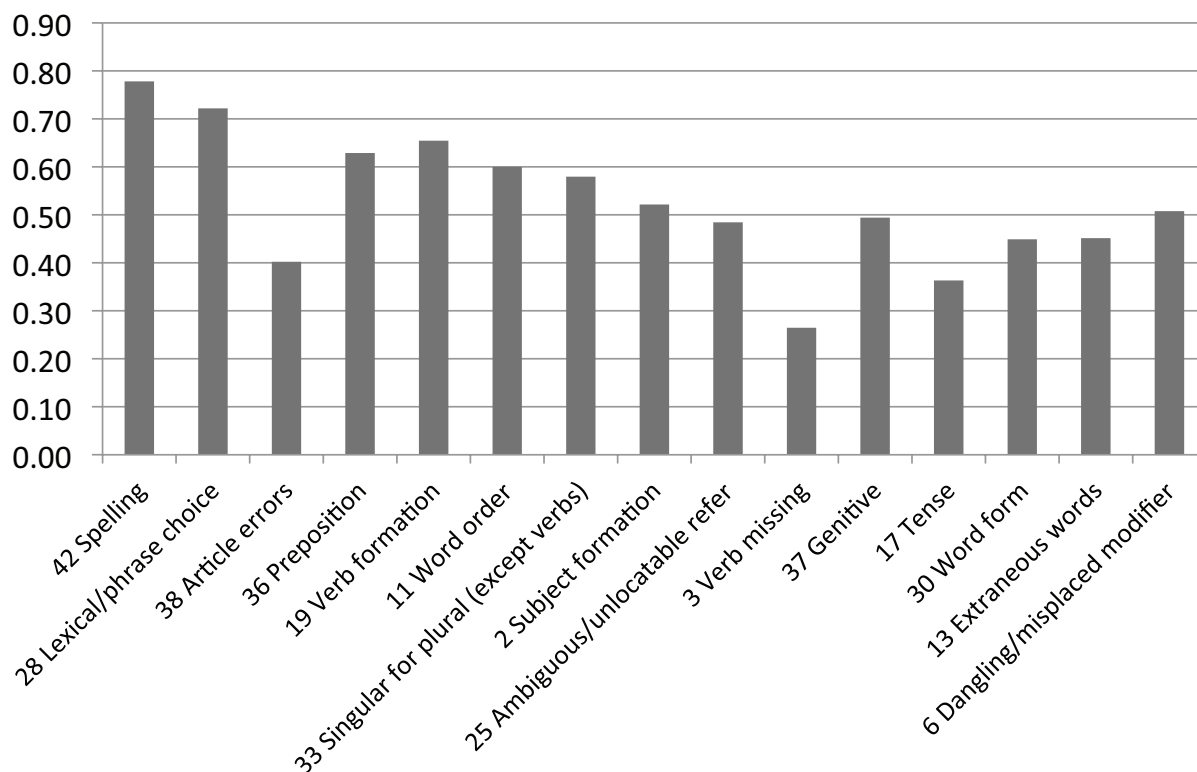


Figure 3.2: Error classification evaluation for each category (10-fold cross-validation, Accuracy).

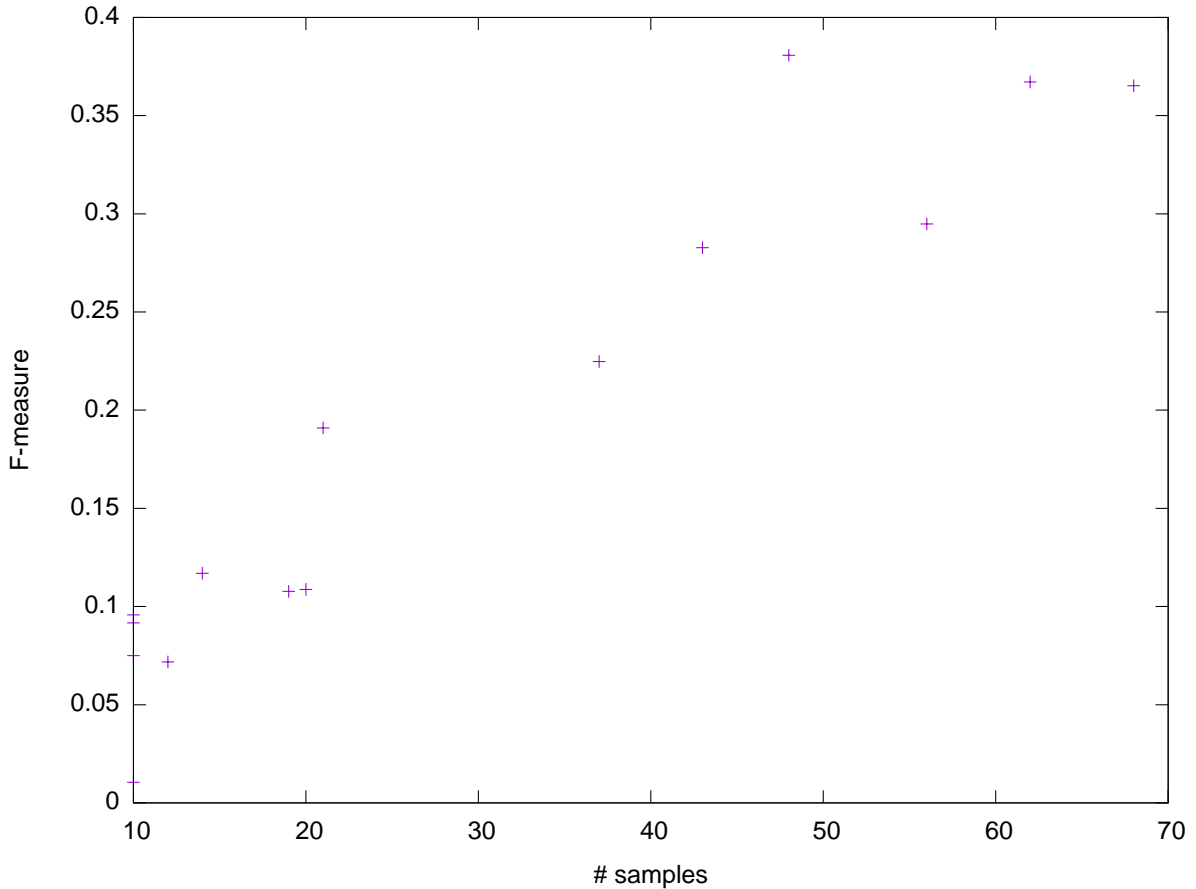


Figure 3.3: Correlation between the number of data samples and the F-measure of the evaluation.

samples for each error category were increased to an adequate number then the prediction performance of the classifier would also increase accordingly.

Looking at the results in Figure 3.3, the error category models that were trained using a small number of samples generally have a smaller F-measure than those with a greater number of samples. Therefore one can expect if 100 manually categorised samples were used to train each error category it would result in an F-measure of around 80%.

A similar correlation can also be seen in Figure 3.4 with the accuracy of models increasing along with the number of samples.

3.4 Optimal Feature Selection

We applied SVM^{light} using all of 399 data as training and test data to construct 15 models with respect to each error category. Then each model was applied to an imaginary

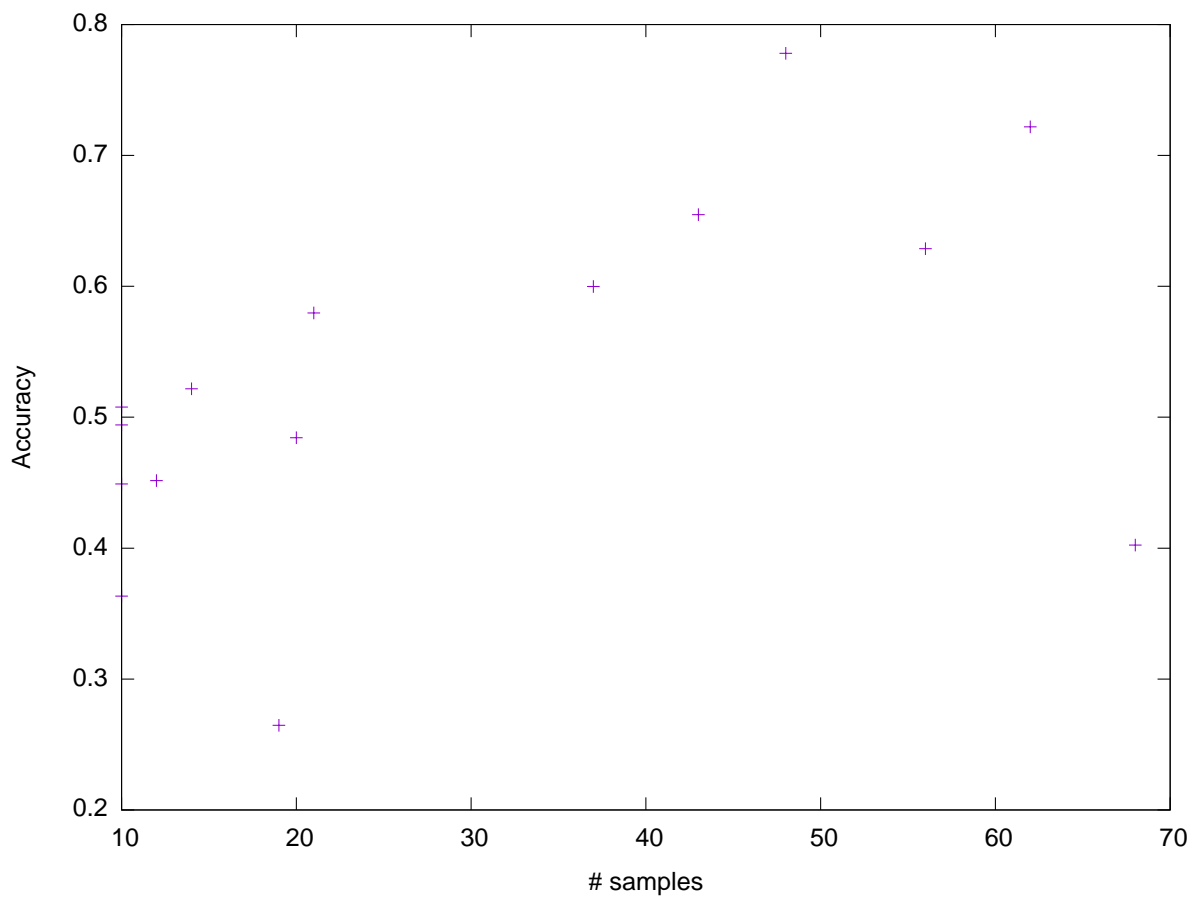


Figure 3.4: Correlation between the number of data samples and the Accuracy of the evaluation.

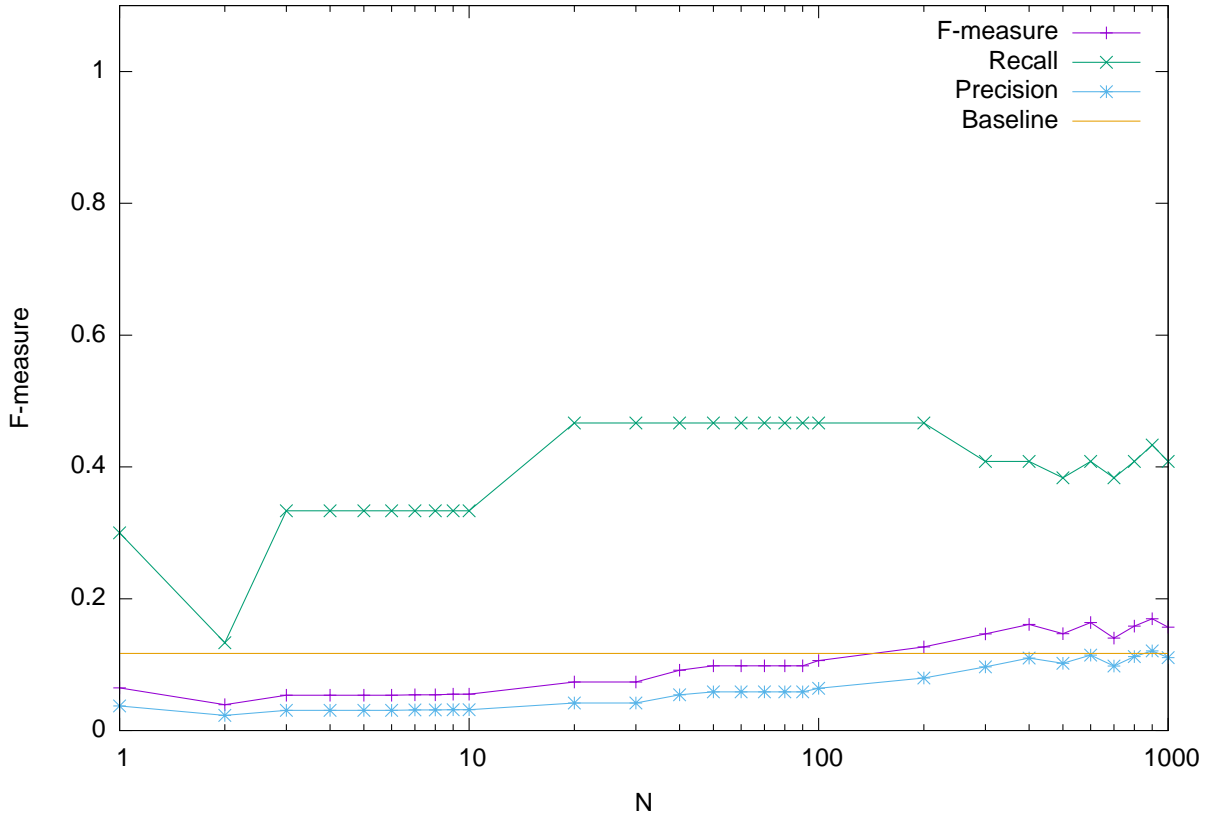


Figure 3.5: Effect of Feature Selection for Error Category 2.

sentence that consists of a single word. The score is used as the predicted score of the word with respect to the error category. The feature words with respect a category c is determined using the formula in Equation 3.1, where $Positive(c, N)$ is the set of words whose score for category c is positive and ranked with in the top N features, and $Negative(c, N)$ is the set of words whose score for category c is negative and ranked with in the bottom N features. The sentences are vectorized to include only word features within the selected set, and the classification performance is evaluated by F-measure with 10-fold cross validation. Then the optimal F-measure is evaluated among $N = \{1, 2, 3, \dots, 10, 20, 30, \dots, 100, 200, 300, \dots, 900, 1000\}$. Figure 3.5 and 3.6 are plots of F-measures for error categories 2 and 42, where the optimal choice is $N = 900$, and $N = 800$ respectively. We also investigated the existence of an optimal N for the other error categories.

$$FeatureWords(c, N) = Positive(c, N) \cup Negative(c, N) \quad (3.1)$$

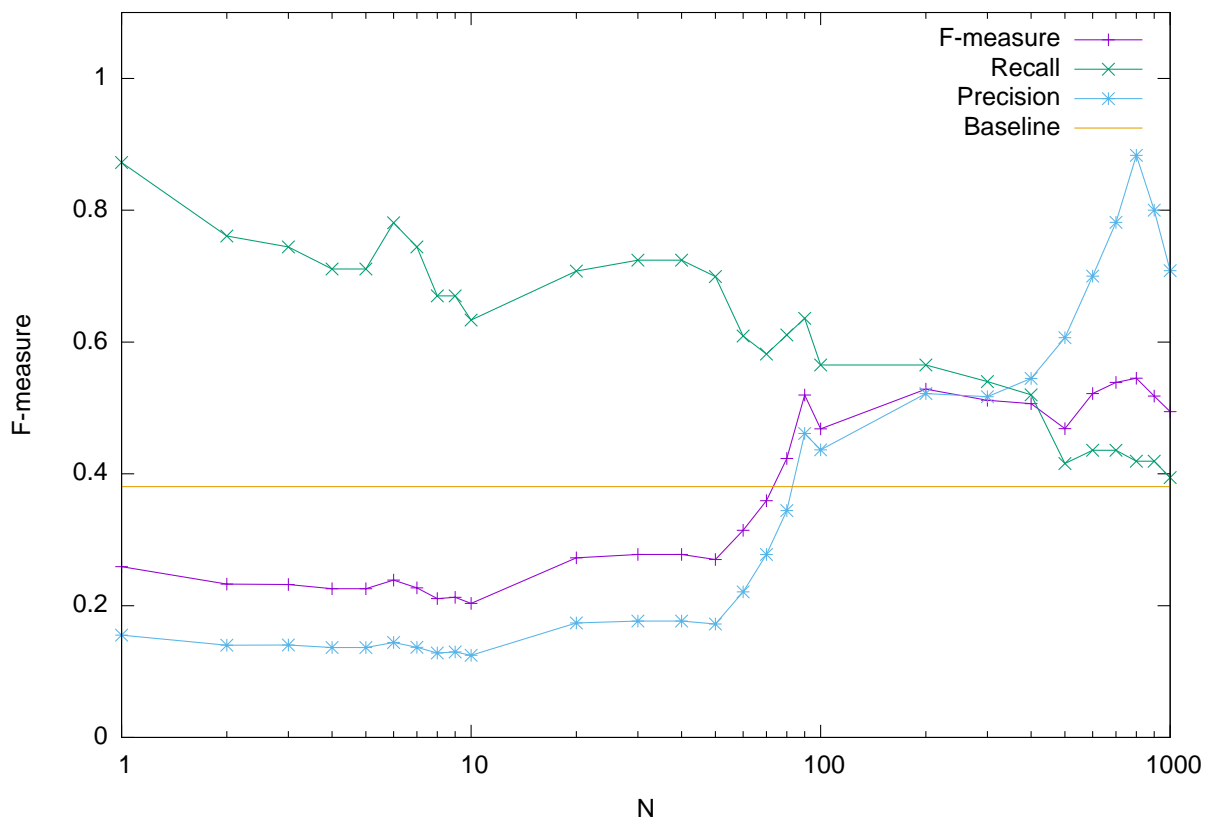


Figure 3.6: Effect of Feature Selection for Error Category 42.

Table 3.5: The words and tags from the model created using SVM.

Err	Feature words
42 Spelling	shopping e went e:e i:e phrase china day friend what
28 Lexical/phrase choice	which m it am would student in d:in here girl
38 Article errors	e:the i:the e:a the i:a a man e:A university e:This
36 Preposition	i:in e:in d:at at e:for e:at e:on on i:on two
19 Verb formation	i:ing e:ing ing didn e:to entrance d e:eat d:eating collage

3.4.1 Detailed Analysis

A score for each word or tag can be extracted from the model created by applying SVM to the training data. As shown in Table 3.5, error category 38 (article) has the features that consist of tags, such as “e:the, i:the, e:a, and i:a”. Error category 36 (preposition) has the following tags as the features of the error “i:in, e:in, d:at, e:for, e:at, e:on, and i:on”. The ability to extract such information from the model enables the confirmation of the features associated with the error types in the corrections. The feature “ing” can be expected for error category 19 (verb formation). The error features associated with error category 42 (spelling) are “e”, “e:e”, and “i:e” can be seen as common spelling errors in words such as conv-a-rsation, and ev[e]ryone.

3.4.2 Improvement of Prediction Performance by Optimal Feature Selection

Table 3.6 and Fig. 3.7 shows the F-measures by baseline and the proposed method with respect to the 15 error categories. The F-measures is greater than 0.4 in the five categories (19 Verb formation, 28 Lexical/phrase choice, 36 Singular for plural, 38 Article errors, 42 Spelling). In all cases, the prediction performance is improved.

3.5 Discussion

In this chapter, we manually classified the errors contained in sample sentences from diaries written in the mutual correction language-learning site Lang-8. The errors were classified into categories based on previous research [41, 71]. The sample sentence pairs had tags indicating the edits in the corrections, however it was determined that these did not always correctly reflect the true corrections, and were removed. An alignment algorithm was then used to programmatically identify the corrections that had been made,

Table 3.6: Prediction Performance Compared by F-measure

Category	Optimal N	Description	Feature Selection	Baseline
2	900	Subject formation	0.1695	0.1169
3	200	Verb missing	0.1490	0.1077
6	20	Dangling/misplaced modifier	0.0403	0.0105
11	500	Word order	0.2843	0.2248
13	300	Extraneous words	0.1552	0.0718
17	500	Tense	0.1040	0.0917
19	800	Verb formation	0.4508	0.2828
25	700	Ambiguous/unlocatable refer	0.1746	0.1087
28	200	Lexical/phrase choice	0.5001	0.3672
30	500	Word form	0.1172	0.0750
33	300	Singular for plural	0.3100	0.1910
36	700	Preposition	0.4688	0.2948
37	700	Genitive	0.2115	0.0957
38	500	Article errors	0.5264	0.3652
42	800	Spelling	0.5452	0.3807

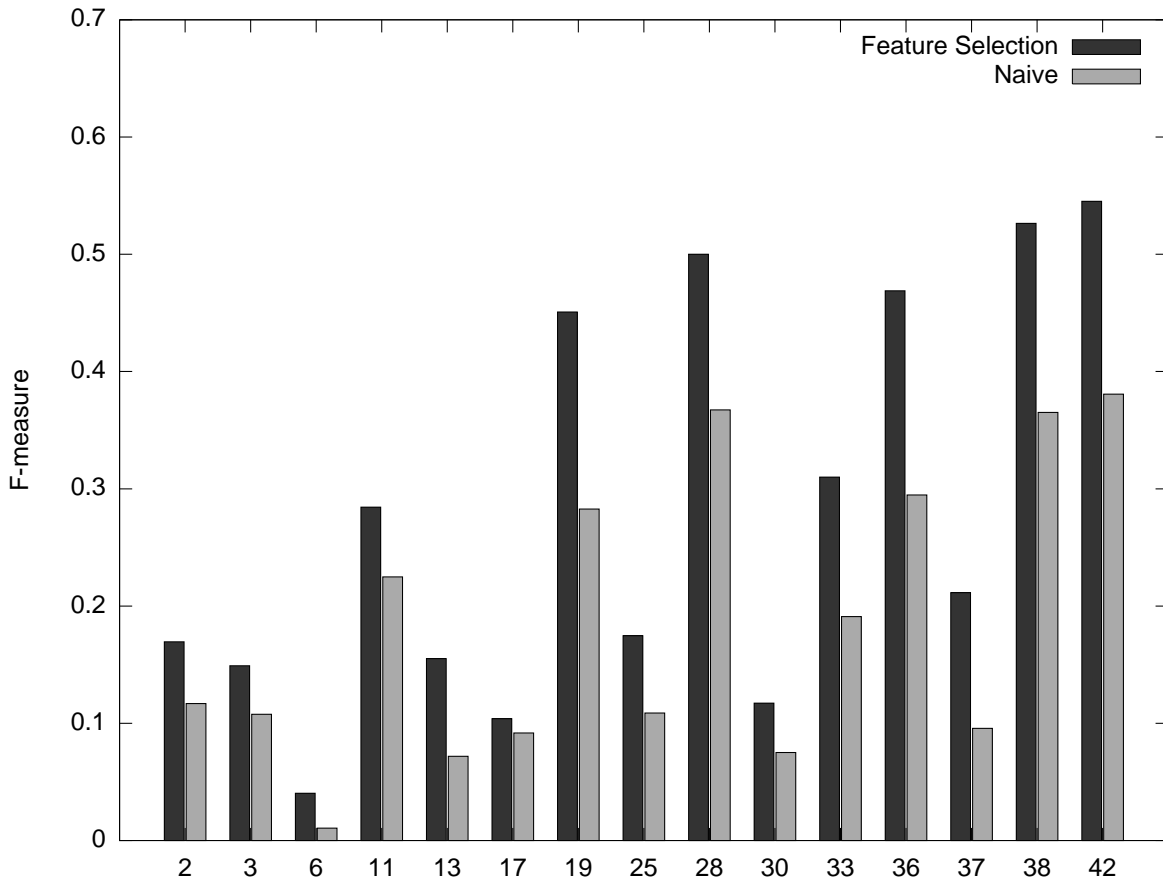


Figure 3.7: Comparison of Prediction Performance

and the edited words were then tagged as 'inserted' or 'deleted' accordingly. These tags, along with the manually classified error categories and the other words in the original sentence, were then indexed to build a special use search engine. This search engine index was then used as training data for SVM machine learning to create a model for error category classification.

This model was then evaluated using ten-fold cross-validation. 399 sentences used as sample data were divided randomly into ten even groups, with 90% of the sample data used for training and the remaining 10% used for model verification. The F-measure for each error category was less than 40%. However, the results did show a significant positive correlation between the number of data samples, F-measure and accuracy of the model. Thus it can be expected that if the number of samples is increased to 100 manually identified samples, then it is expected that the model will produce an F-measure of roughly 80%. Therefore by increasing the training data it is expected to produce a reasonable level of performance for error category classification. A method of optimal feature selection was examined to improve the classification performance. This improved the classification performance of all of the error categories. An F-measures performance of greater than 40% was achieved in the following five categories: 19 Verb formation, 28 Lexical/phrase choice, 36 Singular for plural, 38 Article errors, 42 Spelling.

Chapter 4

Clustering and Co-occurrence Analysis of English Writing Errors Based on Native Language

In this chapter, we examine the error characteristics of language learner writing from the perspective of their native language. Previous research into the analysis of learner writing errors based on native language, most notably the seminal work by Swan and Smith [60], have examined an error as a singular problem within a sentence. This analysis has uncovered the tendencies of different native language groups to make certain errors, and is based on the theory of language transfer where some expression is not present in the native language but used in the target language and vice versa. In comparison, the interaction between two or more errors that co-occur within the same sentence has received little attention. The aim of this chapter is to investigate the differences in the co-occurrence of errors of language learner writing between native language groups and propose a distance to measure the differences between the groups.

The 15 SVM error category models that were trained in chapter 3 are applied to predict the errors scores in 142,465 corrected sentences that were collected from the language learning SNS website, Lang-8¹. These scores are used as a vector representation of the sentences and divided into data subsets by native language of the learner as reported on Lang-8. The subsets are then clustered to analyze the co-occurrence and independence of foreign writing errors based on the native language of the learner.

¹<http://www.lang-8.com>

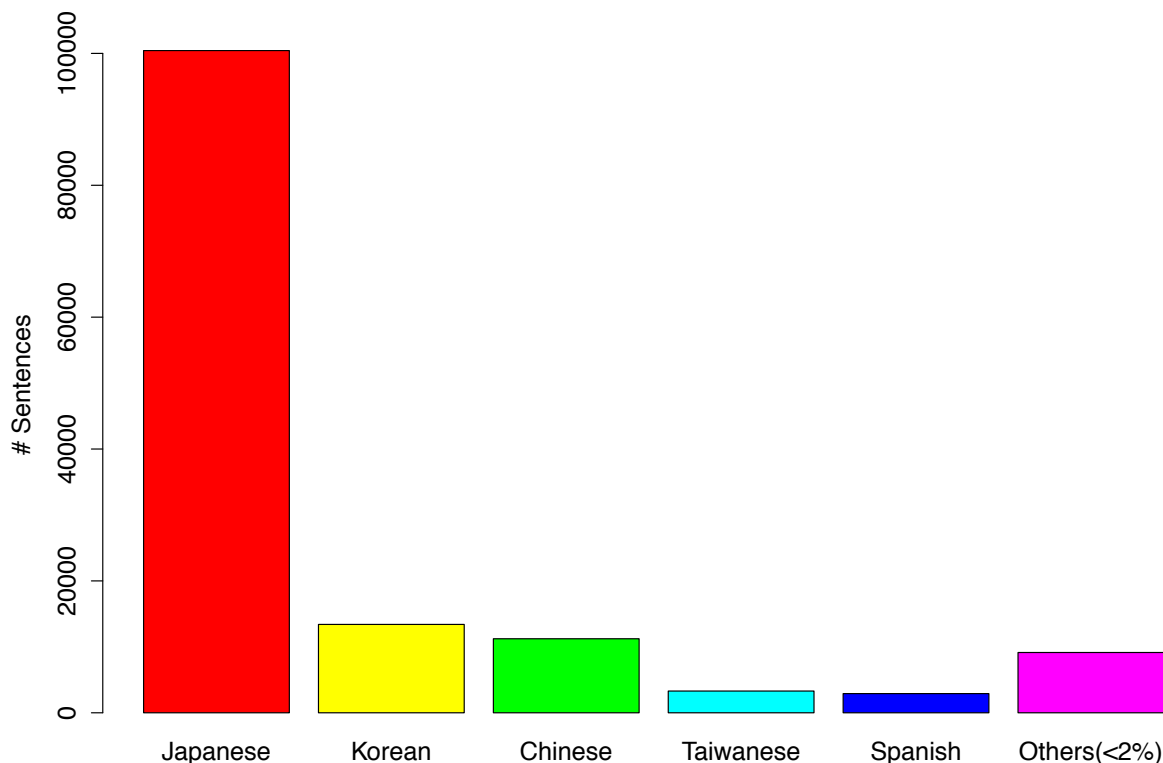


Figure 4.1: Number of sentences grouped by native language.

4.1 Clustering based on Error Category Prediction

4.1.1 Data Collection

Two sets of data were collected for analysis: a set of 142,465 sentences, posted on Lang-8 from October 9, 2011 to January 6, 2012, which are written in English and are corrected in some way. Each sentence is tagged to identify the native language of the author. Figure 4.1 shows the major groups within the collected data set, and it can be seen that Japanese users wrote roughly 100,000 sentences, which is the largest subset in the collected data.

Other main subsets include: Korean, Chinese, Taiwanese and Spanish. The analysis in this chapter will focus on these five major native language subsets within the collected data.

The second set consists of 399 corrected sentence pairs that have been manually classified into error categories. We analyzed the data to train and test SVM classifiers for 15 error categories in previous research [10, 14, 11]. The error category numbers and descriptions that were used are shown in Table 4.1. 10 models constructed for each error

Table 4.1: Error Category Numbers and Descriptions.

Category	Description
2	Subject formation
3	Verb missing
6	Dangling/misplaced modifier
11	Word order
13	Extraneous words
17	Tense
19	Verb formation
25	Ambiguous/unlocatable referent
28	Lexical/phrase choice
30	Word form
33	Singular for plural
36	Preposition
37	Genitive
38	Article
42	Spelling

category. Each sentence from the first set of data was scored with respect an error category is calculated as the average of the score of the result obtained by applying the 10 models. 15 scores corresponding to 15 error categories form a vector representation of a sentence.

4.1.2 Error Co-occurrence Analysis by Clustering

We have previously investigated the co-occurrence of errors from an overall perspective [11], and did not take into account other factors, such as the learners' native language, etc. The score vector representations of the sentences were analyzed by clustering into 20 clusters using the high-dimension clustering tool CLUTO [37]. This research identified co-occurring and non-co-occurring errors, an overview of which can be seen in the dendrogram (clustering tree) in Figure 4.2. The darker colored squares represent clusters of sentences with high averages in parts of the score vector. For example, cluster 0 has a high average score for error category 38 (Article errors), and cluster 3 has a high average score for error category 36 (preposition errors). On the vertical axis is a cluster hierarchy tree of the 20 resulting clusters. The clusters are leaf nodes of the tree where the number of sentences in the cluster is represented in the brackets next to the cluster ID. The error categories are represented on the horizontal axes at a cluster hierarchy tree. This visualization is very helpful to understand the huge amount of target data. We can see

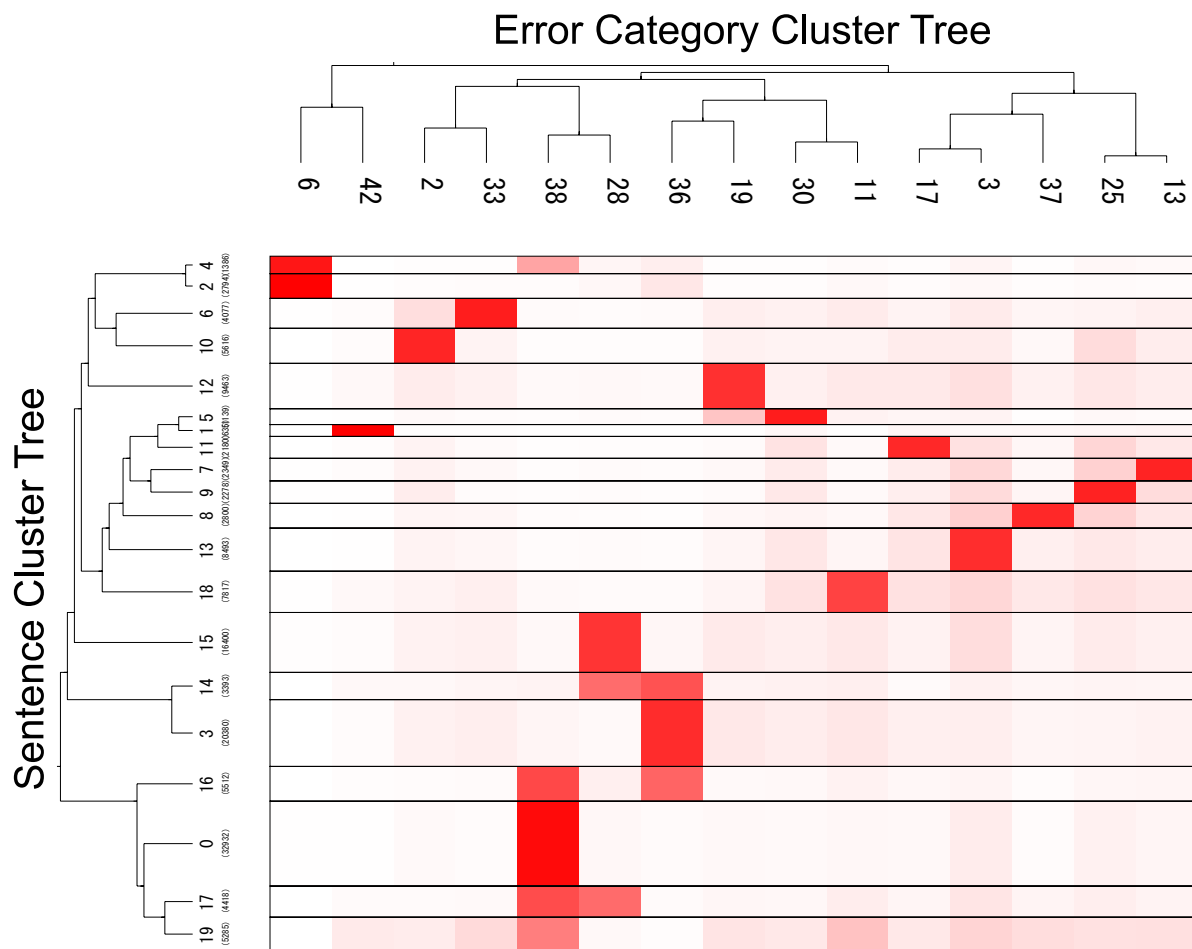


Figure 4.2: Clustering of all writing error data.

that the lower branch of the tree contains 1/3 of all the data and corresponds mainly to error category 38 (article errors). Cluster 0 is a core part of this branch whose sentences contain mainly article errors. Cluster 17 contains lexical or phrase choice errors (category 28). The cluster 16 contains preposition errors (category 36). Thus, the tree represents not only the clustering of sentences but also the clustering of error categories. Indeed, we can interpret that article errors (category 38) are the largest errors and occur with preposition error (category 36) and lexical/phrase choice error (category 28). For further details please refer to [11].

While presenting this previous research, it was recommended by an attendee that it is important to perform the analysis of error characteristics with regard to the native language of the learner. Therefore this chapter investigates co-occurring error categories by native language.

4.2 Co-occurrence Analysis By Native Language

4.2.1 Principal Component Analysis (PCA)

To investigate if there are any underlying correlations between native languages and the predicted error category scores, we analyzed the score vector and the native language of the learners using Principal Component Analysis (PCA).

The results of the PCA are shown in Figure 4.3, where there is a slight association between error category 36 (preposition) and Japanese when compared to other languages: Korean, Chinese, Taiwanese and Spanish. However other than this observation there are no significant correlations between error categories and native languages.

4.2.2 Error Co-occurrence Analysis of Native Languages by Clustering

The dataset of predicted error category score vectors was divided into subsets based on the native language of the learner. The top five native languages by number of sentences were then clustered into 20 clusters to analyze possible differences in error co-occurrence. Figure 4.4 shows the clustering results for sentences written by Japanese natives.

Clusters 15 and 16 are both made up of sentences that contain error categories 36 (preposition) and 28 (lexical/phrase choice) respectively that co-occur with error category 38 (article). Clusters 12 and 18 are also both made up of sentences that contain error category 36 (preposition) that co-occur with 19 (Verb formation) and 28 (Lexical/phrase choice). Other notable co-occurrences are seen in cluster 1 which contains error category 6 (Dangling/misplaced modifier) and 38 (article), and cluster 17 which contains error category 42 (spelling) and 30 (word form).

The results of the clustering analysis of Korean natives, as shown in Figure 4.5, share some similarities with Japanese natives. It contains the same co-occurring errors as Japanese, except instead of error categories 19 (Verb formation) and 36 (preposition) co-occurring, Korean has a more prominent co-occurrence between error category 11 (word order) and 36 (preposition).

For Chinese natives the results displayed in Figure 4.6 have some similarities to the results for Korean and Japanese as error category 38 (article) co-occurs with 6 (Dangling/misplaced modifier), 28 (Lexical/phrase choice), and 36 (preposition), and also error category 28 (Lexical/phrase choice) with 36 (preposition).

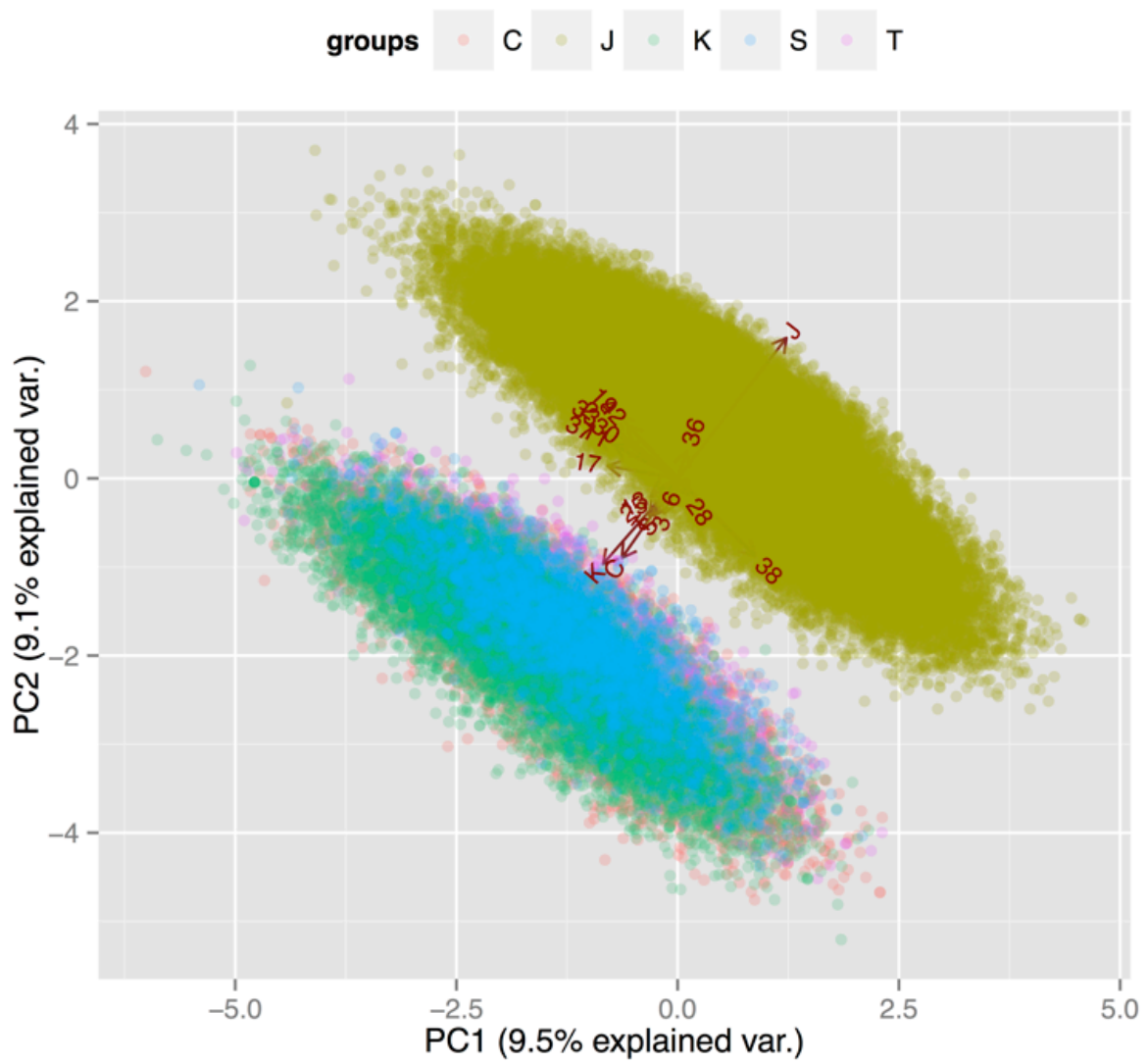


Figure 4.3: Principal Component Analysis of all writing error data.

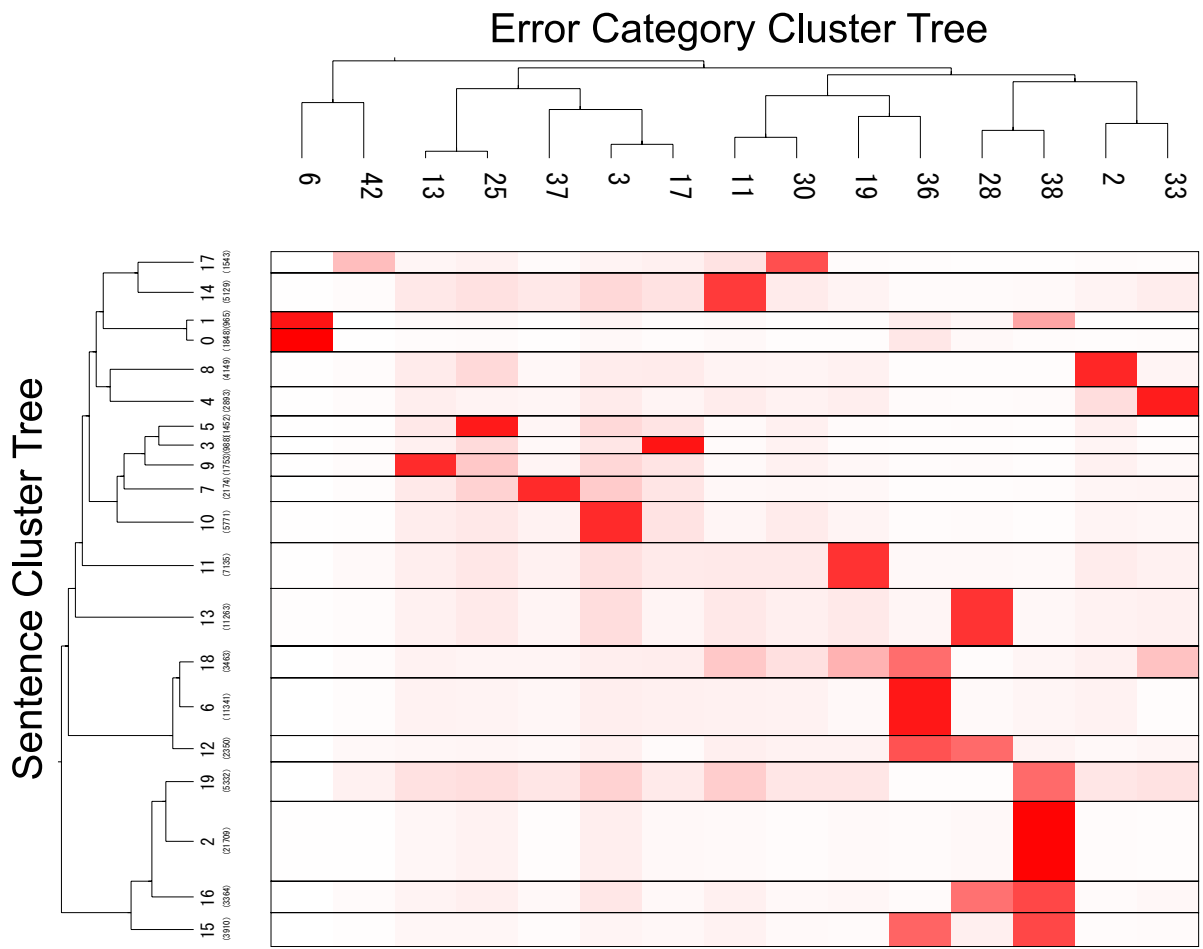


Figure 4.4: Clustering of writing errors by Japanese natives.

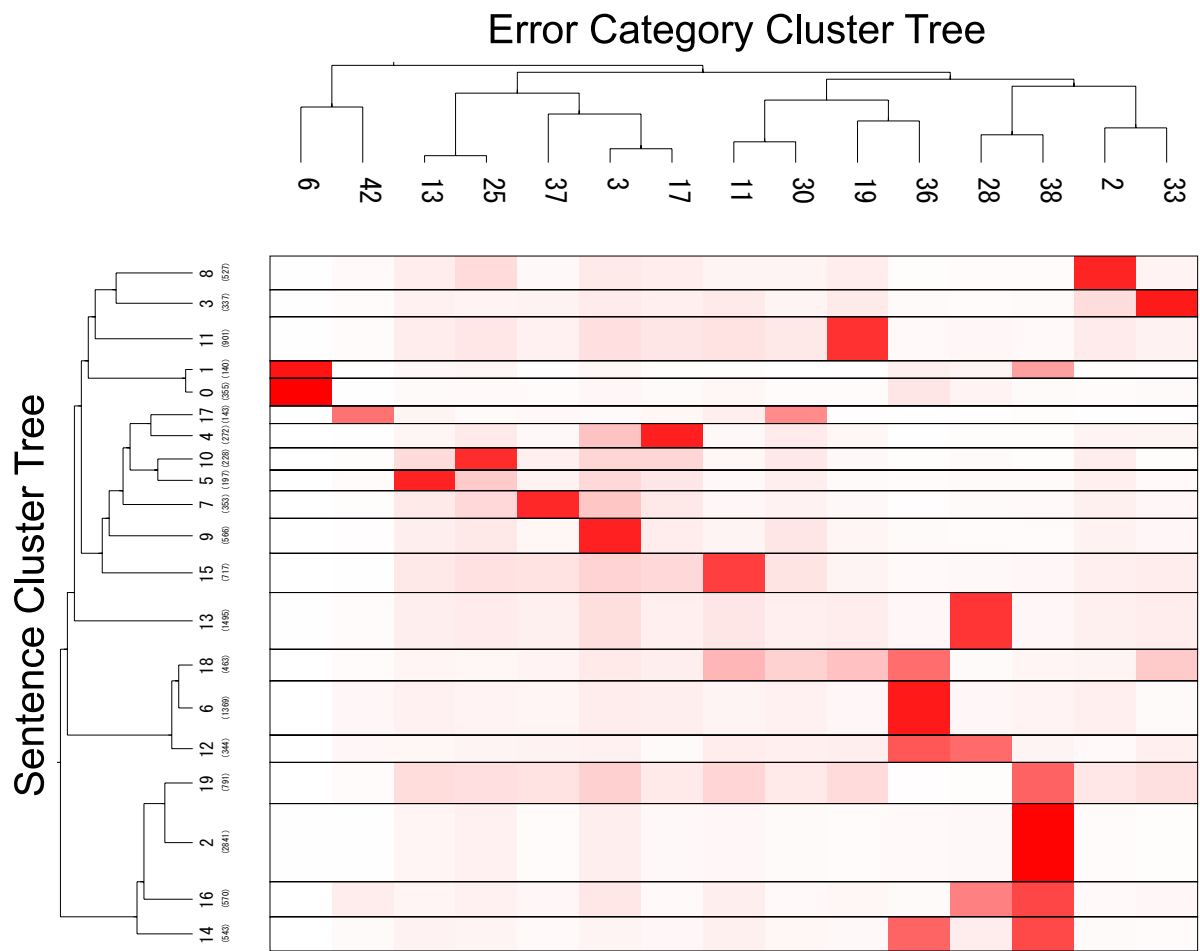


Figure 4.5: Clustering of writing errors by Korean natives.

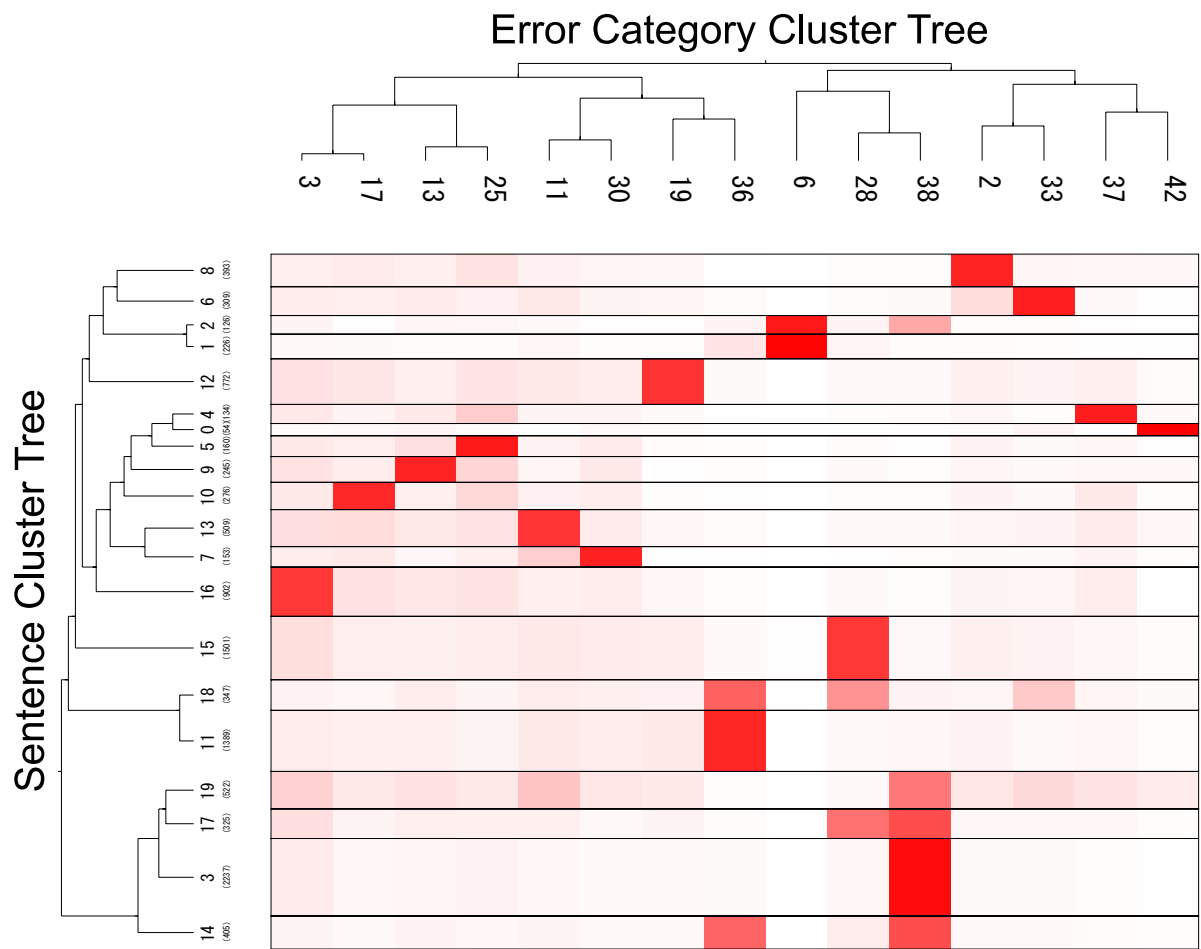


Figure 4.6: Clustering of writing errors by Chinese natives.



Figure 4.7: Clustering of writing errors by Taiwanese natives.

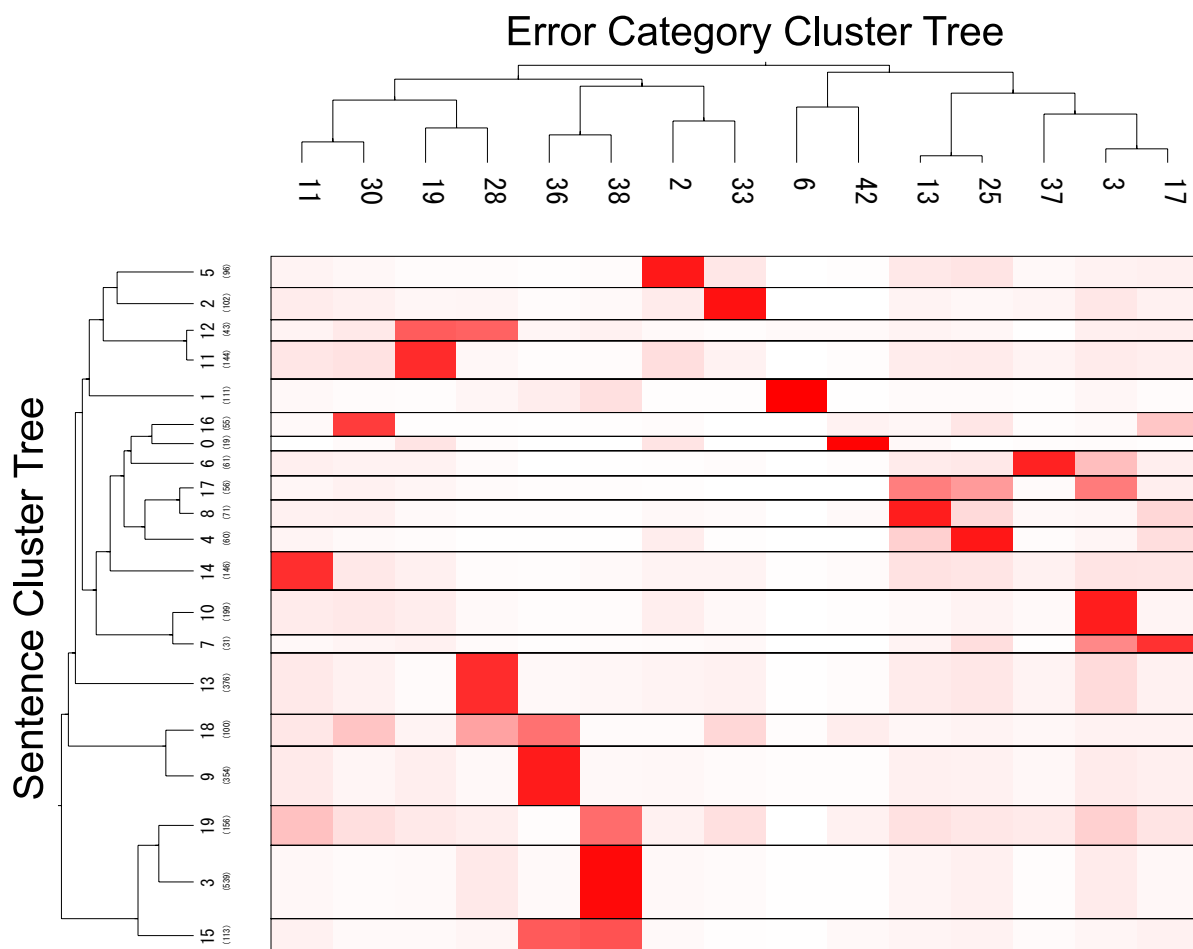


Figure 4.8: Clustering of writing errors by Spanish natives.

As with all the preceding results, sentences written by Taiwanese natives, as seen in Figure 4.7, have error category 38 (article) co-occurring with 6 (Dangling/misplaced modifier), 28 (Lexical/phrase choice), and 36 (preposition), and also error category 28 (Lexical/phrase choice) with 36 (preposition). In addition it can be seen that error categories 3 (Verb missing), 13 (extraneous words), and 25 (ambiguous/unlocatable referent) co-occur in the same sentences.

In contrast with the previous results, sentences by Spanish natives that contain error category 38 (article) co-occur only with 36 (preposition) as displayed in Figure 4.8. Another similarity is the co-occurrence of error category 28 (Lexical/phrase choice) and 36 (preposition). There are some similarities with Taiwanese in that they share the same error co-occurrence between error categories 3 (Verb missing), 13 (extraneous words), and 25 (ambiguous/unlocatable referent). A co-occurrence that is unique to Spanish is that

Table 4.2: Characteristic co-occurring errors for each of the 5 main native languages.

Co-occurring Errors	J	K	C	T	S
3, 13, 25				○	○
3, 17					○
6, 38	○	○	○	○	
11, 36		○			
19, 28					○
19, 36	○				○
28, 38	○	○	○	○	
28, 36	○	○	○	○	
30, 42	○	○			
36, 38	○	○	○	○	○

between 19 (Verb formation) and 28 (Lexical/phrase choice), and 3 (Verb missing) and 17 (Tense).

The results of the error co-occurrence clustering analysis for the 5 main native languages are shown in 4.2. All of the native languages have 36 (Preposition) and 38 (Article) as characteristic co-occurring errors. Asian native languages also share 38 (Article) co-occurring with 6 (Dangling/misplaced modifier) and 36 (Preposition), and 28 (Lexical/phrase choice) co-occurring with 36 (Preposition). Chinese and Taiwanese have some similarities with the only difference being that Taiwanese natives have 3 (Verb missing), 13 (Extraneous words), and 25 (Ambiguous/unlocatable referent) as characteristic co-occurring errors. Korean contains the same co-occurring errors as Japanese, except instead of error categories 19 (Verb formation) and 36 (Preposition) co-occurring, Korean has a more prominent co-occurrence between error category 11 (Word order) and 36 (Preposition).

In contrast with the results from other native languages, sentences by Spanish natives contain 19 (Verb formation) and 28 (Lexical/phrase choice), and 3 (Verb missing) and 17 (Tense) as characteristic error categories. There is also a similarity with Taiwanese in that they share the same error co-occurrence between error categories 3 (Verb missing), 13 (Extraneous words), and 25 (Ambiguous/unlocatable referent).

4.2.3 Analysis by Tree Distance

This analysis aims to identify differences in error categories between native languages. Initial analysis was conducted for similar and different category distances, however it was

found that a large number of error categories shared similar tree distances across most of the native languages. Therefore we will focus on the difference of error category tree distances in this analysis.

$$\arg \max \{|d(C_i, C_j, N_p) - d(C_i, C_j, N_q)|\} \quad (4.1)$$

The tree distance between two leaves of the error category cluster tree was calculated to investigate the difference in error category tree distances. Equation 4.1 was used to search for the errors with the greatest difference between languages, where $N_p \neq N_q$ and $C_i \neq C_j$, and $d(C_i, C_j, N_p)$ is the distance between the nodes of error category C_i and C_j for the native language N_p .

The results of this analysis are shown in Table 4.3 as a distance matrix. It can be seen that there are no discernable differences in the distances of error category nodes between Japanese and Korean. This would suggest that the errors of Japanese and Korean learners have similar characteristics. The difference between Taiwanese and Spanish is also low, only differing in the distance between errors 17 (tense) and 19 (verb formation). These two similar groups seem to be at extremes as they have the greatest number of differences. Chinese has two distances that are different when compared to all the other languages. The differences between Chinese and the Taiwanese/Spanish group are error categories 19 (verb formation) and 36 (preposition), along with 36 (preposition) and 38 (article). The Japanese/Korean group has different distances to Chinese in error categories 3 (verb missing) and 37 (genitive), and 17 (tense) and 37 (genitive).

4.3 Conclusion

The present chapter clustered the predicted scores of the writing error categories of 142,465 sentences by English learners' writing from the language learning SNS web site Lang-8. To investigate the differences in error characteristics of native languages, the data was divided into subsets based on the native language of the learner. These subsets were then each clustered based on the predicted error category score vectors. The clustering results of five major subsets: Japanese, Korean, Chinese, Taiwanese and Spanish were then analyzed and compared to determine the error characteristics of the native languages. All of the native languages have two co-occurring errors in common (28 and 36, 36 and 38). Asian languages have the co-occurring errors 6 and 38, and 28 and 38 in common. Japanese

Table 4.3: Error category difference of native languages based on tree distance.

	J	K	C	T	S
J		NA	3,37; 17,37;	3,37; 17,19; 19,36;	19,28; 19,36; 28,38; 36,38;
K	NA		3,37; 17,37;	3,37; 17,19; 19,36;	19,28; 19,36; 28,38; 36,38;
C	3,37; 17,37;	3,37; 17,37;		19,36; 36,38;	19,28; 36,38;
T	3,37; 17,19; 19,36;	3,37; 17,19; 19,36;	19,36; 36,38;		17,19;
S	19,28; 19,36; 28,38; 36,38;	19,28; 19,36; 28,38; 36,38;	19,28; 36,38;	17,19;	

and Korean also have co-occurring errors 30 and 42 in common. Taiwanese and Spanish have three error categories 3, 13, and 25 that co-occur. The error categories of each of the results was clustered and analyzed by the tree distance between nodes. There were no observed differences between Japanese and Korean. For Taiwanese and Spanish, only the distance between errors 17 and 19 is different, suggesting a degree of similarity. Chinese has two different error pairs difference when compared with all other languages. These results suggest that the error characteristics of Japanese and Korean learners are quite similar, as are those of Taiwanese and Spanish learners to a lesser degree. These differences in co-occurring errors are characteristic of the learner's native language. This could be used in teaching and learning to focus on co-occurring errors that are characteristic of their native language.

Chapter 5

Classification of Native Language by Writing Error Analysis and Characteristic Feature Extraction

Native language identification (NLI) is a process of determining the native language of a foreign language learner by analyzing a piece of their writing. Fundamentally, this problem can be thought of as the process of identifying characteristic features that represent the application of a learner's native language knowledge in the use of the language that they are learning. Previous research has shown that learners from different native language backgrounds have different characteristics in their use of foreign language [60]. Recently, research into the automation of NLI has been gaining in popularity and there are several practical applications to which the process could be applied, such as: providing targeted feedback on detected and potential errors in learner writing based on known problems for native language groups, and forensic linguistic author profiling where the native language of the author can be an important feature for investigation [64].

In this chapter, we approach the problem of identifying characteristic differences and the classification of learner native languages from the perspective of writing errors. The basis for this is that learner writing can contain words, in particular nouns, that have a strong relationship with the learner's native language. While these words can be a good indicator of the learner's native language, the use is highly dependent on the subject or theme of the writing and less to do with the language learning process. An example this might be the differences in the nouns used by a learner writing a personal diary versus those used in an essay on a subject that requires specialist nouns, such as computer science. Analysis on learner writing errors is less dependent on the subject of the writing as the

target of analysis is based on writing error concepts rather than the actual words of the learners' writing.

A set of 15 predicted writing error scores made from the normalized output of 15 different support vector machine (SVM) classifiers trained in chapter 3 are used as the basis of this analysis. We refer to these predicted writing error scores as a 15 dimension error prediction vector. Preliminary investigation by clustering will be used to show the differences of co-occurring writing errors between native language groups. The error prediction vector will then be analyzed by SVM machine learning to classify a learner's native language. As a naïve baseline for comparison we will classify the native language using all words to compare the effectiveness of the proposed method. In the final section of this chapter, we will examine the influence of words that have strong cultural or nationalistic relations, such as nouns representing: people, places, food, religion, etc. A method of removing words that are characteristic to a native language will be proposed. This method will then be applied to filter out cultural or nationalistic words from the corpus to provide an alternative "non-biased" baseline for critical evaluation of the proposed error prediction vector method.

5.1 Related Work

5.1.1 Native Language Identification

As explained in the introduction, the NLI process aims to ascertain the native language of a language learner by analyzing their writing in a foreign language. Most of the research to date has focused on the prediction of the native language of a English learners. This task was examined in early work by Tomokiyo and Jones [66] that investigated the use of native/non-native detection based on text features. The aim was to classify speakers based on the content of their utterances and then switch acoustic models to improve the accuracy of automatic speech recognition. A Naïve Bayes classifier was used to analyze the words and parts of speech in the utterance transcripts in 2 and 3 native language class classification tasks. Jarvis et al. [32], used discriminant analysis classifiers to predict the native language of 500 English learners from different backgrounds. Jarvis and Crossley [33] published the first book on NLI which investigated various features and methods for identifying a learner's native language using text classification techniques. It also covers the automated analysis of the language transfer hypothesis, which is a linguistic theory

about the interaction of languages in the learners mind. Brooke et al. Brooke [5] suggested that the International Corpus of Learner English (ICLE) corpus, which is commonly used in NLI research, has problems that can lead to misleading performance evaluation due to the differences in topic bias across the corpus. It was argued that the problem stems from the way the corpus was built, and proposed other methods and sources to collect data that might be useful in the task of native language prediction. An evaluation was undertaken on data collected from a language learning SNS, Lang-8.com, and it was shown to be useful for the task. In this chapter, we analyze data collected from Lang-8.com for the purpose of native language prediction by writing error prediction vector. In 2013, Tetreault et al. [64] organized the first shared task on native language identification. A new corpus called TOEFL11 was proposed for the task, which contains essays in English by learners from 11 different native languages and was provided as the shared data set on which the participants conducted analysis. Jarvis et al. [34] was a participating group with a high identification performance. A variety of features were analyzed in the identification task, such as: word n-grams, parts-of-speech n-grams, character n-grams, and lemma n-grams. An SVM classifier was trained and the prediction performance was evaluated of several different models with varying combinations of features. Bykh and Meurers [6] examined an ensemble method based on non-lexicalized and lexicalized CFG (Context Free Grammar) production rules (CFGR) features, which outperformed all submissions to the previous shared task in 2013. Ionescu et al [28], further improved classification performance on the same task by analyzing character n-grams using string kernels for Kernel Ridge Regression and Kernel Discriminant Analysis.

5.1.2 Native Language Prediction by Error Analysis

Koppel et al. [40], investigated predicting a learner’s native language by analyzing writing errors detected with MS Word and a Brill based parts-of-speech tagger in addition to other features, such as: function words, letter n-grams, and rare part-of-speech bigrams. They analyzed a sub-corpus of ICLE containing learner writings by learners with native languages from: Russia, Czech Republic, Bulgaria, France and Spain. It was found that most classification errors occurred between writings from Slavic languages. An overall accuracy of 80% was achieved using all features.

Kochmar [39], predicted the native languages of Indo-European learners through binary classification tasks performed with linear kernel SVM. Compare to previous studies a

larger set of learner native languages were examined. These native languages were divided into two main groups: Germanic and Romance, with intergroup prediction performance accuracy ranging from 68.4% to 100%. The features analyzed for prediction ranged from general words and n-grams, to different error types that had been manually tagged within the corpus.

Bestgen et al [3], investigated the used of error patterns in NLI. The occurrence of 46 error categories that had been manually tagged within the ICLE corpus were used to predict the native language of 223 learner writings. Three groups of native languages were chosen: French, German, and Spanish. They identified that using just errors as a predictor of native language, an accuracy of 65% could be achieved. Discriminative error types for the three native languages were identified by comparing the mean relative frequency significance difference of each error category. The impact of proficiency on the results was also examined and lead to an improvement in predictive discrimination between French and German learners. In concluding it is mentioned that it still remains to be seen if the same prediction performance can be achieved through the automatic detection of writing errors, instead of relying on manual classification by hand. In this chapter, we investigate the prediction performance of automatic error detection as a predictor of the native language of learners. This is then compared to the prediction performance of an SVM classifier trained with basic word features.

5.2 Data Collection

The data analyzed in this chapter is based on a corpus of foreign language writing that was collected from the language learning SNS (social networking service) Lang-8¹ during the period from October 9 2011 to January 6 2012. The data consists of journals that have been written by an English language learner and then corrected by an English native speaker at the sentence level. All samples that were collected were tagged with the native language of the learner based on there personal profile on the website. The raw data consists of a pair of sentences: the original sentence that the learner has written, and a version of the same sentence that has been corrected by an English native speaker. The corrected sentence also contained markup identifying the edits that had been made. However the markup is input by users without rules or constrains. Because of these

¹<http://www.lang-8.com>

Table 5.1: Error Category Numbers and Descriptions.

Category	Description
2	Subject formation
3	Verb missing
6	Dangling/misplaced modifier
11	Word order
13	Extraneous words
17	Tense
19	Verb formation
25	Ambiguous/unlocatable referent
28	Lexical/phrase choice
30	Word form
33	Singular for plural
36	Preposition
37	Genitive
38	Article
42	Spelling

inconsistencies the markup was cleaned from the data and a simple alignment algorithm was used to extract the positions of insertions and deletions made during the editing [43]. These features were tagged with the prefix “i:”, “d:”, and “e” for insertion, deletion, and general edit (both insertion and deletion) respectively.

5.2.1 Writing Error Tagged Corpus

In previous work [10], a sample of 399 sentences were randomly selected from the corpus and each original/corrected sentence pair was manually tagged into 15 error categories which are shown in 5.1. It should be noted that the set of error categories is a subset of 42 error categories that had been originally defined by Kroll [41] and Weltig [71] in their previous research into the performance of student writing in academic settings. The 15 error categories were chosen because of the availability of sentences containing examples of those error types. The error tagging process involved examining the types of writing errors that were corrected by the English native speaker on Lang-8. Sentences that contained multiple languages, comments, or different character encodings were removed from the corpus to reduce possible noise that might affect the training of machine classifiers.

Table 5.2: Distribution of sentences grouped by native language.

Native Language	Number of Sentences
Japanese	100,432
Chinese	14,526
Korean	13,402
Taiwanese	3,298
Spanish	2,912

5.2.2 Native Language Tagged Corpus

A corpus of 142,465 sentences were selected that had been corrected by native English speakers on lang-8. The native and target languages of the learner were also collected and each corrected sentence was annotated with this information accordingly. 5.2 shows the sentence distribution of the five main learner native languages who’s English journals were corrected by a native English speaker. A large majority are Japanese natives who have written 100,432 corrected sentences. The other main learner native languages are in descending order: Chinese, Korean, Taiwanese, and Spanish, with each language being greater than 2% of the total corpus sentence count.

In this chapter, we predict the errors of sentences by SVM models that were trained and evaluated using 10-fold cross validation. As a result of this evaluation there are 10 models for each writing error type. The prediction for each error type is made up of the average of the 10 scores from the models. The predictions are then combined to form an error vector representation for each sentence as seen in 5.1. 15 scores corresponding to 15 error categories form a vector representation of a sentence. The distribution of predicted errors for each of the five main learner native languages is displayed in 5.2.

5.3 Error Vector Analysis to Prediction of Learn Native Language

5.3.1 Biased Words

Initially an SVM model was trained to predict the native language of learners just by analyzing the words in their writings, however the prediction performance was unexpectedly high, so we investigated the characteristic feature words for each language. An SVM model was trained for each learner native language by analyzing all of the data. These models were analyzed to calculate and rank all of the feature words by weight.

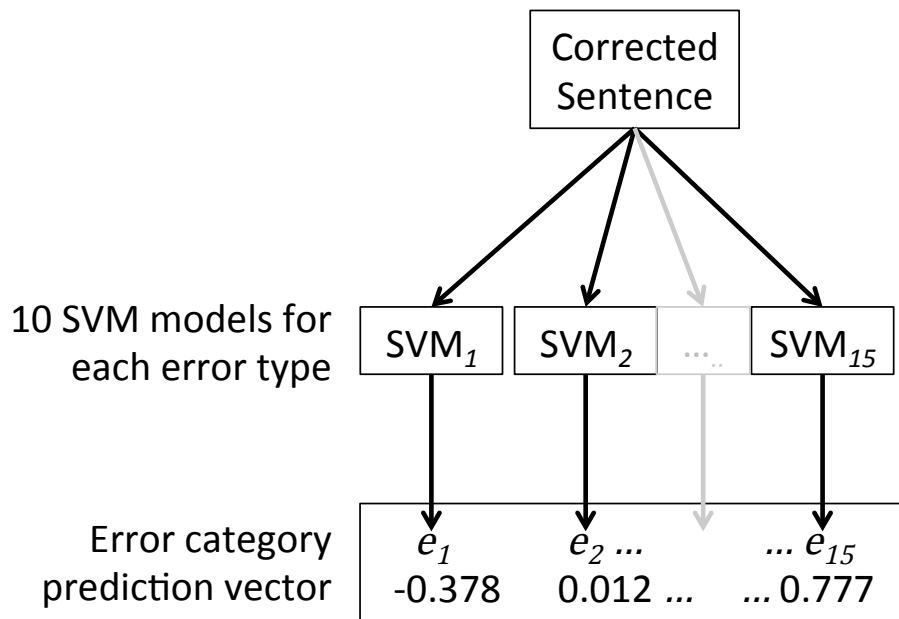


Figure 5.1: The process of creating error vector representations of each sentence.

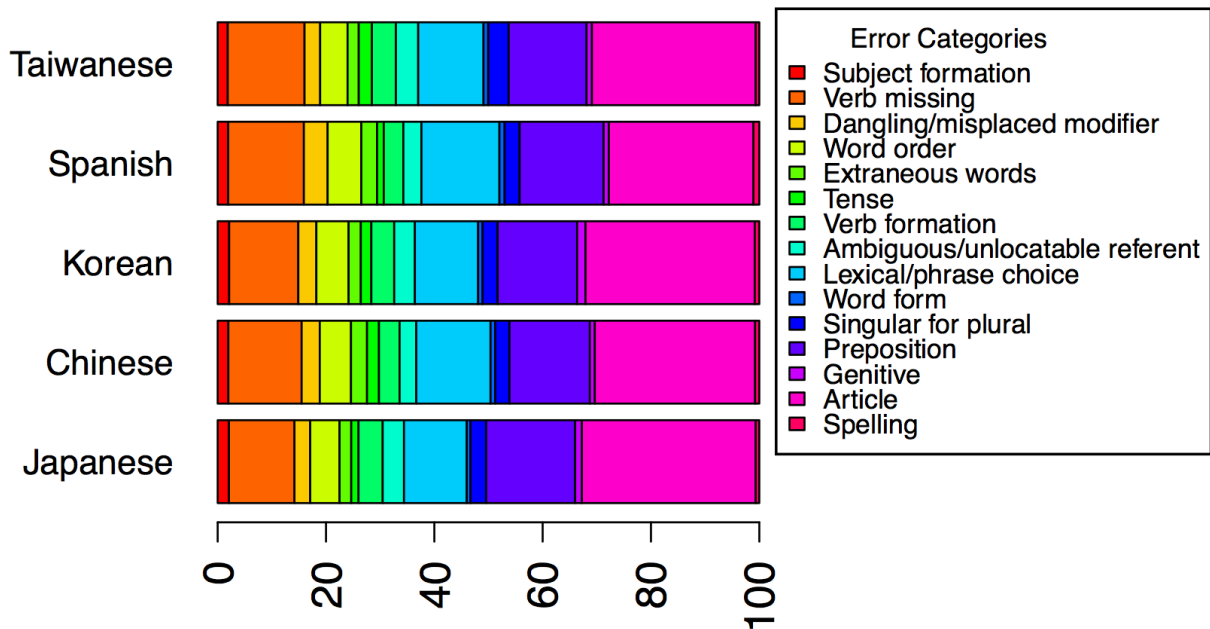


Figure 5.2: Distribution of predicted errors for each language

Table 5.3: Top 10 positive and negative feature words by weight for native Japanese learners of English.

Top Positive Words	Weight	Top Negative Words	Weight
north	1.0305	taiwan	-1.2025
japan	1.0073	campus	-1.2510
tokyo	0.6735	soju	-1.2600
japanese	0.5720	beijing	-1.3393
peninsula	0.5502	pepero	-1.3534
jong	0.5223	korean	-1.5220
kara	0.5032	kimchi	-1.5315
kyoto	0.4653	l	-1.7565
thailand	0.4447	korea	-1.7737
algerian	0.4447	seoul	-1.8214

Table 5.4: Biased words in the model for Japanese native language learners.

Rank	Weight	word
13	0.3602	earthquake
...
24	0.3093	radiation
...
42	0.2943	nuclear

Feature words with a high positive weight are characteristic of that particular learner group. In Table 5.3, the top 10 positive and negative weight feature words for native Japanese learners of English are shown. Many high positive words are directly related to Japan, were as low negative words are related to other countries. Therefore, these words have a strong bias that is influenced by the nation or culture of the learner. The characteristic feature words for each learner native language group also contained similar influences.

Other sources of biased words included events that had occurred just before the collection of data from the lang-8.com website (October 2011 – January 2012). Table 5.4 contains feature words that we believe are related to the 2011 Tohoku Earthquake and Tsunami that occurred in Japan.

To reduce the influence of trivial biased words and provide a fair comparison between the proposed method of language prediction by error vector and the baseline method of prediction by words, feature words with a high frequency distribution difference between the native language groups were removed. The relative standard deviation for each word

was calculated as follows:

$$TDR(w, l) = \frac{TF(w, l)}{DF(l)} \quad (5.1)$$

$$\mu(w) = \frac{\sum_{l \in L} TDR(w, l)}{|L|} \quad (5.2)$$

$$\sigma(w) = \sqrt{\frac{\sum_{l \in L} TDR(w, l)^2}{|L|} - \mu(w)^2} \quad (5.3)$$

$$RSD(w) = \frac{\sigma(w)}{\mu(w)} \quad (5.4)$$

Where Equation 5.1 is the term document ratio for the word w in language set l , and $TF(w, l)$ is the term frequency and $DF(l)$ with the document frequency. The standard deviation and mean of the term documents ratio between languages is calculated in Equation 5.3 and 5.2 respectively. Then finally the relative standard deviation is shown in Equation 5.4.

A list of words ranked by RSD was manually checked for words that might identify the culture or nation of the five main groups of native languages. Through these manual checks it was estimated that words with an RSD of greater than 1.25 were trivially biased towards one or more of the native languages. 5.3 shows a plot of all words ranked by RSD in descending order, with the horizontal line at 1.25 RSD representing the maximum threshold for non-biased words used in the analysis of this chapter.

5.3.2 Method and Results

To provide a fair evaluation of the two feature sets, the same method was used for training and evaluating prediction performance of error prediction vectors and word vector features. For additional comparison, we also include the classification performance for word vectors that contain all the words of the original learner writing, including those that were identified as trivially biased in the previous section. For the word vectors, the words of each sentence were vectorized as a bag-of-words model. The error prediction vector consists of the values of 15 error prediction scores.

Separate SVM classifiers were trained for five different native languages across all three data sets. The native language classification performance of each of these classifiers was evaluated by randomly sampled 10-fold cross validation, with 9:1 training to test data ratio for each of the data sets.

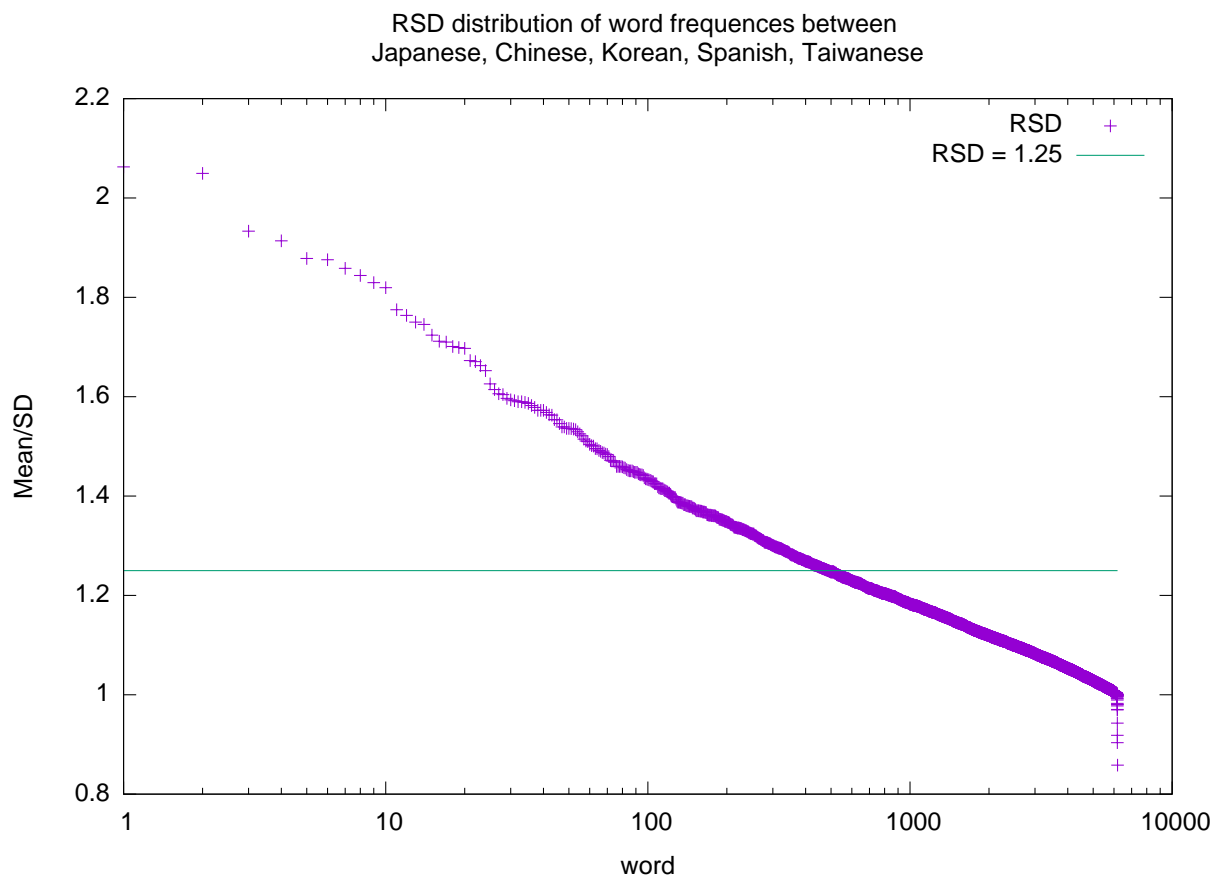


Figure 5.3: The RSD distribution of word frequencies between five native languages.

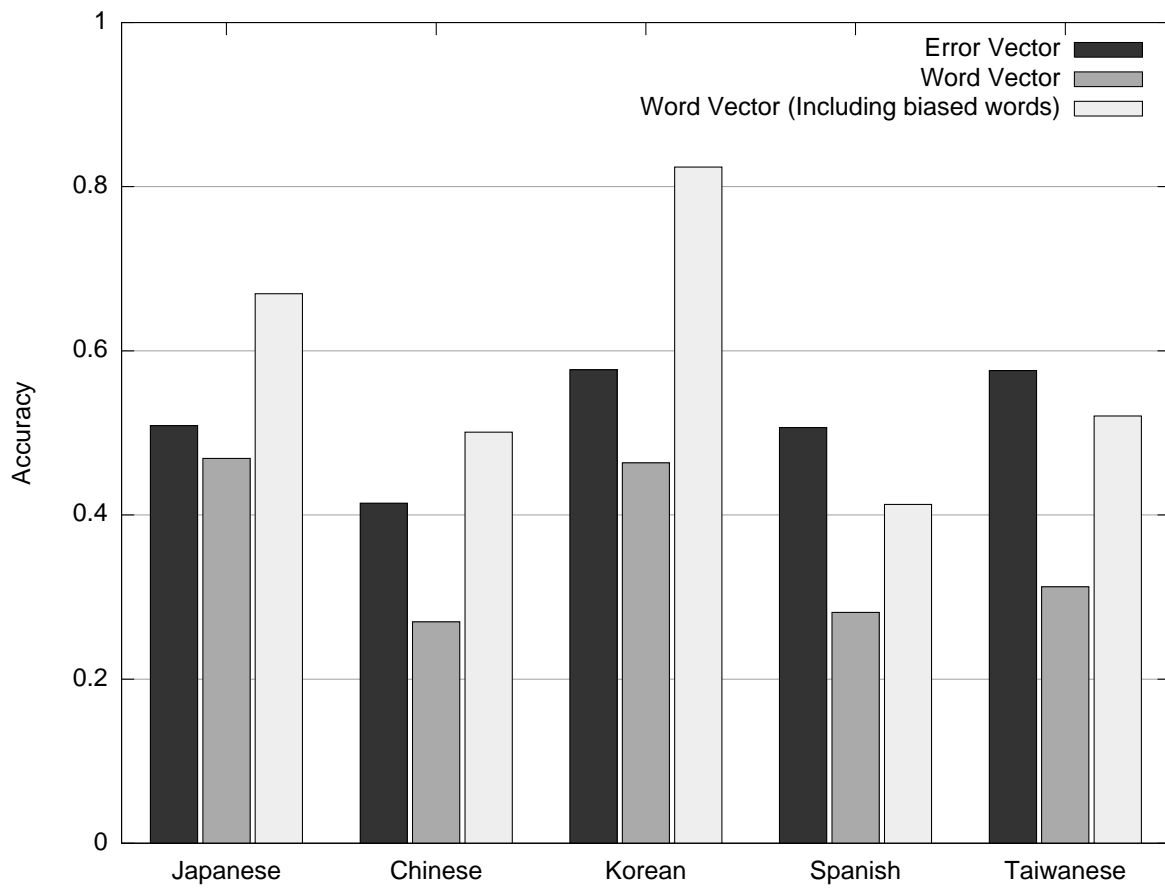


Figure 5.4: Native language prediction evaluation for each vector (Accuracy, 10-fold cross validation).

A comparison of the classification performance evaluation on all three data sets for each of the five native languages is shown in 5.4. The classification performance of the word vectors that include biased words is high, especially for writings by native Korean learners. This would suggest that there are biased words that are highly characteristic of native Korean learners. The word vectors that do not contain biased words have a classification performance ranging from 36% lower in the case of Korean, to 13% lower for Spanish. The native language classification performance by error prediction vector is higher than the performance of the unbiased word vector. However the classification performance for two out of the five native languages is lower than that of the word vectors that contain all the words of the original learner writings, which we argue is influenced by biased words.

5.4 Discussion

In this chapter, we approach the problem of identifying characteristic differences and the classification of learner native languages from the perspective of writing errors. A set of 15 predicted writing error scores were used as the basis of this analysis. The error prediction vector made up of the 15 predicted writing error scores was analyzed by SVM machine learning to classify a learner’s native language. The effectiveness of this method was compared to two baseline evaluations: a naïve baseline using all words, and a baseline trained only on words that do not have a strong bias. In the final section of this chapter, we proposed a technique used to filter words that have strong cultural or nationalistic relations based on the relative standard deviation of occurrence across native languages in the corpus. The classification performance for models trained on error prediction vectors were superior to the unbiased word vectors for all native languages. However, models trained on all words including biased words performed better in three out of five native languages.

Chapter 6

Automatic Extraction and Prediction of Word Order Errors From Language Learning SNS

In recent years, research into writing tools to support foreign language learners of English has been growing. However, most research to date has focused on the prediction/correction of prevalent errors in learner writing, such as: preposition and article errors[65]. While the prediction/correction of these errors would have a great impact on learner writing, other less prevalent errors, such as word order errors, have received little attention. Word order differs significantly across languages[49], which poses a particular problem for learners from L1 languages that have a fairly different word order to the L2 language. In this chapter, we examine the prediction of word order errors in foreign language writing of learners from a Japanese L1 background learning English. In chapter 3, we examined automated error prediction of 15 different error categories in learner writing on the language learning SNS, Lang-8.com. However the samples of manually tagged sentences available for some error categories was minimal, such as word order errors, and was problematic when training error models and resulted in low prediction performance. To overcome this problem, we propose that a large amount of word order error samples can be automatically extracted from a corpus of corrected learner writing by comparing the edit distance between original and corrected sentences. We then train and evaluate the prediction performance of a Support Vector Machine (SVM) classifier by analyzing a corpus constructed using the proposed method.

The method of analyzing the edit distance between original and corrected learner writing sentences has been examined in previous work to automatically identify errors[67]

and extract L2 criterial lexicogrammatical features from learner corpora[68]. We extend the use of this method to data that has been collected from an language learning SNS to automatically predict word order errors by machine learning.

6.1 Automatic Word Order Error Sample Extraction by Edit Distance

In this section, we will introduce a method of automatically extracting sentences written by foreign language learners that contain word order errors from a corrected language learning writing corpus. An edit distance of the difference between the original and corrected sentence can be analyzed to identify the corrections that have been made. In particular, we analyzed the Levenshtein distance[43] to find insertions and deletions in corrected sentence pairs.

A word order error can be thought of as a sentence pair that contains the same frequency of insertions and deletions identified by the edit distance for each corrected word. Conversely, a sentence pair that only contains either insertions or deletions for each corrected word can be thought as not containing a word order error. In Equation 6.1, we define the conditions used to select a set of sentence pairs that contains word order errors.

$$WO(S) = \{s_i | w_j \in s_i; ins(w_j) = del(w_j), ins(w_j) > 0\} \quad (6.1)$$

Where S is the set of all sentence pairs, w_j is the j^{th} word in sentence s_i , and $ins(w_j)$ and $del(w_j)$ are the number of insertions and deletions of the word w_j identified in sentence s_i by the edit distance. Equation 6.2 defines the conditions to select a set of sentence pairs that does not contain word order errors.

$$NotWO(S) = \{s_i | w_j \in s_i; ins(w_j) \oplus del(w_j)\} \quad (6.2)$$

6.2 Data Collection

In this section we will analyze the raw data from the Lang-8 Learner Corpora[48] to extract word order errors by the edit distance method described in the previous section. The corpus contains both the original sentences written by learners and sentences corrected by other users of Lang-8 that are proficient in the target language. The learners' L1 and L2 are tagged for each document made up of a number of sentences. Firstly, we extracted

Table 6.1: Number of corrected sentence samples extracted.

Error Type	# Sentences Pairs
Word order error only	7043
Other error only	742064
Word order and other error	122325

sentences from the corpus that were written by Japanese L1 learners learning English that had been corrected one or more times. After removing comments and styling tags from the corrections, we then filtered to remove invalid corrections containing multiple languages which resulted in 871,432 original/corrected sentence pairs. The edit distance between the original and corrected sentence was then calculated for each of the sentence pairs. This was then analyzed to extract sentence pairs that contain word order errors, sentences that do not contain word order errors, and sentence pairs that contain a combination of errors, and therefore do not fall into either of the defined sets. The size of the extracted sets is shown in Table 6.1.

As shown in Figure 6.1, a majority of corrected word order error sentence samples have been corrected by only moving the position of one word in the sentence. The number of samples by edit distance decreases at an almost exponential rate as the edit distance length increases along the x axis of the plot.

We created a corpus for machine learning by selecting all of the sentences in the word order error set as the positive class, and then selected at random using the GNU shuf utility¹ an equal amount of sentences (7043 sentence pairs) from the other error only set as the negative class. All of the original and corrected sentences were then processed using TreeTagger[56] for Parts of Speech (POS) tagging. Words in the corrected sentence that were identified by the edit distance analysis to be either an insertion or deletion were included as both untagged and tagged words as follows: insertions were prefix tagged with “i:”, deletions with “d:”, and all edited words were prefix tagged with “e:”. N-grams of lengths ranging from 2 to 4 Words/POS tags were also used for analysis and will be referred to as n-gram features. This corpus contains features from both the original and corrected sentences and we will refer to it as the *Parallel corpus*. An additional corpus containing features only from the original learner written sentences, that we will refer to as *Single corpus*, was created for the prediction of word order errors in non-corrected

¹<http://www.gnu.org/software/coreutils/shuf>

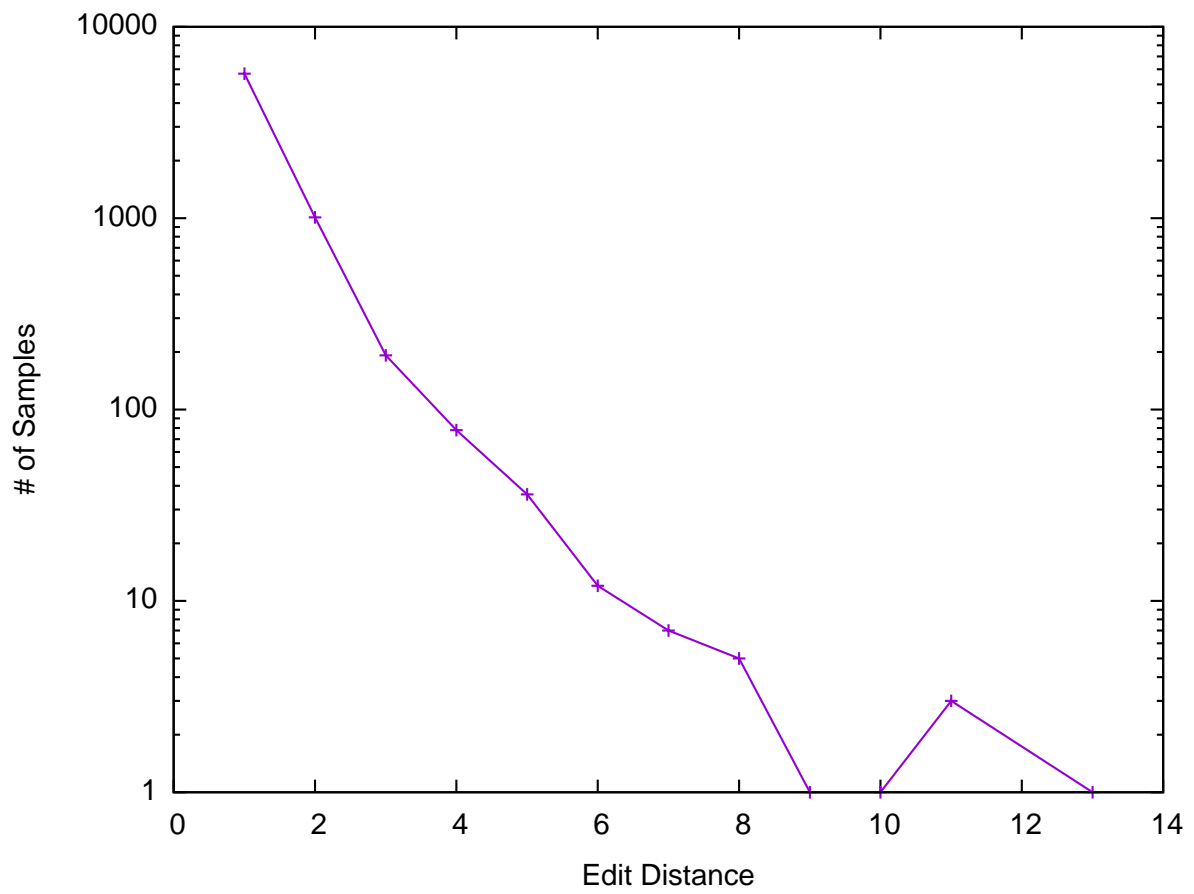


Figure 6.1: Distribution of the edit distance of corrected word order error sentence samples.

Table 6.2: *Parallel* corpus baseline prediction performance.

Features	F	Accuracy
Word	0.8745	0.8777
Word, N-gram	0.6184	0.7178
Word, POS	0.9037	0.9043
Word, N-gram, POS	0.3305	0.5981

learner writing.

6.3 Word Order Error Prediction by SVM and Feature Selection

The *Parallel* and *Single* corpora were indexed using GETAssoc² to create a search engine for the retrieval of features and vectorization of sentence data.

6.3.1 Method

The SVM^{light}[36] linear kernel classifier was used for model training and evaluation. Initially an SVM model was trained on all of the corpus data only for the purpose of feature scoring. The feature score was extracted by analyzing the weights of features in the SVM model trained on all the data. The corpora were then split into train and test sets at a ratio of 9:1 for evaluation by 10-fold cross validation. The prediction performance of an SVM model trained on all of the features was evaluated as a baseline. Feature selection was then performed by selecting increasingly larger sets of N top positive and N top negative score features and evaluating the prediction performance of each set. The set with the best prediction performance is therefore the optimal feature selection.

6.3.2 Baseline Prediction Performance Evaluation

An SVM model trained on all features was evaluated as the baseline of prediction performance. The baseline prediction performance results are shown in Table 6.2 for SVM models trained by analyzing all of the features in sub-feature set of the *Parallel* corpus. The best performing SVM model by Accuracy and F-measure was trained and tested on word and POS tag features of the *Parallel* corpus. The prediction performance is high,

²<http://getassoc.cs.nii.ac.jp/>

Table 6.3: *Single* corpus baseline prediction performance.

Features	F	Accuracy
Word	0.6750	0.5979
Word, N-gram	0.6813	0.5997
Word, POS	0.6839	0.6074
Word, N-gram, POS	0.6942	0.6207

Table 6.4: Optimal feature selection prediction performance for the *Parallel* corpus.

Features	N	F	Accuracy
Word	1000	0.9234	0.9250
Word, N-gram	40000	0.8905	0.8950
Word, POS	20000	0.9049	0.9056
Word, N-gram, POS	100000	0.8553	0.8705

however this is to be expected as the corpus contains features from both the original and corrected sentences along with tags indicating edits in the corrected sentence. The baseline prediction performance results for SVM models trained and tested on features from the *Single* corpus are worse as only the original learner writing features are analyzed, and lacks any information on corrections made. The baseline prediction performance results are shown in Table 6.3, with word, N-gram, and POS tags producing the best prediction performance.

6.3.3 The Effect of Feature Selection on Prediction Performance

In this section, we will examine the effectiveness of feature selection on the prediction performance of SVM models on different sub-feature sets of the corpora.

The evaluation of the optimal feature selection prediction performance on the *Parallel* corpus is shown in Table 6.4. Interestingly the top performing sub-feature set was that made up of words. The optimal N shows that a feature set of 1000 top positive and negative word features produces optimal prediction performance. Feature selection did not have much of an effect on the best performing baseline feature set of words and POS tags.

The optimal prediction performance for the *Single* corpus is shown in Table 6.5. As with the baseline prediction performance, the best prediction performance was by the sub-feature set made up of word, n-gram, and POS tag features. Optimal feature selection

Table 6.5: Optimal feature selection prediction performance for the *Single* corpus.

Features	N	F	Accuracy
Word	800	0.7115	0.6414
Word, N-gram	4000	0.7494	0.7107
Word, POS	700	0.7116	0.6625
Word, N-gram, POS	8000	0.7509	0.7154

Table 6.6: Single corpus top 10 positive and negative features.

Score	Feature	Score	Feature
0.5827	rb	-0.2664	2:jj_nn
0.3514	only	-0.2130	rb_so
0.2950	2:jj_pp\$	-0.2129	2:vb_dt
0.2753	2:nn_rb	-0.1950	rb_very
0.2726	more	-0.1689	very
0.2682	3:jj_pp\$_nn	-0.1558	2:vbp_nn
0.2668	wrb	-0.1503	3:nn_vbz_nn
0.2666	2:jj_dt	-0.1485	never
0.2546	in_up	-0.1485	rb_never
0.2525	2:i_and	-0.1390	any

was achieved at an N of 8000 top positive and negative features, resulting in a gain of 0.0947 by Accuracy.

A plot of the effect of feature selection relative to the baseline prediction performance is shown in Figure 6.2. The performance of the SVM model is greater than the baseline at $N = 200$ when measured by Accuracy. As N increases and more top positive and negative features are used for training the prediction performance by Accuracy increases until it reaches optimal feature selection at $N = 8000$. After this point, overfitting reduces the prediction performance of the SVM model as more features are added for training.

The top 10 positive and negative scoring features from the SVM model for the *Single* corpus is shown in Table 6.6. Positive scoring features are indicative of word order errors. The top scoring feature “rb” is the adverb POS tag which suggests that sentences that contain word order errors would also contain an adverb, but it doesn’t explain how the adverb interacts with other features in the same sentence. The POS tag bi-gram in the positive column “2:jj_pp\$” and “2:nn_rb” are combinations of adjectives with possessive pronoun and noun, singular or mass with adverb respectively. This suggests that word order errors usually occur in sentences written by Japanese learning English where an

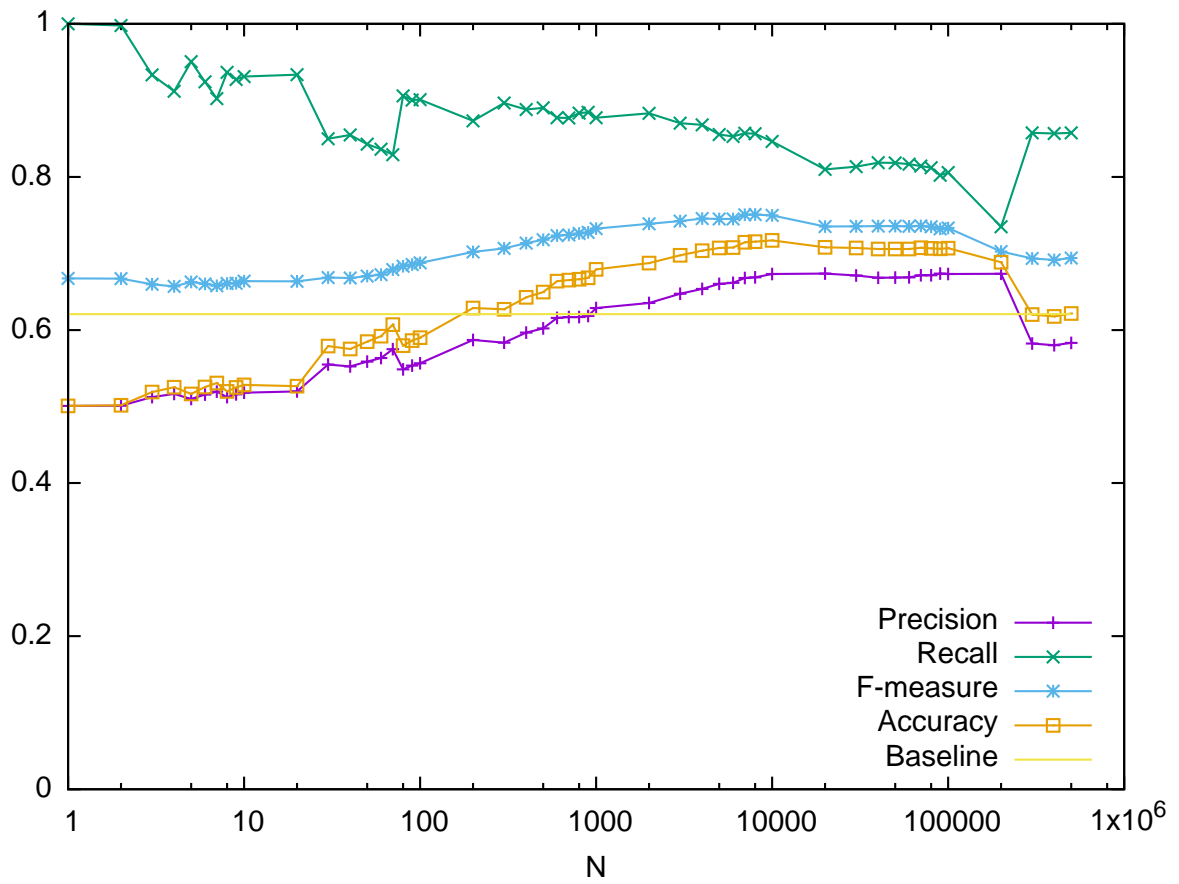


Figure 6.2: The effect of optimal feature selection on prediction performance for the *Single* corpus with the sub-feature set of words, n-grams, and POS tags.

Table 6.7: Trigram word order error characteristic feature examples.

Original
“Here are <jj> <i>new</i> </jj> <pp\$> <i>my</i> </pp\$> <nn> <i>temple</i> </nn> pictures.”
Corrected:
“Here are <pp\$> <i>my</i> </pp\$> <jj> <i>new</i> </jj> <nn> <i>temple</i> </nn> pictures”
Original:
“ <i>The</i> <jj> <i>next</i> </jj> <pp\$> <i>her</i> </pp\$> <nn> <i>show</i> </nn> will be held on 5th and 6th of December in Tokyo”
Corrected:
“<pp\$> <i>Her</i> </pp\$> <jj> <i>next</i> </jj> <nn> <i>show</i> </nn> will be held on <i>the</i> 5th and 6th of December in Tokyo”

adjective is followed by a possessive pronoun, or a noun (singular/mass) is followed by an adverb. The trigram “3:jj_pp\$_nn” is an extended form of the bigram described above and ends with a noun (singular/mass) which suggests that this particular combination in a sentence is characteristic of word order errors. In Table 6.7 are two example corrected sentences from the corpus that contain the trigram “3:jj_pp\$_nn” characteristic feature. POS tags have been inserted around the words to indicate where the feature occurs. Words that have been corrected are displayed in *italic* text.

In both of these examples, the order of the adjectives and possessive pronouns are inverted, with the corrected sentence placing the possessive pronoun before the adjective.

6.4 Discussion

In this chapter, we examined the use of edit distance analysis in the automatic extraction and prediction of word order errors from a Language Learning SNS. We extracted 7043 word order corrected learner writing sentence pairs from a raw corpus and combined it with 7043 randomly selected sentence pairs that do not contain word order errors to create a balanced word order error corpus for machine learning.

We then evaluated the prediction performance of an SVM model and feature selection in classifying word order errors on a *Single* and *Parallel* corpus. As expected, the results were high for the *Parallel* corpus as it contains information from the corrected sentence. The prediction performance on the *Single* corpus was improved by optimal feature selection.

Chapter 7

Classification of Speaking Proficiency Level by Machine Learning and Feature Selection

At present there are many machine readable data that are publicly available, and this has increased the application of machine learning to the task of supporting language learning. In this chapter, we analyze the NICT-JLE corpus¹ to investigate which words describe and discriminate different speaking proficiency levels by applying a method of machine learning called SVM (Support Vector Machine) to the classification task. The corpus consists of 1280 transcribed recordings of the Standard Speaking Test[29, 30, 31] (herein referred to as SST) English language learner exam. Each exam contains 3 different tasks and the transcriptions are made up of the dialogue between the examiner and examinee. The proficiency level for each examinee was determined by an expert examiner and ranked on a scale from 1 to 9, from beginner to advanced respectively. In this chapter, the focus of the classification analysis will be on the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) (Council of Europe, 2001)[9] which is utilized internationally, rather than the SST proficiency levels that are applicable only within Japan. The equivalent proficiency levels of SST, CEFR, and CEFR-J (a version of the CEFR that has been tailored to the needs of Japanese learning English) as defined by Tono et al. [69] are shown in Table 7.1. It should be noted that SST level 4 can be assigned to either CEFR level A1 and A2, and we will refer to these as CEFR1 and CEFR2 respectively. In this chapter, the evaluation of the classification method was performed with SST level 4 included in the CEFR level A2. The classification of SST level 4 included

¹http://alaginrc.nict.go.jp/nict_jle/index_E.html

Table 7.1: Equivalent levels of CEFR, CEFR-J, and SST

CEFR	-	A1			A2		B1		B2		C1	C2
CEFR-J	Pre A1	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	C1	C2
SST	1	2/3	3	4	4	5	6/7	8	9	9	9	9

in the CEFR level A1 should be investigated in future work. SST level 9 is included only in CEFR level B2.

For each of the 1280 examinee’s in the SST data there are 5 stages of the interview that have been transcribed. In this chapter, the results for each examinee were represented as one document, and there were 1280 sample documents for which the proficiency level classification problem was analyzed. Examinees who have an SST proficiency level of 1 were excluded as it would be equivalent to Pre A1 CEFR level. A total of 9,626 words were analyzed along with 11 parts of speech (POS) from Lancaster University’s CLAWS5 and CLAWS7 tag sets².

Automated language scoring using a computer was first proposed by Page in 1968 [50]. Since then research into the prediction of foreign language proficiency has focused on a number of different approaches. Supnithi et al. [59], analyzed the vocabulary, grammatical accuracy and fluency features of the NICT-JLE corpus. SVM and Maximum Entropy classifiers were trained to automatically predict the proficiency level of the learner, with SVM achieving the best prediction accuracy of 65.57%. There has also been research into extracting features that can be useful in classifying proficiency levels in the NICT-JLE corpus [1, 18]. In this chapter, analysis by SVM and feature selection is used to not only improve the accuracy of proficiency classification, but also identify optimal sets of characteristic features that can describe learners from different proficiency levels.

7.1 Proficiency Level Classification by SVM and Feature Selection

The occurrence frequency (tf) of each word was used to vectorize each of the transcripts. This was realized by creating a term document matrix of the exam transcripts using GETA³.

To evaluate the performance of classifying documents into two classes of proficiency

²<http://ucrel.lancs.ac.uk/claws5tags.html>, <http://ucrel.lancs.ac.uk/claws7tags.html>

³<http://geta.ex.nii.ac.jp>

Table 7.2: Measures used for Feature Selection.

Symbol	Measure	Symbol	Measure
<i>w.o</i>	$weight(w_i)$	<i>w.a</i>	$abs(weight(w_i))$
<i>d.o</i>	$weight(w_i) * df(w_i)$	<i>d.a</i>	$abs(weight(w_i) * df(w_i))$
<i>l.o</i>	$weight(w_i) * \log(df(w_i))$	<i>l.a</i>	$abs(weight(w_i) * \log(df(w_i)))$

levels, the documents of level X were represented as positive examples, while the documents of level Y were represented as negative examples to train a machine learning model. SVM^{perf} [35] was used to train and test models on the data of the corpus. The experiment process can be broken down into 3 main steps. All features (words, POS tags) are used to train a model in step 1. The ranking $weight(w_i)$ scores for each feature are then extracted from the model in step 2. These feature weights are then ranked in step 3 where the classification performance of models trained and evaluated using feature selections of increasingly larger sets of $N = 1, 2, \dots, 10, 20, \dots, 100$ is analyzed. The optimal feature selection is the best performing model trained on N features. The classification performance of each model was evaluated using 5-fold cross validation.

The feature $weight(w_i)$ score extracted in Step 2 represents the distance from the SVM hyperplane that separates the positive and negative classes on which the model was trained. Models were trained with the upper proficiency level learner data as the positive class, and the lower level learner data as the negative class. Features that have positive $weight(w_i)$ are characteristic of upper level learners, and a negative feature $weight(w_i)$ are characteristic of lower level learners.

7.1.1 Feature Selection Measures

The classification performance of a model trained using all features for A1 and A2 were: Precision 0.8923, Recall 0.8117, F-measure 0.8491, and Accuracy 0.7830. Although the classification performance is quite high, we do not know which grammar items are effective for discriminating between different proficiency levels. In this chapter, we apply the method from Sakai and Hirokawa [55] to the problem of feature selection to find a set of optimal discriminating features.

The feature score $weight(w_i)$ extracted in Step 2 was calculated using 6 different evaluation measures as shown in Table 7.2. $df(w)$ is the number of documents in which the word w occurs, and abs returns the absolute value of the enclosed value.

In the case of measures that do not take the absolute value of the score: the top

Table 7.3: Classification performance when using all feature words.

Acc	A2		B1		B2	
A1	all	0.8188	all	0.9675	all	0.9966
	$N = 20$	0.7607	$N = 20$	0.8099	$N = 20$	0.7763
	$N = 50$	0.7837	$N = 50$	0.8615	$N = 50$	0.9333
	$N = 100$	0.8171	$N = 100$	0.9269	$N = 100$	0.9760
A2			all	0.8673	all	0.9879
			$N = 20$	0.6292	$N = 20$	0.8774
			$N = 50$	0.7750	$N = 50$	0.9657
			$N = 100$	0.8393	$N = 100$	0.9846
B1					all	0.8512
					$N = 20$	0.4493
					$N = 50$	0.6809
					$N = 100$	0.8405

N positive *weight* features are selected along with the top N negative *weight* features for vectorization. For measures that do take the absolute value of the score the top $2N$ positive *weight* features are selected for vectorization.

7.2 Proficiency Classification Performance

This section explains the results of the proficiency classification performance by accuracy that are shown in Table 7.3, and plots of feature selection results for all measures shown in Figure 7.1. The x-axis in these plots represents $2N$ number of features selected. The results for A2 vs B1, on the right of Figure 7.1, show that as the number of selected features increases, the accuracy increases following a curved line, suggesting that as the number of features increases the accuracy will steadily get higher. In other words, it is not possible to classify these classes with few features. Conversely, in the results of A1 vs A2, on the left of Figure 7.1, the accuracy rises quickly at around $N = 10$. This indicates that a decent level of classification performance can be achieved using a small number of features.

The baseline classification performance of a model that was trained using all of the features is shown in Table 7.3. It can be seen that the classification performance of adjacent proficiency levels is low. The classification accuracy of feature selection is shown in Figure 7.1, where for the plot on the left A1 is the positive class and A2 is the negative class, and the x-axis is $2N$ top ranking features selected. When $N = 200$ or greater the

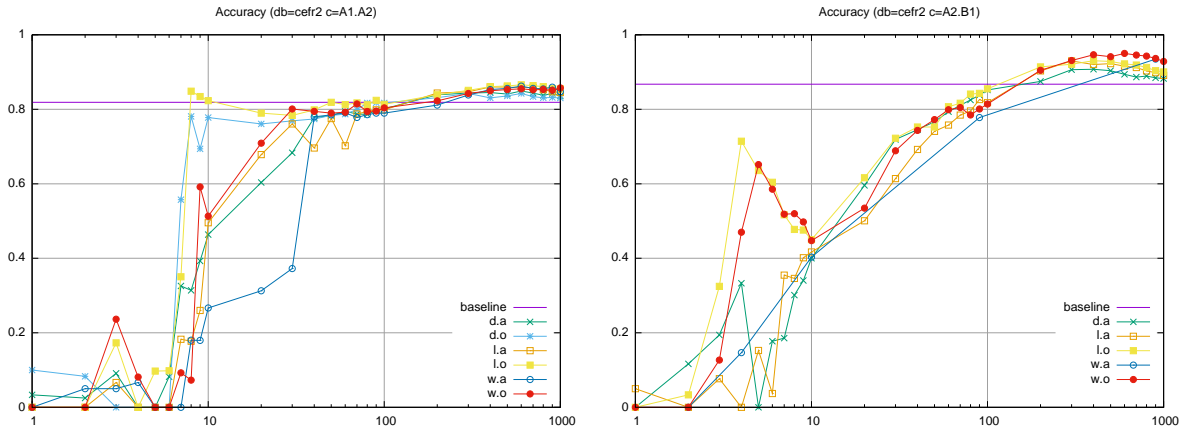


Figure 7.1: Performance accuracy of feature selection

accuracy of the model is slightly better than a model trained using all features. The two measures: *l.o* and *d.o* outperform the baseline at $N = 9$ which indicates that classification can be achieved with a small number of features.

7.3 Characteristic Features of Level A1

The top 10 characteristic features of level A1 are compared to other levels in Table 7.4. The feature "jp" represents a Japanese word that was said in the exam and has been replaced during transcription. Regardless of which level A1 is compared to, the nouns: cat, theater, boy, zoo, lion, and monkey are frequent. This is most likely effected by the contents of picture cards on which conversations are based in certain SST tasks. On the other hand, other levels have higher numbers of verbs, adverbs, and adjectives. However, more high level parts of speech features such as VERB and ADJ are not seen as characteristic features. Therefore, discrimination between levels is not possible using simple parts of speech. Even though the POS tag information was analyzed, looking at the top ranking features when comparing A2 and B1, only 3 POS tags appear as characteristic features: C7=RGQ (adverb expressing a degree) for A1, and C7=RRR (comparative adverb) and C7=DA (adjective used as pronoun) for B1. Also in Table 7.4 is can be seen that different characteristic features are chosen when comparing A1 to different levels. An unexpected result is that classification can be achieved with just 20 features.

Table 7.4: Characteristic features of A1 and other comparative levels.

Characteristic Features of Level A1	Comparative level characteristics
look, please, jp, first, work, just, picture, what, friend, cat	(A2) home, find, when now, will, ask, other eat, think, if
ten, c7=RGQ, story, speak, theater, boy our, bring, anonym., favorite	(B1) really, also, ask call, actually, different c7=RRR, your, stay c7=DA
theater, pardon, cold color, zoo, lion, monkey shinjuku, recently, tv	(B2) an, into, drive brother, anything, club fun, once, teacher explain

7.4 Discussion

In this chapter, we analyzed the transcripts of a speaking test corpus by applying SVM machine learning to the problem of classifying the differences between CEFR proficiency levels. Feature selection was used to find an optimal feature set by evaluating the model accuracy. It was found that a set of about 20 features produced the same performance as a model trained using all words and an accuracy of greater than 90%. For adjacent levels the classification accuracy was around 10% less. Classification of levels A1 vs B1 and B1 vs B2 were difficult and decent accuracy could not be achieved using small numbers of features. The characteristic features of level A1 contained numerous Japanese words, proper nouns, and simple nouns. In this chapter, when assigning the equivalent levels of SST and CEFR we made the assumption that SST level 4 was contained within CEFR level A2.

Chapter 8

Predicting and Visualizing Wine Characteristics Through Analysis of Tasting Notes From Viewpoints

Areas of specialty often require a set of expressions that are tailored to meet the need of a specific genre. As these expressions are not used commonly in everyday communication, for people that are not familiar with the specialty terminology or expressions it can be quite baffling and difficult to understand. An area of particular interest to the authors is the language that is used to describe and express complex emotions and senses. A good example of this can be seen in the description of food and beverages that consist of complex aromas, flavors, and many other characteristics as they usually are expressed using specialist terminology used in a subjective manner. Within this area, the descriptions of wine are notorious for the use of specialist terminology and the expression of commonly used words in an uncommon manner. This is formally known as winespeak, and is used by wine reviews/tasters and also in the descriptions on the back label of wine bottles. To the uninitiated, it might be difficult to understand what a wine with “slightly pungent notes of green tomato or crushed tomato leaf” might be like, as used by Joe Czerwinski in his review of a Villa Maria 2009 Sauvignon Blanc¹.

In this chapter, we propose a method for the automatic visualization of wine characteristics from viewpoints based on the sense sentiment analysis of a corpus of wine tasting notes. A subset corpus consisting of wine tasting notes that have been manually classified into four sense sentiment viewpoints will be analyzed to train and evaluate Support Vec-

¹<http://buyingguide.winemag.com/catalog/villa-maria-2009-taylors-pass-vineyard-sauvignon-blanc-marlborough>

tor Machine (SVM) classifiers for sentiment analysis. By analyzing target wine tasting notes with these classifiers, a score will be predicted from each of the sense sentiment viewpoints. These predicted scores will then be visualized in the form of Radar Charts so that the characteristics of wines may be compared.

8.1 Related Work

There are many papers on research into the language that is used to describe wines, called winespeak. Some of this research is dedicated to analyzing wine tasting notes from different points of view. Paradis and Eeg-Olofsson [51] examined tasting notes to identify expressions and words that are related to the viewpoints of vision, smell, taste, and touch. 39 typical phrases of these sensory expressions were identified. Caballero [7] focused on how manner-of-motion verbs are used from the point of view of describing a wine's intensity and persistence, and collected 56 typical sentences that contain such verbs. In this chapter, wine sentiment analysis is conducted using the four sensory viewpoints defined by Paradis and Eeg-Olofsson [51].

There is also related research into the visualization of wine tasting notes for linguistic analysis. Kerren et al. [38] visualized wine tasting notes using word trees generated from parts of speech and words. Their system enables the analysis of linguistic patterns within single wine reviews or based on regions and varieties. However the system is highly specialized and not intended for general use. In previous research, we examined the relations of Winespeak expressions and visualized these as mindmaps [27]. In this chapter, the language used in tasting notes is automatically analyzed from different sensory viewpoints. The results are then visualized as radar charts so that the sensory sentiment content of the wine tasting note can be conveyed without having an understanding of winespeak.

8.2 Data Collection

In this chapter, we propose that tasting notes can be analyzed to predict the classification of wines from various points of view. The target data for analysis is a corpus that consists of 91,010 wine tasting notes, or 255,966 sentences, that were collected from the Wine Enthusiast website². The attributes of each wine, such as: winery, region, and grape

²<http://buyingguide.winemag.com/>

Table 8.1: Top 10 positive and negative score words for the smell sense.

Positive		Negative	
Score	Word	Score	Word
1.2935	aroma	-0.4143	flavor
1.1981	note	-0.3202	juice
0.9632	nose	-0.3189	tannic
0.6752	smell	-0.2660	finish
0.5676	accent	-0.2503	chewy
0.4608	oak	-0.2359	bitter
0.4601	scent	-0.2298	card
0.4450	smoky	-0.2235	richness
0.4381	spice	-0.2218	sweet
0.4011	perfume	-0.2166	acidity

variety were collected along with the text of the wine tasting notes. This data was then indexed to construct a special use search engine using GETA³. A subset of the data consisting of 992 sentences from wine tasting notes was randomly selected for use in the training, testing and evaluation of sentiment models. This data subset was manually classified by hand into four different sensory category viewpoints, as defined by Paradis and Eeg-Olofsson [51].

8.3 Sensory Viewpoint Analysis and Prediction

An overview of the analysis in this chapter, which involves training SVM models to predict four sense sentiments for visualization as radar charts, is shown on the left in Figure 8.1. Firstly, a data subset of 992 manually classified sentences was vectorized, with each feature vector consisting of the words contained within a wine tasting note. The feature weights were normalized at the feature vector for each sentence to ensure that the number of features does not have an influence on model training. An SVM classifier for each sense was initially trained using all the data in the subset. The weights from these models was then extracted and used to score feature words for feature selection. An example of the top 10 positive and negative score feature words for the sense smell are shown in Table 8.1.

The words are ranked by the absolute value of the weight score, with the top N ranked words selected for training and testing. For each set of N top words, 5 SVM classifiers

³<http://geta.ex.nii.ac.jp/>

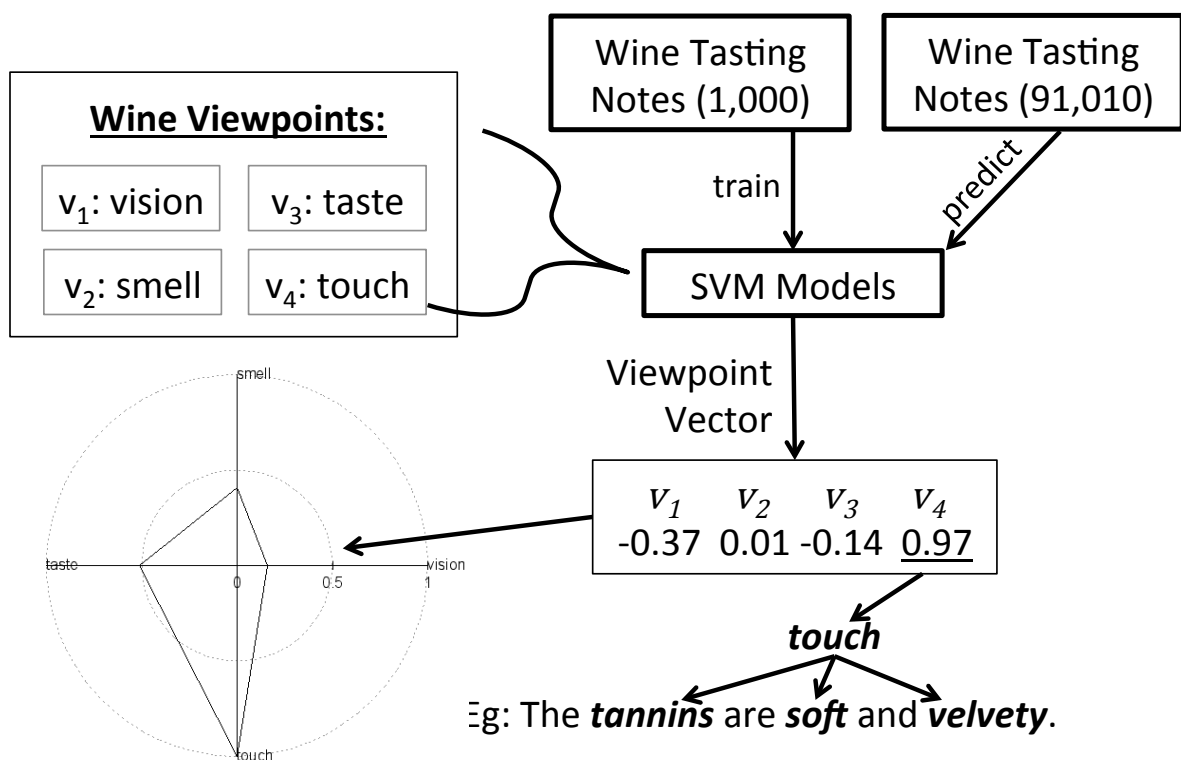


Figure 8.1: An overview of the automatic prediction and visualization of wines from multiple viewpoints.

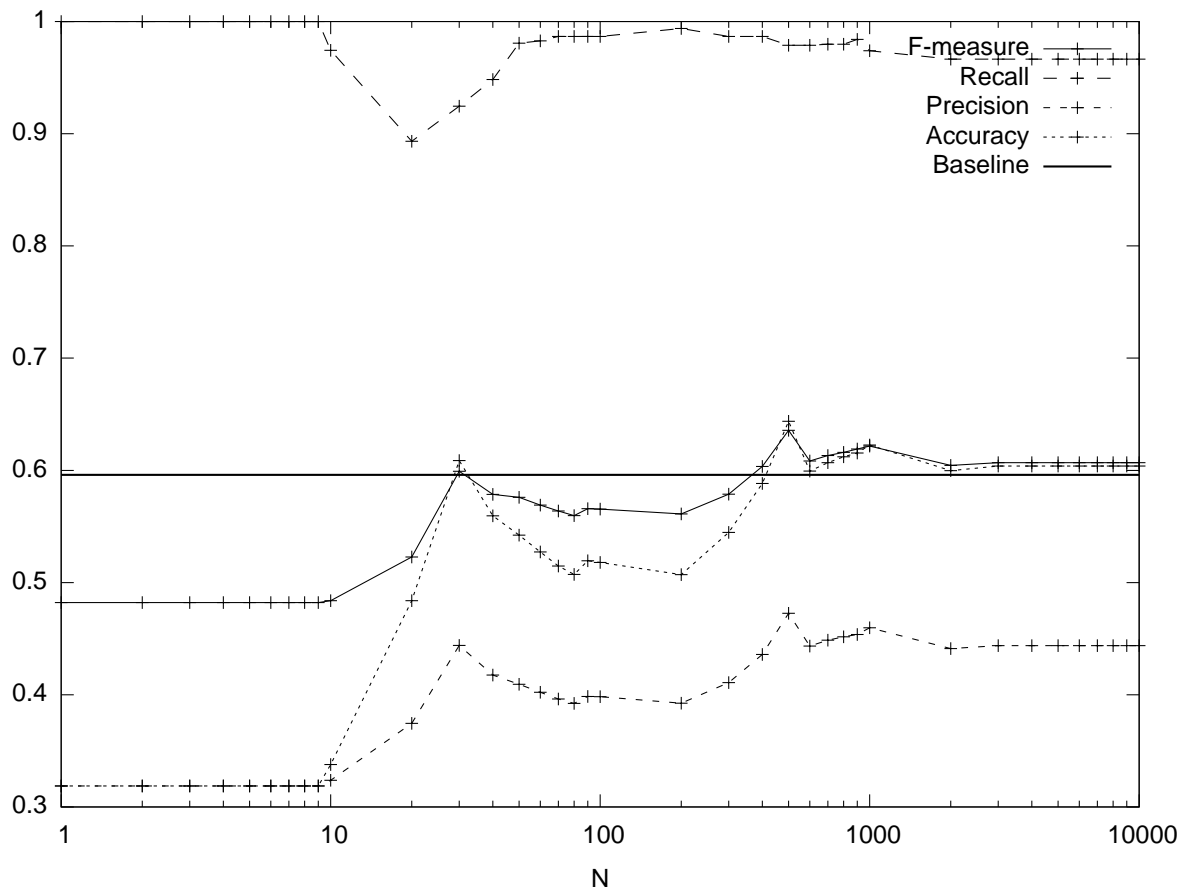


Figure 8.2: Prediction performance of feature selection using the top N absolute value score words for the touch sense.

Table 8.2: Feature selection: Optimal N and evaluation for each of the sensory viewpoints.

Sense	Optimal N	Precision	Recall	F	Baseline	Accuracy
Smell	500	0.4726	0.9788	0.6356	0.5960	0.6438
Taste	600	0.6503	0.9889	0.7839	0.7397	0.7071
Touch	200	0.4754	0.9728	0.6370	0.5646	0.6426
Vision	700	0.2488	0.8942	0.3872	0.2986	0.5755

were trained and tested using 5-fold cross validation with a training/test data ratio of 4:1. Evaluation of the prediction performance for increasingly larger N was calculated, which can be seen in Figure 8.2 for the smell sense. The N with the greatest average prediction performance from the 5 SVM models by F-measure is selected as the optimum model. For the smell sense, the baseline prediction performance is an F-measure of 0.59 for a model created by analyzing all feature words. Prediction performance peaked at an F-measure of 0.63 for a model created by analyzing 500 of the top ranking words. This indicates that the top 500 words are representative features for the smell viewpoint.

The feature selection process was applied to all four sense sentiment models. The optimal N for each of the sensory viewpoint model, the evaluation of the model, and the baseline F-measure are shown in Table 8.2. Models trained on optimal feature selection are used to predict the sense sentiment of wine tasting note visualization.

8.4 Visualization of Sensory Sentiment as Radar Charts

By reading the descriptions on a wine bottle, or a tasting note for a single wine, we might be able to roughly understand some of the wines characteristics without having a mastery of winespeak. However, it is much harder to grasp the characteristics of a wine region without reading about all the different wines produced. Sensory sentiment analysis from different viewpoints can provide an overview of the characteristics of a wine region, and then be plotted as a Radar Chart for easy comparison with other regions.

8.4.1 Model Normalization For Characteristic Prediction and Visualization

If a feature vector of a wine tasting note contains many feature words, then the sum of the predicted scores of these feature words would be greater than the sum of the predicted scores of a wine tasting note that only contains a subset of the same feature words. Also,

because each of the SVM classifiers for each sensory viewpoint were trained by 5-fold cross validation, the feature weights and therefore the prediction score range is different for each model. As the size of the feature vector and the SVM models that classify the sensory sentiments of wine tasting notes can influence the final score given, both the feature vector and SVM model prediction scores need to be normalized before visualization of the results.

When vectorizing the tasting notes of a region, the weight of each word in the feature vector was determined by Equation 8.1,

$$weight(w_i) = \frac{DF(w_i)}{\sqrt{\sum_{w_j \in W} DF(w_j)^2}} \quad (8.1)$$

where $DF(w_i)$ is the document frequency of the word w_i from the search query. This normalization ensures that a feature vector with many terms is not of greater weight than a feature vector that contains only a few terms. Thus, the number of terms does not influence the analysis of the characteristic features. Also the prediction score from each SVM classifier can be over a different range, and therefore the prediction score needs to be normalized so that a fair comparison can be made. Equation 8.2 was used to normalize the prediction scores for each feature vector from each SVM model,

$$norm(v_i, m_j) = \frac{score(v_i, m_j) - min(m_j)}{max(m_j) - min(m_j)} \quad (8.2)$$

where $score(v_i, m_j)$ is the predicted score for the feature vector v_i from the SVM model m_j , and the maximum and minimum model feature weights are represented by $max(m_j)$ and $min(m_j)$ for the model m_j .

8.4.2 Visualization of Sensory Sentiment by Region

In the data that was collected for analysis there are 4,675 regions, including major and sub-region combinations. The characteristics were calculated based on the wine tasting notes for each region as an example of sensory sentiment analysis. The chart with the largest summed score was from the Pelješac region in Croatia, as seen in Figure 8.3.

The chart shows three large positive scores, with the vision sense not scoring highly. The region with the smallest summed score was from the Primorska region in Slovenia as shown in Figure 8.4. The chart for this region shows that few sense descriptive feature words were used in the wine tasting notes, with only a slight emphasis on the taste and touch senses.

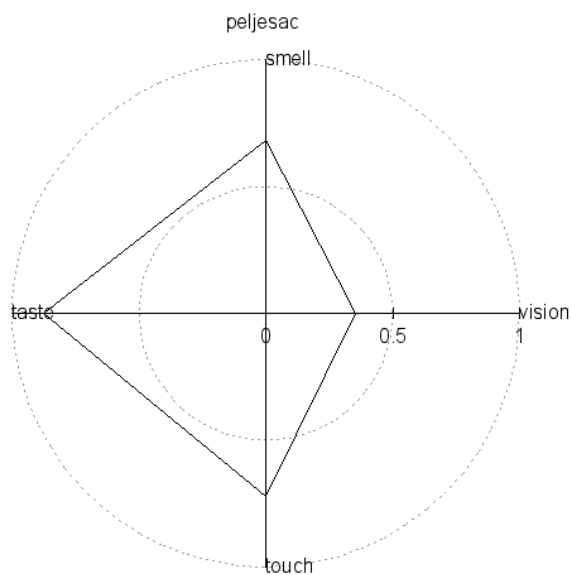


Figure 8.3: Example radar chart of the region Pelješac which has the largest graph area.

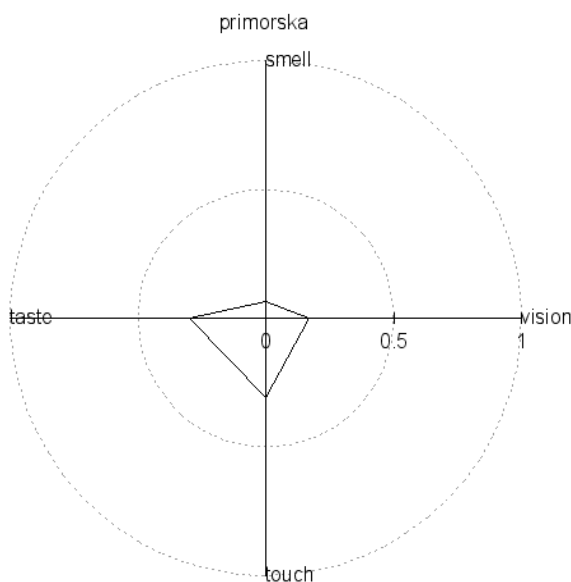


Figure 8.4: Example radar chart of the region Primorska which has the smallest graph area.

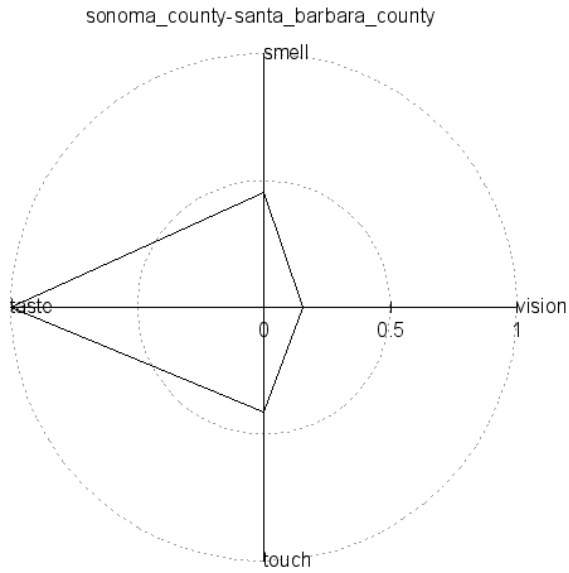


Figure 8.5: Strongest taste sentiment: Sonoma County, Santa Barbara County.

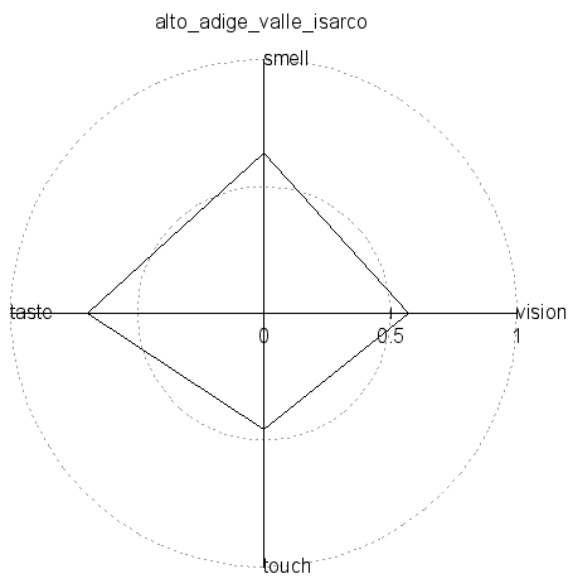


Figure 8.6: Strongest vision sentiment: Alto Adige Valle Isarco.

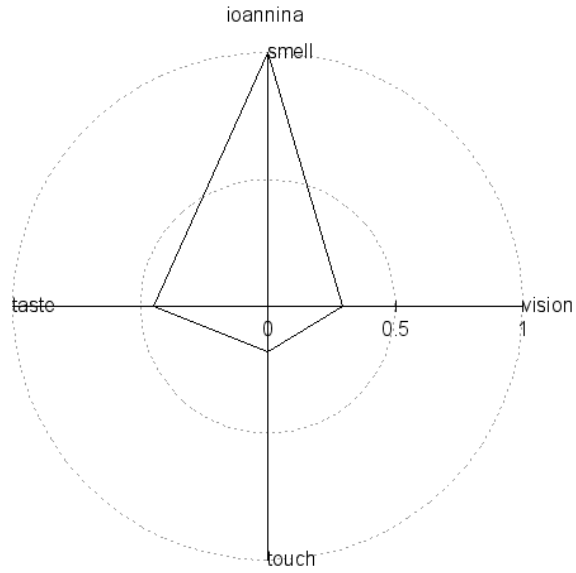


Figure 8.7: Strongest smell sentiment: Ioannina.

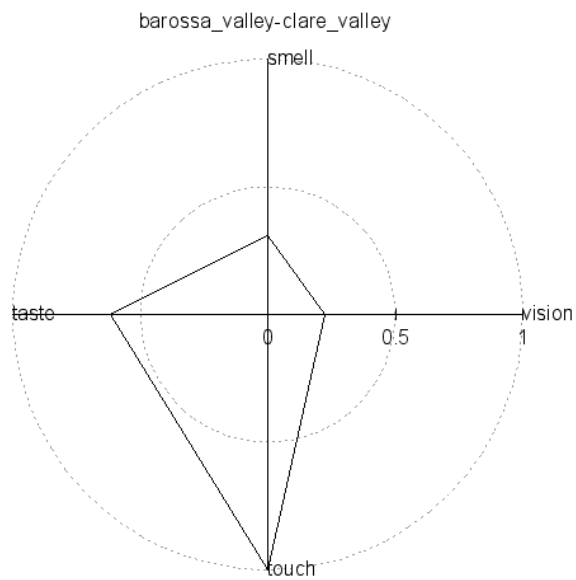


Figure 8.8: Strongest touch sentiment: Barossa Valley, Clare Valley.

Extreme sense sentiments can be seen in the charts of Figures 8.5, 8.6, 8.7, 8.8. These charts represent the highest score values for each of the four sense sentiment viewpoints. The chart for the Sonoma County, Santa Barbara County region in the USA shows a large number of features representing taste were used to describe the wines. This could suggest that the wines from that region have more taste qualities than smell, vision, and touch. In the chart for Alto Adige Valle Isarco region in Italy is the highest scoring region for the vision sense, but it would seem that negative scoring features for the vision sense are also prominent in the wine tasting notes. This would explain why the score is less than seen in other charts. The chart for the Ioannina in Greece scores highly on the smell sense viewpoint. This suggests that aromas play an important point in the description of the wines from that region. Lastly the chart for the Barossa Valley, Clare Valley region in Australia scores highly in sense sentiment for touch, suggesting feature words to do with the texture of the wine are often used.

8.5 Discussion

In this chapter, we analyzed 992 manually classified sentences of wine tasting notes to create SVM models from four sense sentiment viewpoints: vision, smell, taste, and touch. The models were evaluated and a search was performed to find the optimal feature selection for each model. The optimal models were then used to analyze the four sense sentiment viewpoints of 4,675 regions in a corpus consisting of 91,010 wine tasting notes. The results of the analysis were then normalized for fare comparison between models and sense sentiment viewpoints. Six examples of visualizations by Radar Chart were given representing the largest, smallest, and strongest sentiment for all four of the sense viewpoints.

Chapter 9

Analysis of The Diachronic Relations of Adjective Antonym Pairs in Wine Tasting Notes

Adjectives play the role of expressing attributes and are frequently used in wine tasting notes to describe opposite sensory properties of the wines along various different meaning dimensions, often concatenated in sentences to describe complex characteristics, as seen in the following example from Paul Gregutt’s tasting note of the Waterbrook 2002 Syrah¹: “This hits the palate with a crisp, clean, thin seam of flavor, balanced but quite light”. In this chapter, we investigate the change in antonym use in wine tasting notes over time. The diachronic analysis presented in this chapter, targets the antonyms thick and thin which were also part of a study investigating the use of antonymic adjectives in the BNC in Paradis et al. [52]. It was reported that use patterns relating to the width dimension suggested a positive negative antonym relation, expressing opposing poles of the dimension. However, in the case of wine tasting notes, polar opposites of the width dimension express negative qualities of wines. Therefore, it can be thought of as a double negative relation, with the middle of the dimension representing the optimal characteristic of the attribute. An example of this double negative relation can be seen in Steve Heimoff’s tasting note of the Woodbridge 2000 Merlot² expressing negative characteristics with the opposing poles of the width dimension: “Feels dry and thin on the palate, with a thick, unnaturally oaky taste”. Firstly, we will investigate the changes in use of thick and thin by the number of documents relating to the antonym pair by year. As words

¹<http://www.winemag.com/buying-guide/waterbrook-2002-shiraz-syrah-syrah-columbia-valley-wa>

²<http://www.winemag.com/buying-guide/woodbridge-2000-merlot-other>

surrounding a word indicate how and what context a word is used in, we will analyze changes in the frequency and the dissimilarity of adjectives that co-occur with thick or thin in the same tasting note. Secondly, we make use of visualizations of thick and thin to aid the analysis usage patterns and changes of closely related adjectives over time.

9.1 Data Collection

The target of the analysis is a corpus that consists of 91,010 wine tasting notes, or 255,966 sentences, that were collected from the Wine Enthusiast website³. The attributes of each wine, i.e. the date when the note was published or uploaded to the website, winery, region, and grape variety were collected along with the text of the wine tasting notes. The tasting notes in the corpus were published from 1999 to 2014. As many wine tasting notes were published in both the Wine Enthusiast magazine and on the website at different times, the earliest publication date was selected for analysis. An overview of the non-uniform distribution of publication dates by year can be seen in 9.1.

The corpus was indexed to construct a term document matrix based search engine using GETA⁴. The words were stemmed and a list of 2,488 adjectives that frequently occur in the wine tasting notes was used as a mask for word selection. The raw frequencies of the top 10 frequent terms by year are shown in Table 9.1.

As seen in Figure 9.1, there is a non-uniform distribution in the published tasting notes by year, which can cause bias when comparing the frequency of words from different years. To avoid this bias, the term frequency ratio for each year as shown in Equation 9.1 will be analyzed.

$$TR(w_i, y_k) = \frac{TF(w_i, y_k)}{\sum_{j \in W} TF(w_j, y_k)} \quad (9.1)$$

The term frequency $TR(w_i, y_k)$ of a word w_i occurring in a specific year y_k are divided by the sum of all the frequencies of words occurring in the same year. The occurrence percentage for each year of the top 10 frequent terms are shown in Figure 9.2.

³<http://buyingguide.winemag.com/>

⁴<http://geta.ex.nii.ac.jp/>

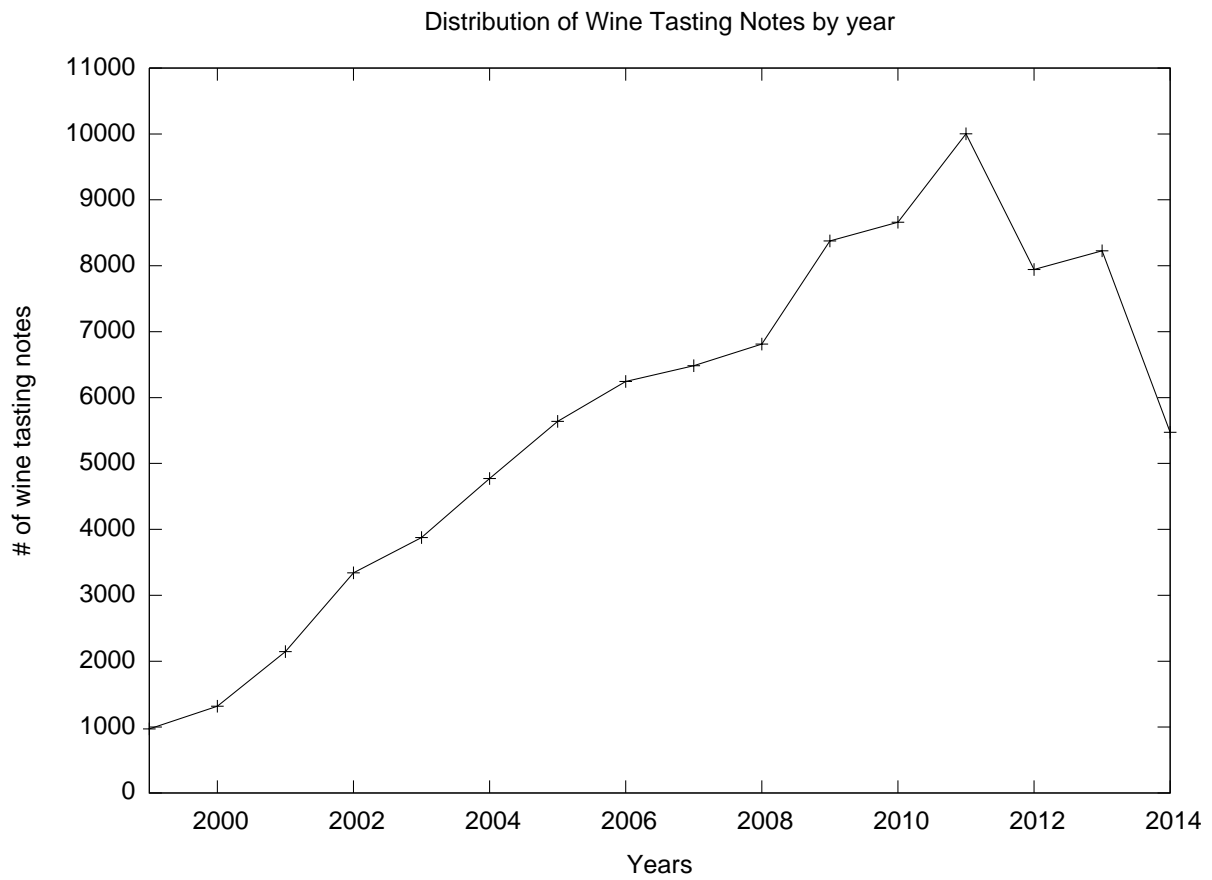


Figure 9.1: Distribution of wine tasting notes published from 1999 to 2014.

Table 9.1: Occurrence frequency of top 10 terms.

year	flavorful	fruitful	finishing	acidity	cherry	tannins	sweet -ness	dry	palate	ripeness
1999	397	559	356	121	160	143	120	124	144	131
2000	718	850	674	249	384	228	235	221	313	197
2001	1419	1236	1135	510	577	477	385	505	656	326
2002	2166	1904	1338	800	741	947	685	796	764	566
2003	2712	1728	1850	982	1024	1104	699	701	996	631
2004	3179	2253	1954	1129	1440	1266	1121	922	972	817
2005	3551	2321	2066	1246	1739	1255	1078	1371	1020	915
2006	4030	2846	2047	1662	1632	1303	1092	1410	1029	1214
2007	4172	2968	1868	1483	1874	1272	1241	1271	974	1084
2008	4634	3315	1731	1597	1862	1504	1411	1155	961	1188
2009	5242	3782	2130	1956	2406	1715	1962	1561	1321	1320
2010	4853	4312	2378	2409	2361	1966	1717	1788	1570	1318
2011	5550	4777	2850	2798	2593	2023	2611	1693	1742	1690
2012	4563	3714	2373	2166	1827	1682	1677	1166	834	1545
2013	4870	3823	2249	2556	1920	1960	1138	1233	1608	1794
2014	3029	2663	1704	2047	1109	1307	560	925	1493	1395

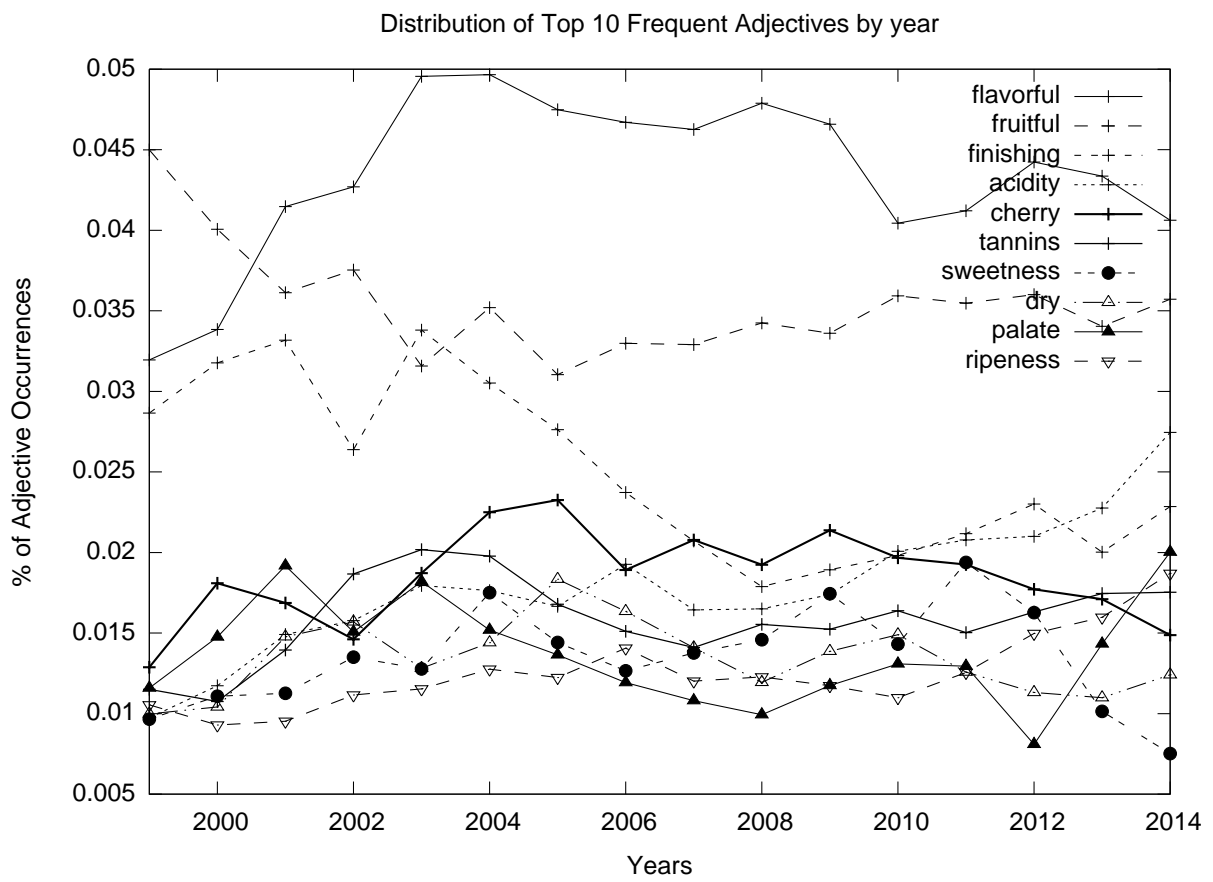


Figure 9.2: Top 10 frequently occurring terms by year.

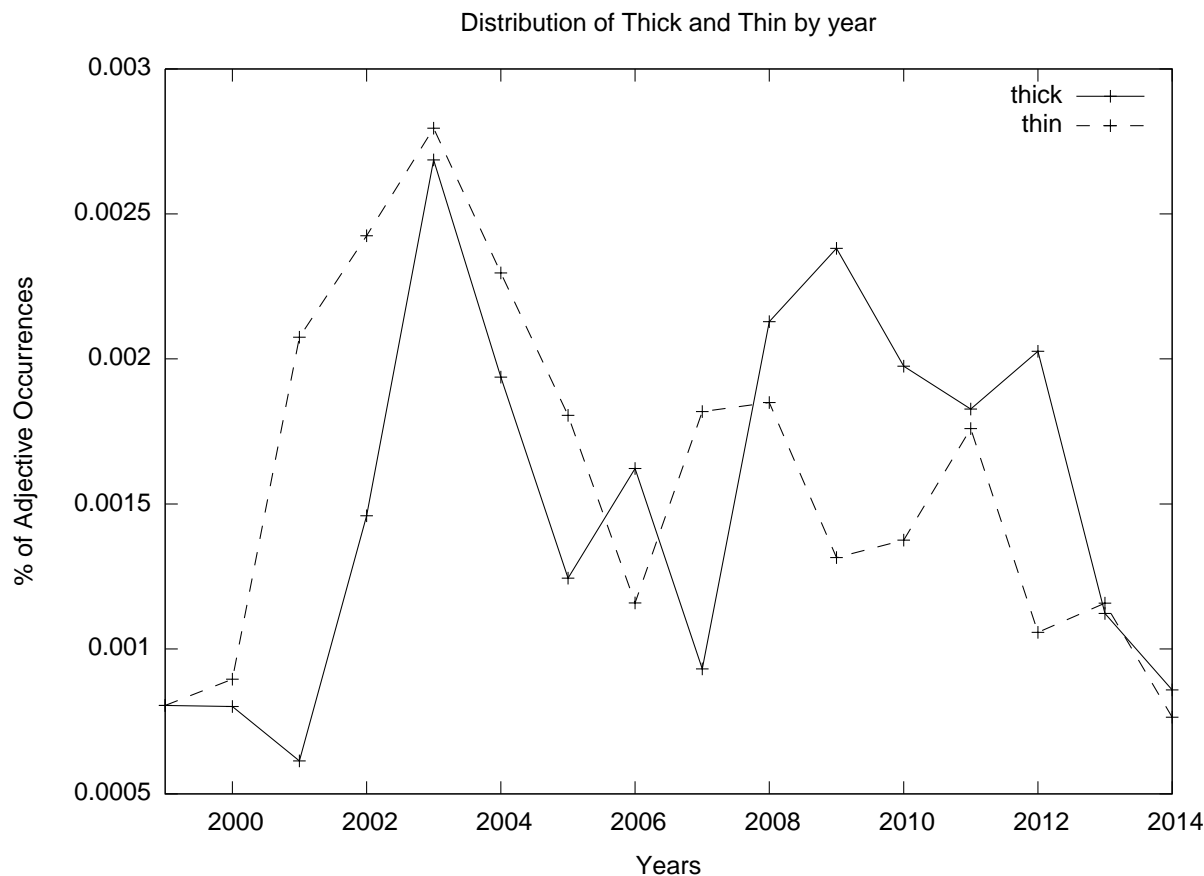


Figure 9.3: Distribution of the term frequency ratio of the adjectives thick or thin by year.

9.2 Diachronic Analysis of Antonyms by Frequency

In this section, we will examine the changes in the usage of the antonym pair thick and thin by analyzing the occurrence frequency of each word over time. Figure 9.3 shows the term frequency ratio of thick or thin in wine tasting notes from the year 1999 to 2014. It should be noted that for most years, wine tasting notes contain either thick or thin, and there are only a few exceptions that occur in the years 2000, 2002 and 2003 where both thick and thin occur in the same wine tasting note.

As mentioned in the previous section, there is a non-uniform distribution in the number of published wine tasting notes. To avoid bias, the term frequency ratio of the occurrence of the words thick and thin for each year are shown in Figure 3. A rise in usage of thin between 2000 and 2002 occurs before the increased usage of thick between 2001 and 2003, suggesting that thin seems to lead in usage or introduce thick to be used.

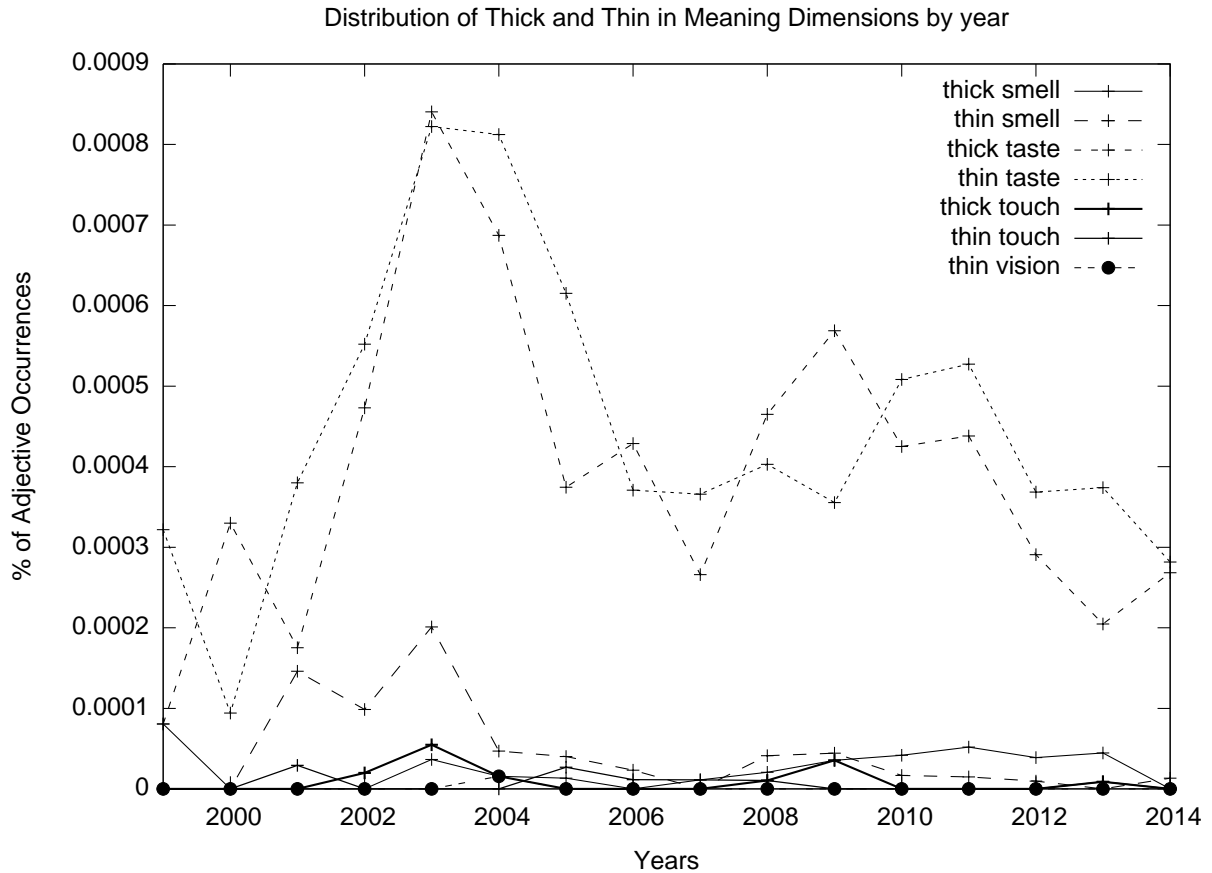


Figure 9.4: Distribution of thick and thin by year from the viewpoints of smell, taste, touch, and vision.

To further investigate the reason behind the changes in usage patterns, we can examine the changes from different sensory viewpoints. In previous work, we have examined the automatic prediction of sensory sentiment characteristics described in wine tasting notes [17]. This method involved training and evaluating support vector machine (SVM) models by analyzing wine tasting note data that has been manually classified into four previously defined [51] sensory viewpoints: smell, taste, touch, and vision.

We predicted the sensory viewpoints by applying the SVM models to each wine tasting note of the corpus. The distribution of thick and thin in relation to the sensory viewpoints is shown in Figure 9.4. The viewpoint of taste has a strong relation to the usage of thick and thin, with the usage of thin rising slightly before thick. It is interesting to see that the usage of only thin from the viewpoint of smell increased from the year 2001 to 2003 which is around the same time as the leading increase in the overall use of thin. This

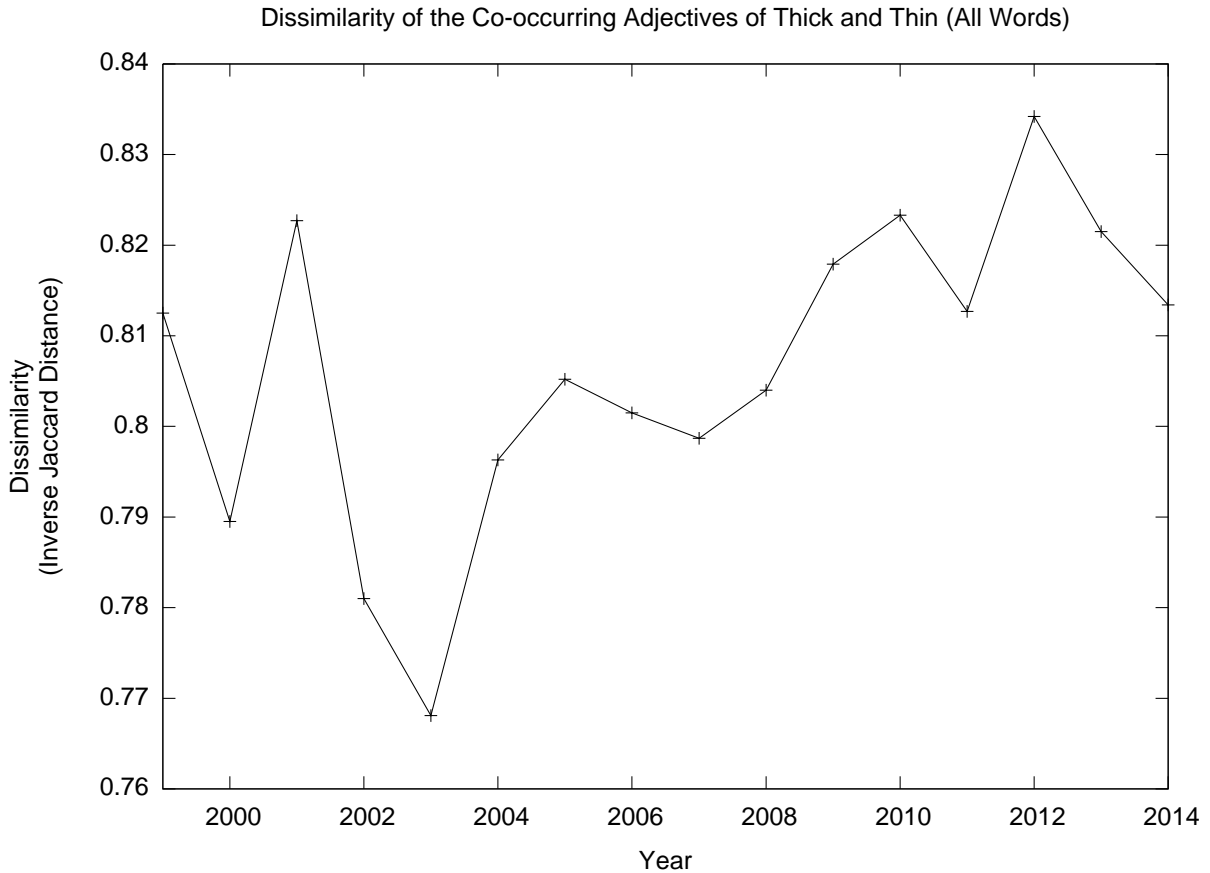


Figure 9.5: Dissimilarity of the co-occurring adjectives of thick and thin by year.

suggests that the use of thin from the smell viewpoint was a contributor to the rise in usage of thin before thick. Examining the relation of thick and thin with other adjectives over time can provide insight into the changes in use patterns in wine tasting notes. By measuring the co-occurring words between thick and thin by year we can show when the antonym pair are being used in similar patterns, and when they are not.

$$Dissimilarity(a, b, y) = 1 - \frac{|W(a, y) \cap W(b, y)|}{|W(a, y) \cup W(b, y)|} \quad (9.2)$$

Equation 9.2 is based on the Jaccard similarity distance and calculates the dissimilarity distance between two words a and b in the year y , where $W(a, y)$ represents the words of the documents that contain the word a and were published in a wine tasting note in year y .

The changes in the dissimilarity over the years of thick and thin are shown in Figure 9.5. In the year 1999 the dissimilarity between the pair is decreasing, until it reaches

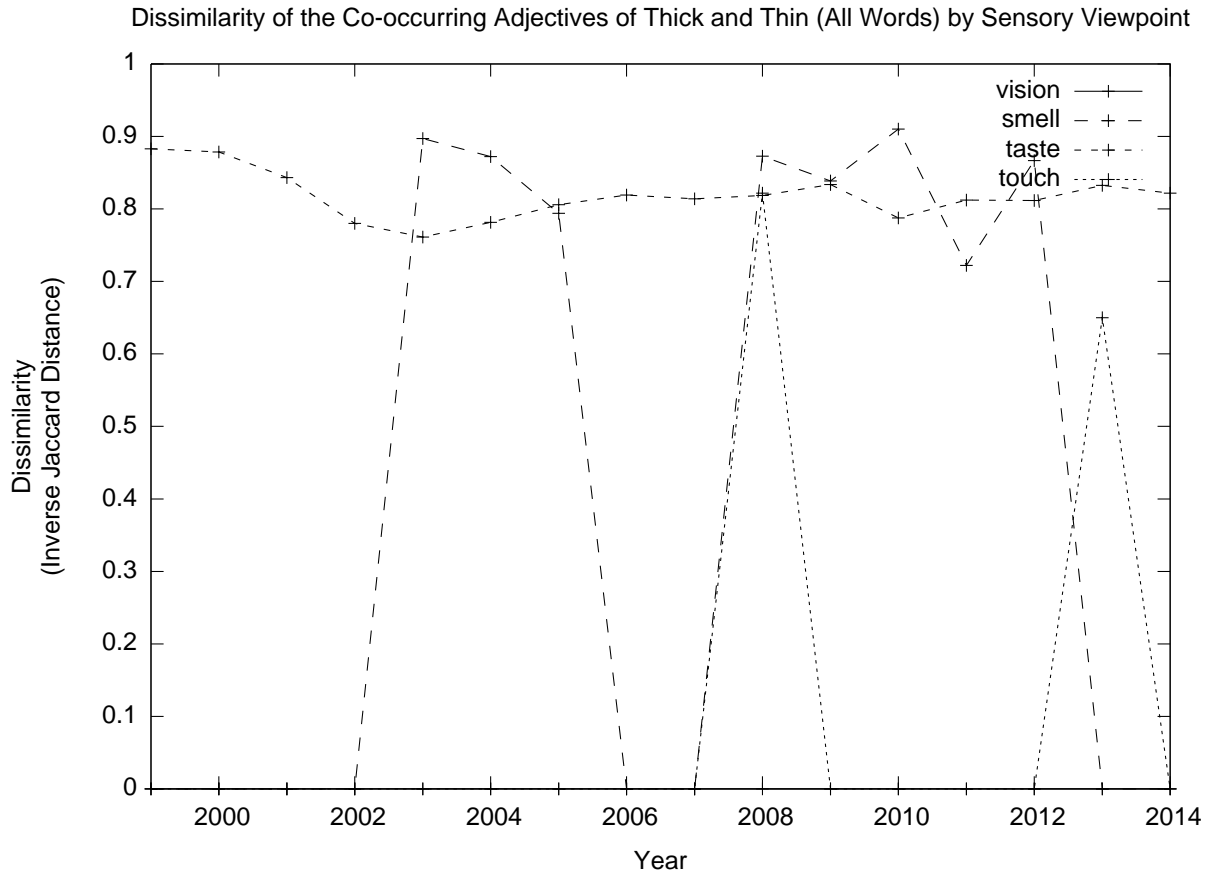


Figure 9.6: Dissimilarity of the co-occurring adjectives of thick and thin by year from the viewpoints of smell, taste, touch, and vision.

the lowest point in 2003. After this point, there is a steady increase in the dissimilarity distance. This could be interpreted as showing the increasing dissimilarity in use patterns of thick and thin.

Analyzing the dissimilarity distance from four sensory viewpoints, we can see that the dissimilarity between the co-occurring adjectives of thick and thin from the taste viewpoint has remained highly dissimilar without much change over the years. This would suggest that another viewpoint which was not the target of this analysis, and to a minor extent the smell viewpoint, are influential in the change of dissimilarity from around the year 2008 to 2012 as seen in Figure 9.6.

9.3 Diachronic Analysis of Antonyms by Mind Map

In this section we will investigate the differences in antonym use over time by observing changes in related words by mind map visualization.

9.3.1 Mind Map Generation

The method presented in this chapter for generating mind maps is based on a system that was proposed in [27, 12]. Firstly, a co-occurrence graph is generated of the words occurring in the documents returned by searching for the query words. The nodes represent the words within the documents, and the edges are weighted by the document frequency of co-occurring words. A mind map is generated by finding the minimum spanning tree of the co-occurrence graph. The root node of the mind map is specified as an initial search query from which the child nodes expand. Map expansion is limited to a set number of nodes or other features of the map. In this chapter, the number of related words (nodes) in the graph was limited to the 10 top ranking words related to the root node. Firstly, simple mind maps were generated for each year of wine tasting notes containing either of the antonyms: thick or thin.

An example of the mind map of wine tasting notes from the year 1999 containing the word thin is shown in Figure 9.7. The root node that contains the word thin and the number of wine tasting notes (documents) in which the word occurs: 9. Strong related word nodes expand out from the root node, in this case: short, flavorful, and fruitful are closely related to the root node thin.

Some of the mind maps generated contained common nodes, suggesting that there are common and distinct relations between antonym pairs. An example of this can be seen by comparing the mind map for thin in Figure 9.7 with the mind map for thick from the corresponding year in Figure 9.8. Both maps contain a fruitful node branching off from the root node. Other common nodes occur further away from the root node.

To identify and visualize the distinct and common characteristics of the thick/thin antonym pair, a merged mind map of the two single word root node maps can be generated by combining common nodes of both the thick and thin mind maps from the same year.

The merging process consolidates nodes that represent the same word in two or more different graphs, as seen in the example showing the merger of two sub maps in Figure 9.9. The consolidated node is connected by all the edges of the nodes it represents. The

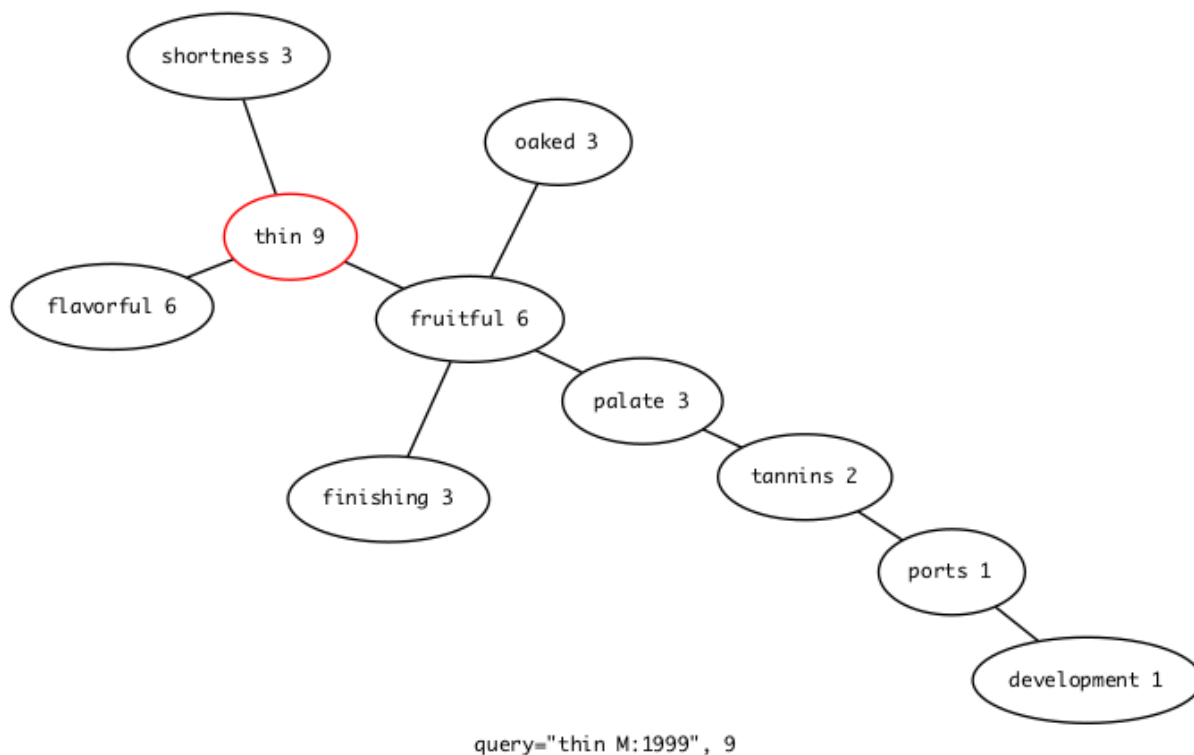


Figure 9.7: A mind map of thin from 1999.

consolidated nodes are represented by a dashed outline and no fill color. Edges connecting consolidated nodes are also represented by dashed lines. In the example, the root node of the merged map contains both of the words from the two sub maps, and they are colored according to the representation of distinct nodes in the merged map, in this case: thick is red, and thin is blue. The distinct nodes in the merged map are filled and outlined accordingly: thick only nodes are pink filled and red outlined, thin only nodes are cyan filled and blue outlined. In the example merged map the fruitful node is dashed representing a common node, and tightness and palate are colored representing distinct nodes from thick and thin maps respectively.

In the full merged mind map of thick and thin from 1999 shown in Figure 9.10, there are 4 common nodes (flavorful, fruitful, oaked, and finishing), 5 distinct thick nodes, and 5 distinct thin nodes near the perimeter of the mind map. Two of the common node words (flavorful, fruitful) are from the top 10 most frequent terms that are listed in Table 9.1.

The mind map of thick and thin from 2002 and 2004 as seen respectively in Figures 9.11 and 9.12 contain the highest ratio of common nodes out of all of the years, with only

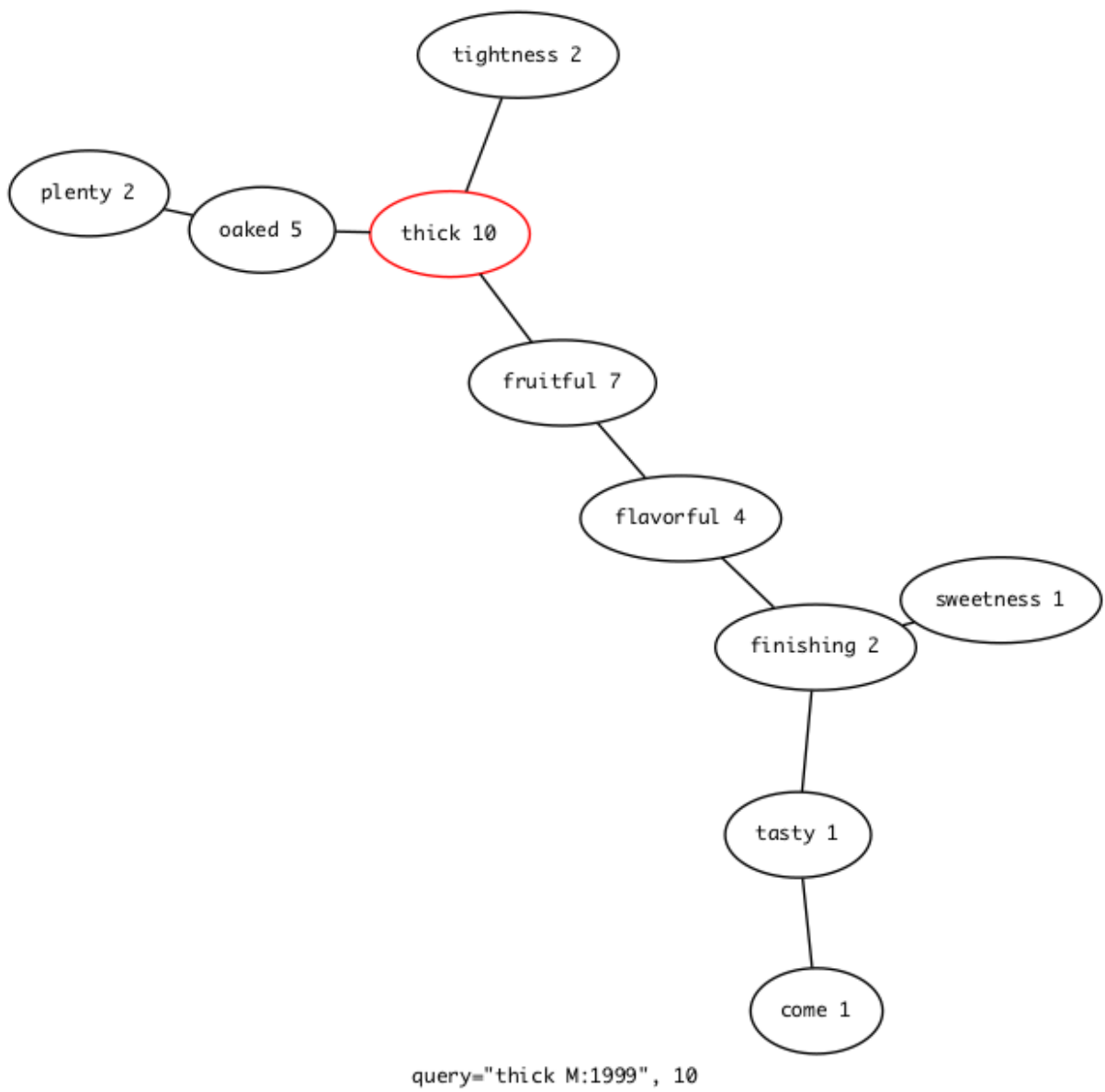


Figure 9.8: A mind map of thick from 1999.

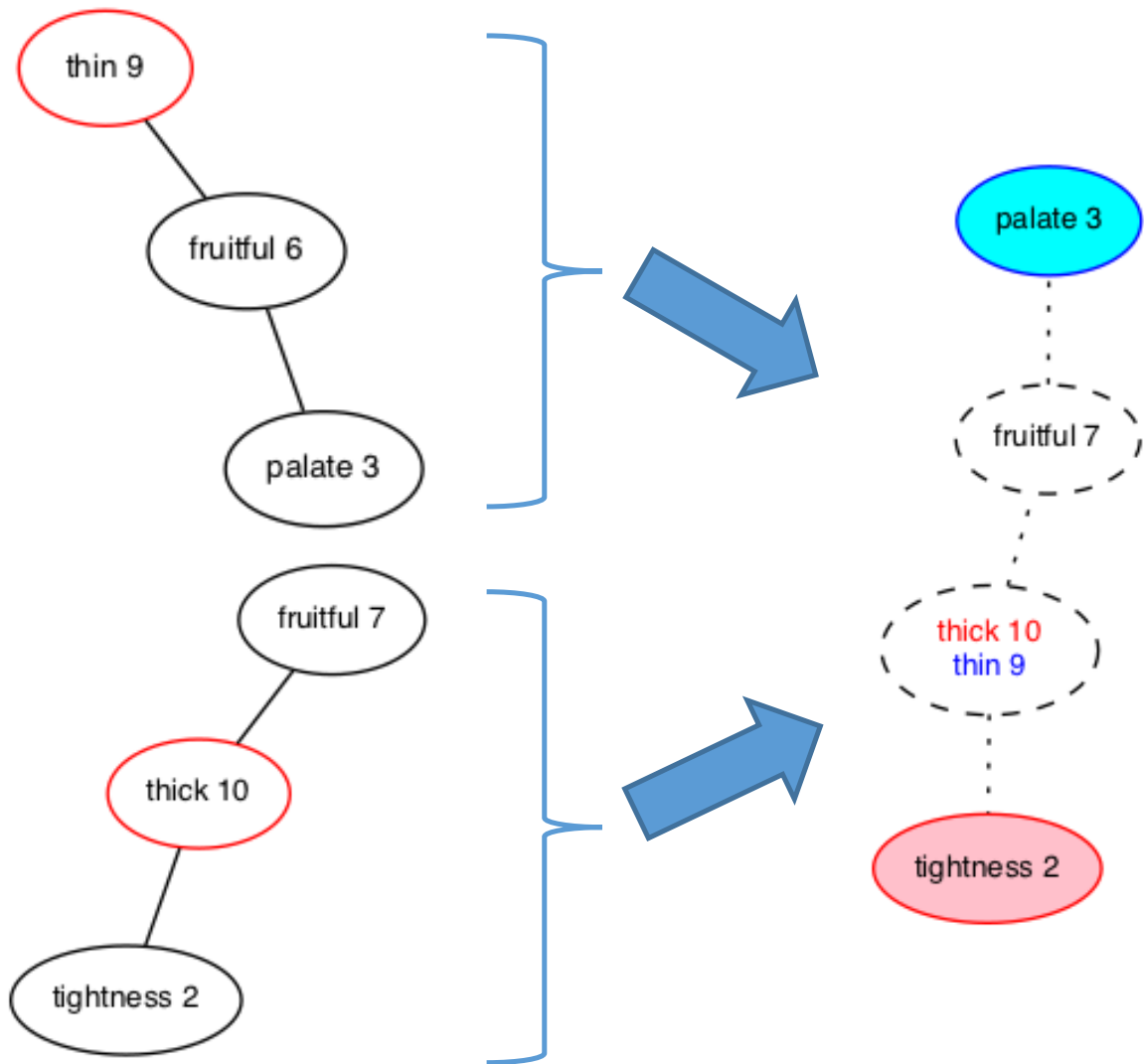


Figure 9.9: Merging mind maps of thick and thin to show distinct and common nodes.

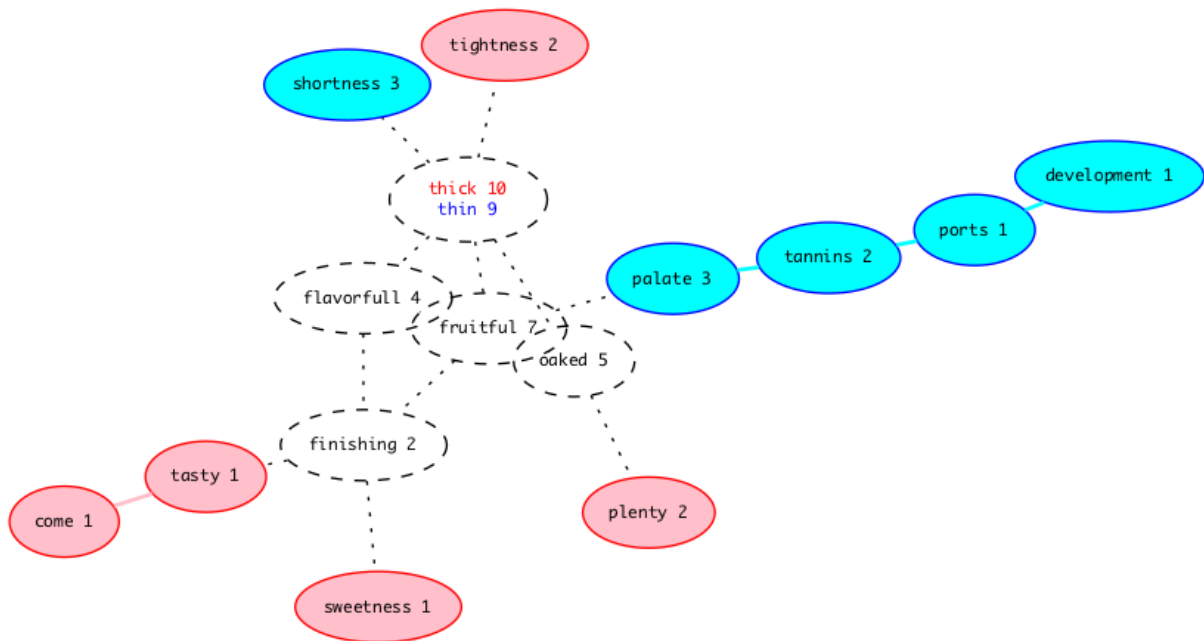


Figure 9.10: The merged mind map of thick and thin from 1999.

2 distinct thick and thin nodes for each year: richness, blackness, dry, and very thick nodes, and little, dry, acidity and lightness thin nodes. It is interesting to see that the dry node which was associated with thin in 2002 has become associated with thick in 2004. This suggests that dry is not affected by changes in the width meaning dimension.

The mind map with the highest ratio of distinct nodes is that from 2013 as seen in Figure 9.13. Only the nodes finishing, flavorful, and fruitful are common between the antonym pair. As with the map from 2002, richness and blackness are associated with the word thick, and dry and little with the word thin. The dissimilarity of mind maps from the same year can be analyzed to investigate how the relation between thick and thin has changed over time.

In figure 9.14, the Jaccard dissimilarity (see Equation 9.2) between the nodes of the mind maps of thick and thin are plotted over time to show the change in distinct and common nodes. As the mind maps are limited to 10 nodes each, only words with high relevance to thick and thin are analyzed. Compared with the dissimilarity analysis in the previous section, the dip in dissimilarity occurs a year earlier in 2002 and later in 2004, and then increases steadily with a pronounced trough in 2012.

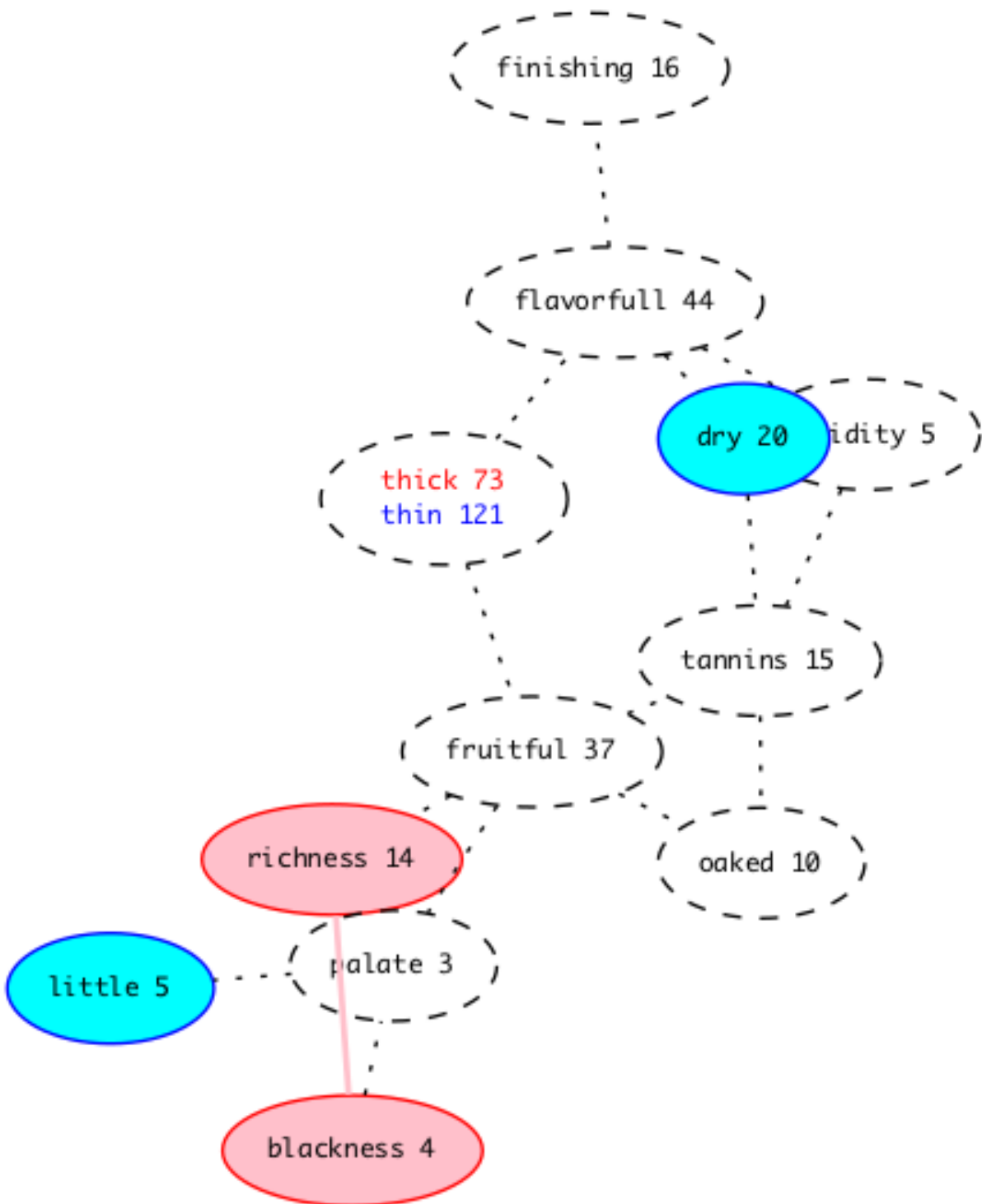


Figure 9.11: Mind map of thick and thin from 2002.

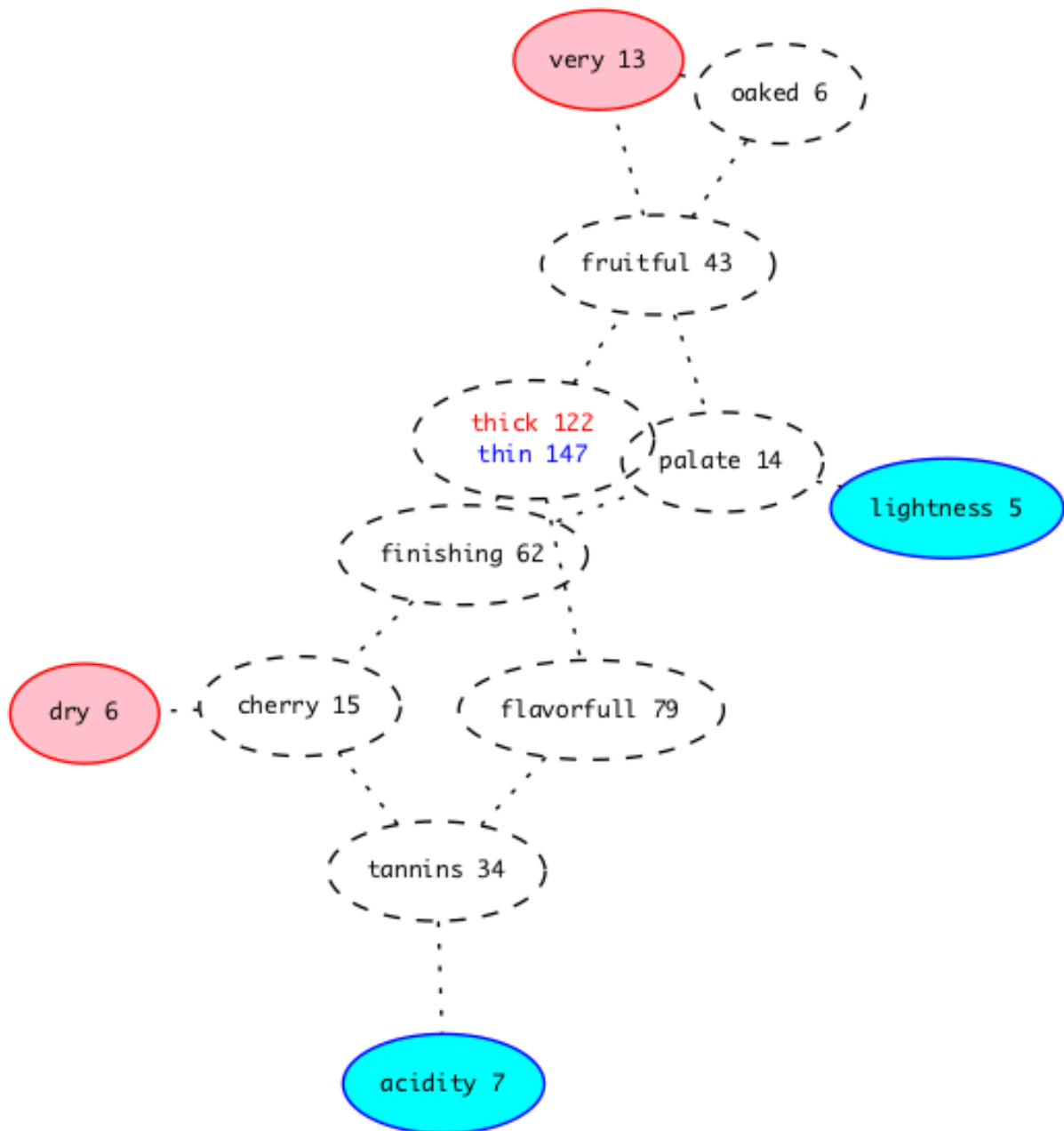


Figure 9.12: Mind map of thick and thin from 2004.

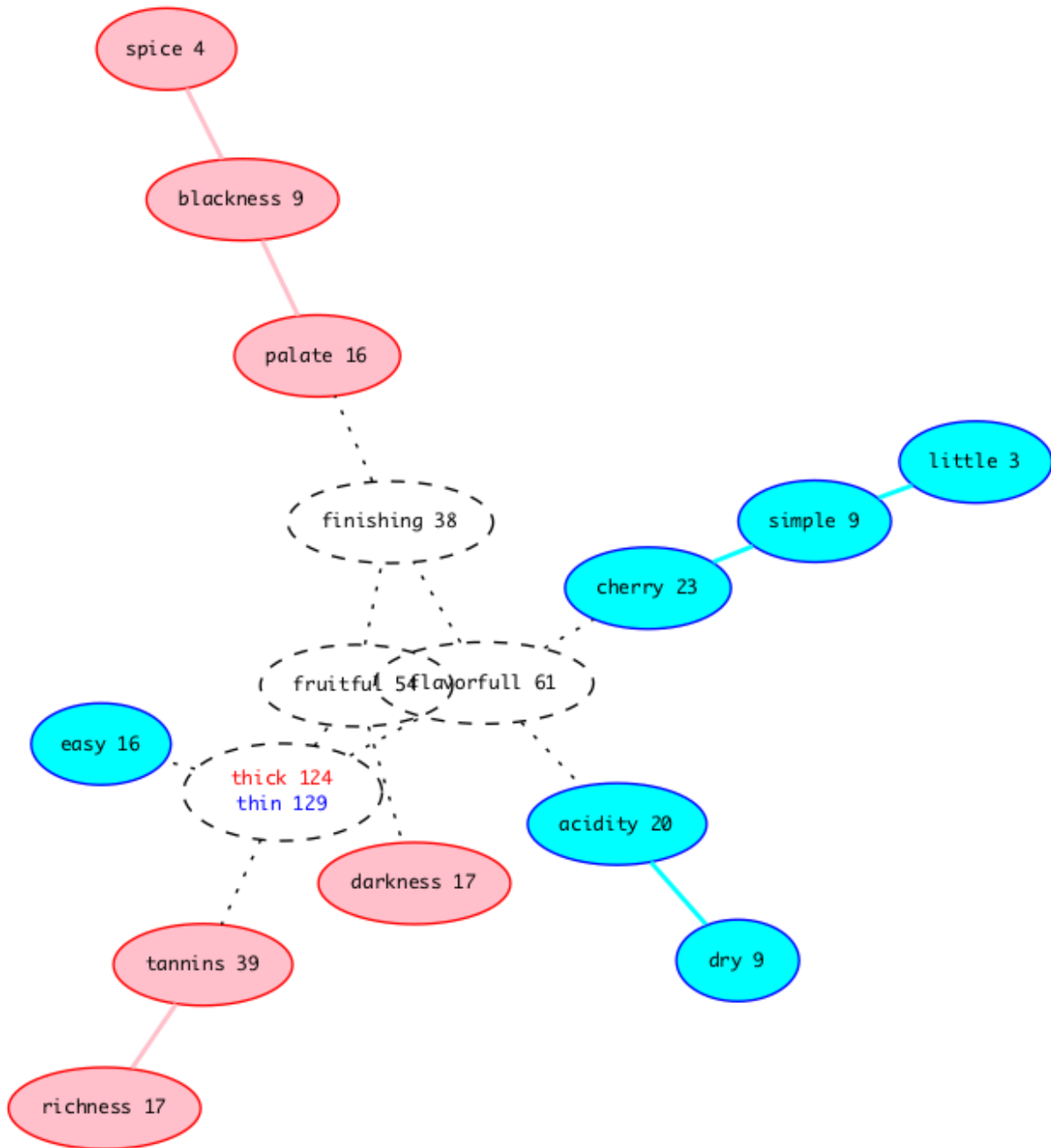


Figure 9.13: Mind map of thick and thin from 2013.

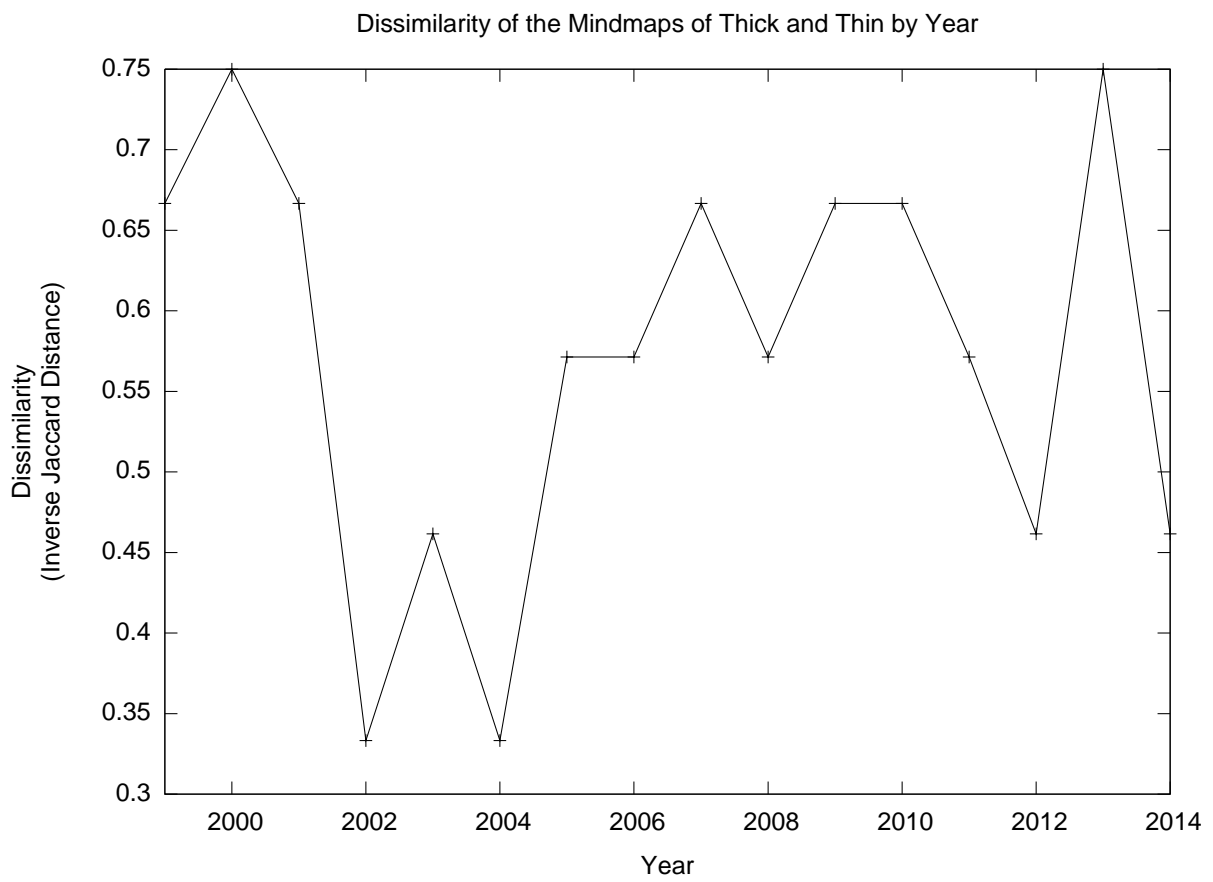


Figure 9.14: Jaccard dissimilarity of the mind maps of thick and thin by year.

9.4 Discussion

In this chapter, a diachronic analysis of the use patterns of the antonym pair thick and thin in wine tasting notes from the year 1999 to 2014 were examined. In particular, the following aspects were analyzed: the frequency of wine tasting notes in which the antonym pair occurred, the dissimilarity of co-occurring adjectives, and visualization of the dissimilarity by mind map. The analysis by co-occurring adjectives suggested an increase in the similarity of use patterns from 1999 to 2003, and then a steady increase in dissimilarity in the following year. The analysis by mind map also suggested a similar trend, however dissimilarity reached its lowest point a year earlier in 2002 and then later in 2004. Further analysis from sensory viewpoints revealed that there was a rise in the use of thick and thin from the taste viewpoint, and only in thin from the smell viewpoint. This suggests that the increased use of thin from the smell viewpoint was a contributor to the rise in usage of thin before thick.

Chapter 10

Analysis of Antonymic Adjective Meaning Dimensions in Winespeak

10.1 Introduction

Adjectives play the role of expressing attributes and are frequently used in wine tasting notes to describe opposite sensory properties of wines along various different meaning dimensions. To describe complex characteristics adjectives are often concatenated in sentences, as seen in the following example from Paul Gregutt’s tasting note of the Waterbrook 2002 Syrah¹: “This hits the palate with a crisp, clean, thin seam of flavor, balanced but quite light”. In this chapter, we examine the characteristics of adjective antonym pair use in winespeak. The analysis presented in this chapter targets the antonyms thick and thin which were also part of a study investigating the use of antonymic adjectives in the BNC in Paradis et al. [52]. It was reported that use patterns relating to the WIDTH dimension suggested a positive negative antonym relation, expressing opposing poles of the dimension. However, in the case of wine tasting notes, polar opposites of the WIDTH dimension express negative qualities of wines. The differences of the WIDTH dimension in the BNC corpus and Winespeak are shown in Figure 10.1.

It can be thought of as a double negative relation, where the optimal characteristic can be found in between the normal characteristic in the center and the outer extremes represented by thick and thin in the WIDTH dimension. An example of this double negative relation can be seen in Steve Heimoff’s tasting note of the Woodbridge 2000 Merlot² expressing negative characteristics with the opposing poles of the WIDTH dimension: “Feels

¹<http://www.winemag.com/buying-guide/waterbrook-2002-shiraz-syrah-syrah-columbia-valley-wa>

²<http://www.winemag.com/buying-guide/woodbridge-2000-merlot-other>

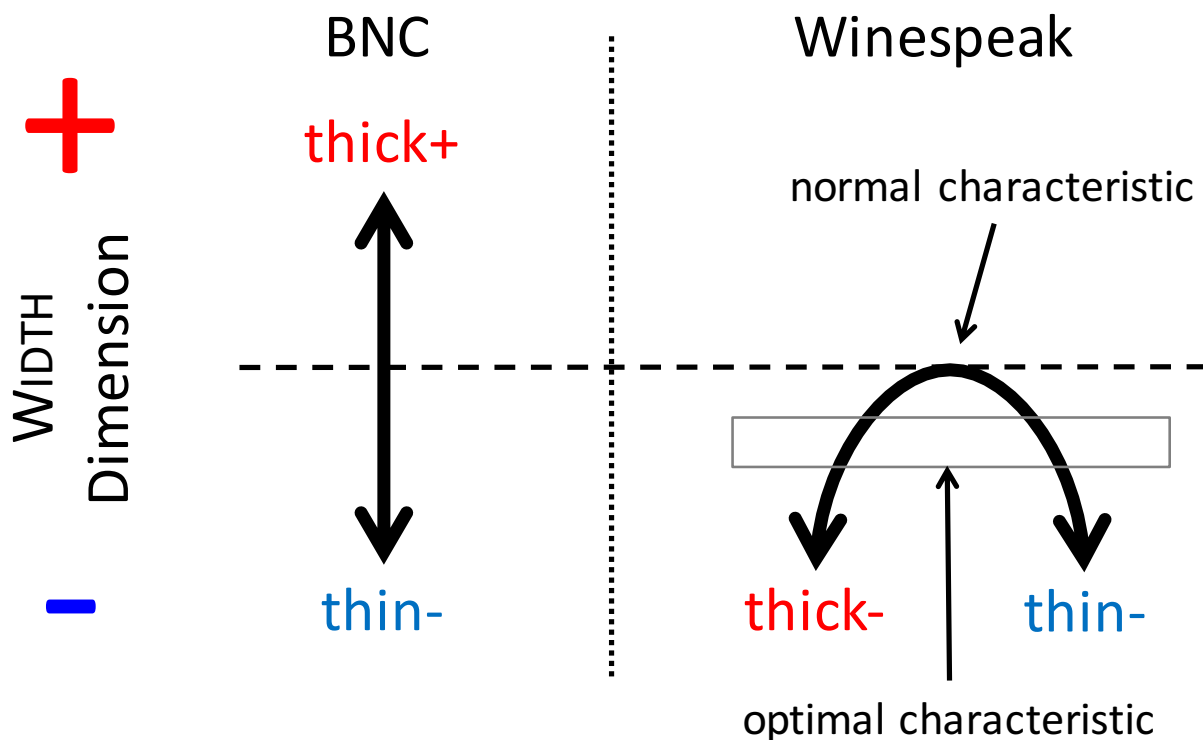


Figure 10.1: Relation of thick and thin: BNC corpus (left), Winespeak (right).

dry and thin on the palate, with a thick, unnaturally oaky taste”. In this chapter, we investigate the double negative relation of thick and thin in the WIDTH meaning dimension that is found in wine tasting notes. In particular, we extract the representative features of thick and thin, and examine the feature space of the WIDTH dimension by analyzing the feature weights of two SVM classifiers representing the two negative extremes thick and thin.

10.1.1 Data Collection

The target of the analysis in this chapter is a corpus that consists of 91,010 wine tasting notes, or 255,966 sentences, that were collected from the Wine Enthusiast website³. We have analyzed this corpus in chapter 9 to investigate the changes of adjectives over time in wine tasting notes. The attributes of each wine, i.e. the date when the note was published or uploaded to the website, winery, region, and grape variety were collected along with the text of the wine tasting notes. The tasting notes in the corpus were published from 1999 to 2014. The corpus was indexed to construct a term document matrix based search engine

³<http://buyingguide.winemag.com/>

Table 10.1: Top 10 positive and negative features of thick and thin.

#	Positive (<i>thick</i>)		Negative (<i>thin</i>)	
	$Score(w_i) \times DF(w_i)$	Feature	$Score(w_i) \times DF(w_i)$	Feature
1	11454.6068	tannin	-9611.2260	c:red
2	7618.9810	rich	-6231.9933	r:us
3	7004.5350	b:michael_schachner	-5465.6878	fruity
4	5906.1065	ripe	-4685.4468	acid
5	4718.4240	black	-4026.3960	b:s.h.
6	4440.8616	flavorful	-2901.4713	light
7	4081.9250	c:white	-2534.4176	dry
8	3901.5460	sweet	-2242.2750	b:roger_voss
9	3865.6640	full	-2190.3577	fresh
10	3649.3768	dark	-2039.0682	crisp

using GETA⁴. The words were stemmed and a list of 2,488 adjectives that frequently occur in wine tasting notes was used as a mask for word selection.

10.1.2 Features of Thick and Thin

In this section, we will examine the extremities of the WIDTH dimension which are represented by the adjectives thick and thin. A dataset for analysis by SVM classifier was created using 2,015 wine tasting notes that contain the word thick as the positive class, and 1,915 wine tasting notes that contain the word thin as the negative class. The words thick and thin were excluded from the feature set. The SVM^{light}[36] linear kernel classifier was used for model training and evaluation. Initially an SVM model was trained on all of the corpus data only for the purpose of feature scoring. The feature score was extracted by analyzing the weights of features in the SVM model trained on all the data. Features that have a high positive and negative score are representative of the thick class and thin class respectively. However some features that only represent a small portion of wine tasting notes are highly scored, such as wine or winery names. To overcome this, we analyze the following score: $Score(w_i) \times DF(w_i)$, where $DF(w_i)$ is the document frequency of the feature w_i . The characteristic features of thick and thin that were extracted using this method are shown in Table 10.1.

Features that are prefixed represent the following attributes of wine tasting notes: “b:” is the author of the wine tasting note, “c:” is the category of the wine, and “r:” is the

⁴<http://geta.ex.nii.ac.jp/>

appellation region of the wine in either singular or hierarchical form. All other features are words from wine tasting notes. These words are often used together in wine tasting notes. Of the top representative features, three of them are authors which suggests that the use of thick and thin is influenced by their writing style, or the regions that they are assigned to cover: Michael Schachner (b:michael_schachner) reviews wines from Argentina, Chile and Spain⁵, Steve Heimoff (b:s.h.) reviewed wines from California⁶, and Roger Voss (b:roger_voss) reviews wines from Portugal and France⁷.

The data was split into training and testing sets at a ratio of 9:1 for evaluation by 10-fold cross validation. The classification performance of an SVM model trained on all of the features was evaluated as a baseline. Feature selection was then performed by selecting increasingly larger sets of N top positive and N top negative score features and evaluating the prediction performance of each set as proposed in [55]. The set with the best prediction performance is therefore the optimal feature selection. The classification performance of each feature selection are shown in Figure 10.2. Feature selection performance peaks at an F-measure of 0.8873, with the optimal set made up of the top 500 positive and negative scored features.

10.2 The feature space of the width dimension

To investigate the complex relation between thick and thin in the WIDTH dimension, we split the data into three subsets: 2,015 notes that contain the word thick, 1,925 notes that contain the word thin, and 86,354 notes that do not contain either of the adjectives thick or thin (which we will refer to as *other*). The words thick and thin were excluded from the feature set. Two SVM classifiers were then trained: one with the thick note data as the positive class, and another with the thin note data as the positive class. The *other* note data was set as the negative class for both models. The weight score for each feature was then extracted from the two models. These two scores represent the relevance of a feature with in the WIDTH dimension. A plot visualizing the relation of the scores in two dimensions is shown in Figure 10.3.

Feature scores that are around or on the origin have weak relevance to the WEIGHT dimension. Features that have a high score in relation to both the thick and thin models

⁵<http://www.winemag.com/taster/michael-schachner/>

⁶https://en.wikipedia.org/wiki/Steve_Heimoff

⁷<http://www.winemag.com/taster/roger-voss/>

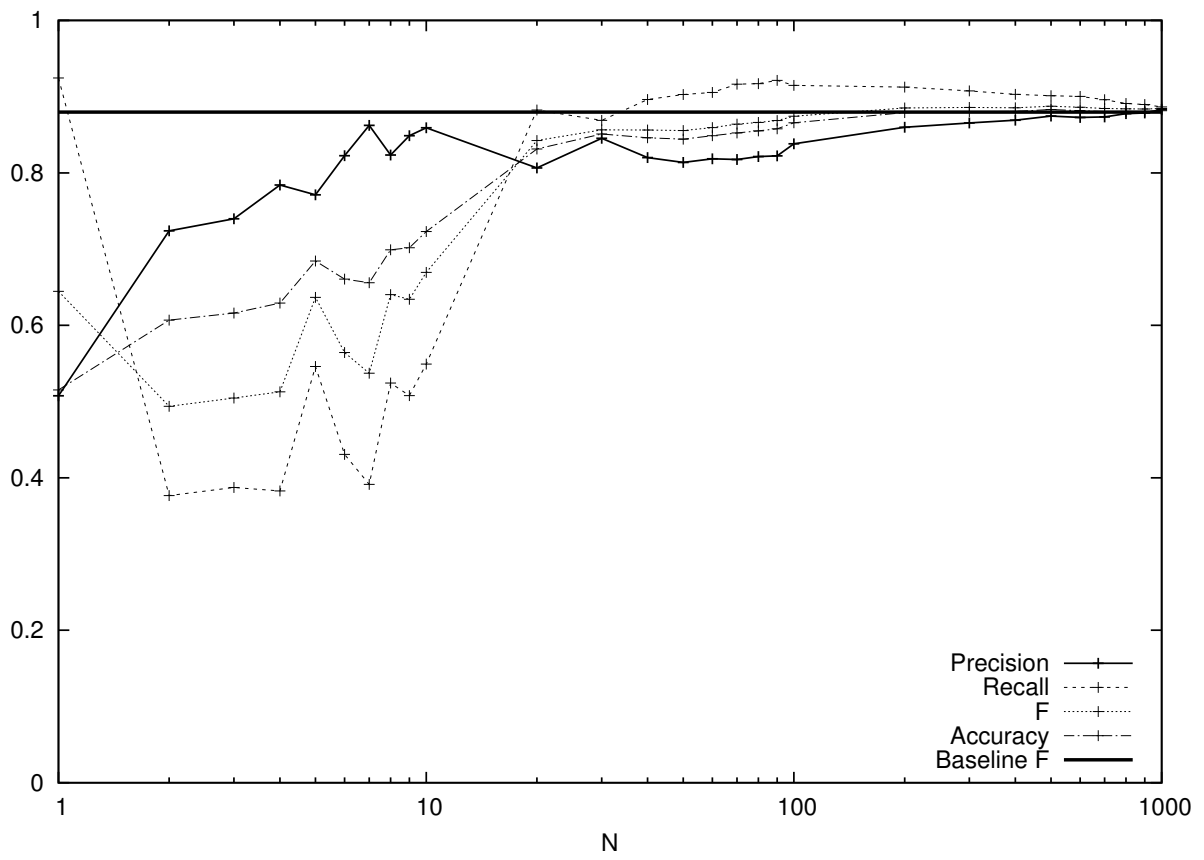


Figure 10.2: Feature selection performance of classifying thick and thin by SVM.

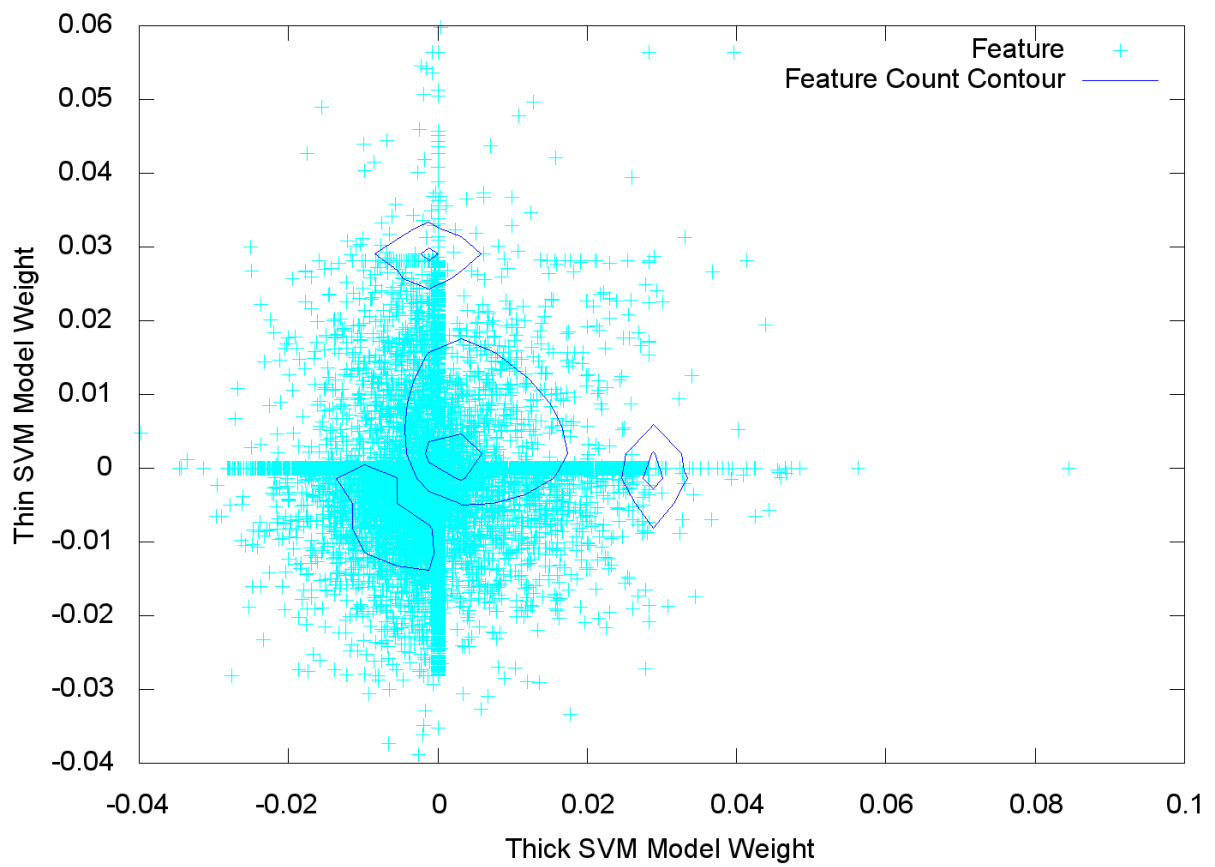


Figure 10.3: Scatter plot of feature scores from the thick model and thin model.

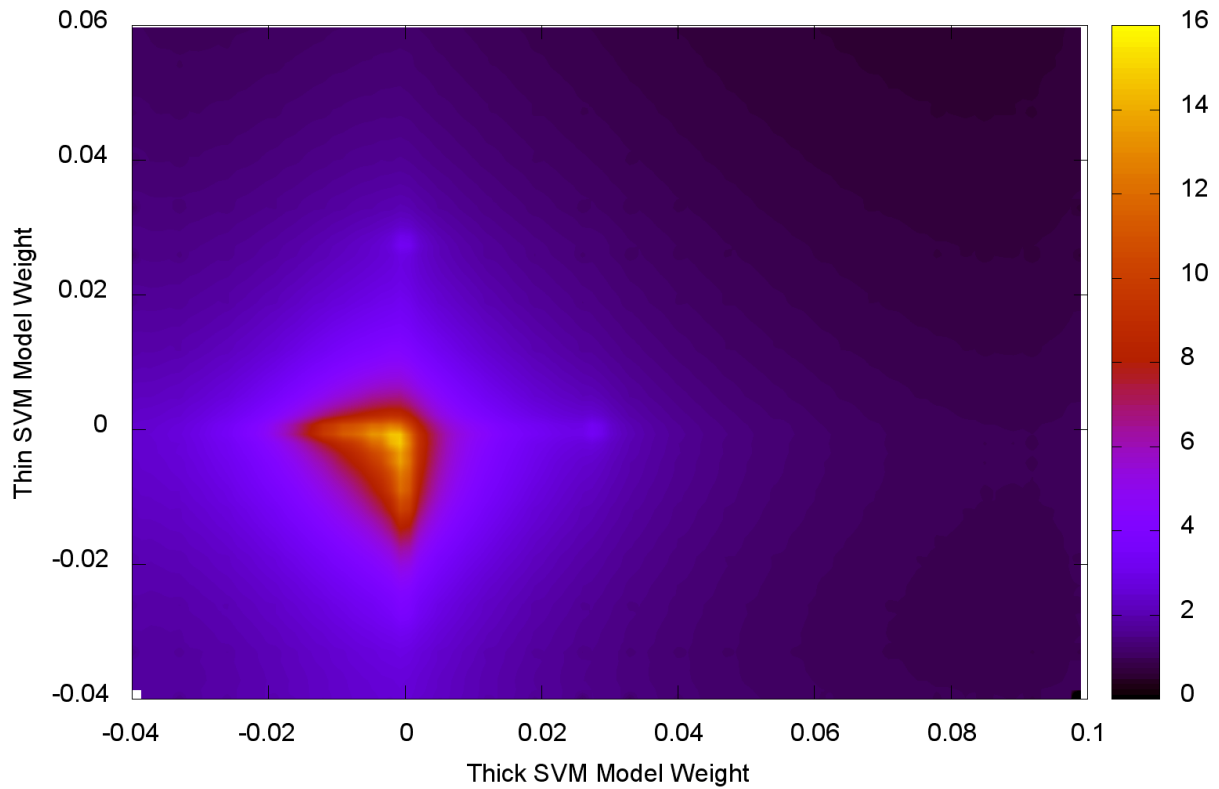


Figure 10.4: Interpolated feature density heat map.

are representative of the optimal characteristic within the WIDTH dimension. Conversely, features that are negatively scored in relation to both the thick and thin models are representative of the normal characteristic. The contour lines in the plot represent the peaks of dense clusters of features around the following areas: a mid-density cluster of features that have negative scores on both the thick and thin axis, a high density cluster around the origin that expands out positively on both axes, and two localized high density clusters of positive scored features on the thick and thin axis respectively. This suggests that there are many features that describe the normal and negative characteristics of the WIDTH dimension, but the cluster of features that represents the optimal characteristic is not well defined.

A heat map of the interpolated feature density was plotted, as shown in Figure 10.4, to examine the clusters in greater detail. The heat map confirms the presence of normal and negative characteristic feature clusters. However the density of features that represent the optimal characteristic of the WIDTH dimension decreases gradually away from the origin.

10.3 Discussion

Adjectives play a key role in the description of attributes in wine tasting notes. In this chapter, we analyzed the relation of the thick and thin adjective antonym pair in the WIDTH dimension. Representative features were extracted by analyzing the weight scores of a SVM model trained to classify wine tasting notes that contain the word thick or thin. A majority of the top 10 positive and negative features were adjectives that co-occur with thick and thin in wine tasting notes. To examine the feature space of the WIDTH dimension, two SVM models were trained to classify wine tasting notes that contain thick and thin respectively. Dense feature clusters were identified for the negative and normal characteristic of the WIDTH dimension. However a cluster representing the optimal characteristic was not well defined.

Chapter 11

Conclusion

This work investigates text mining techniques to analyze the support of foreign language learner expression from two perspectives: intention and emotion. The perspective of intention focused on errors that foreign language learners make when expressing *syntactic intention*. We proposed a method of analyzing foreign language learner writing errors in Chapter 3 using a corpus that was created by collecting data from a language learning SNS and manually tagging error occurrences. The significance of this method can be seen when comparing it to costly traditional techniques that are used to create learner writing corpus in controlled environments. The data was then analyzed by SVM to create 15 error category classification models. The performance of the models was then improved through optimal feature selection and the characteristic error features for each category were extracted. Previous work [60] had shown how learners from the same native language background tend to make similar characteristic errors. Based on this assumption, we investigate the interaction of co-occurring errors, and the differences of these error occurrences between native language groups by clustering analysis. The predicted error scores from the previous chapter were combined to form a 15 dimension vector representation of each corrected sentence of learner writing. These representations are then analyzed to: extract characteristic co-occurring errors by clustering analysis and a method of measuring the distance between different native language groups was proposed. In Chapter 5, we analyze the 15 dimension error vector representation to automatically predict the learners native language based on their writing errors. To the best of our knowledge, this is the first work that has proposed the automatic prediction of the learner's native language based on automatic error classification. In Chapter 7, we investigate the word features that are characteristic of different levels of proficiency, which is also another

learner characteristic that can influence errors in the expression of intention. The contribution of these researches is the ability to automatically profile learner characteristics that affect the expression of intention, and predict other characteristics based on common trait association. In Chapter 6, we address a problem that plagues research into writing error classification: the lack of data samples for training, testing and evaluating models and systems. We examine the automatic extraction and creation of a corpus of word order error samples from the logs of a language learning SNS. A large balanced corpus is automatically compiled and the technique is evaluated through the training, testing, evaluation of an SVM classification model. A sample corrected sentence from the corpus is examined in detail to reveal the effectiveness of extracting word order errors using the proposed technique.

A majority of research into sentiment analysis has focused on simple positive/negative sentiment relations. For the analysis of emotion expression in this thesis, we examined complex sentiment relations that are not usually the subject of research. These relations are of emotional sensory expressions in wine tasting notes. The main contribution of this work is that we examined beyond the usual sentiment analysis where there are just positive/negative sentiment relations, onward to different dimensions and therefore can't be analyzed using the usual techniques, such as: the double negative antonym dimension of thick and thin where the optimal characteristic resides in between the opposite poles. In Chapter 8, we investigate the emotional sensory expressions in wine tasting notes. Firstly, we present a method of classifying by SVM the four types sensory expressions that are commonly used in tasting notes. The models are then applied to a large corpus of wine tasting notes to predict the characteristics of wine regions and provide visualizations of the degree of sensory expression used in the review of wines in the region. In chapter 9, we examine changes in emotional sensory expressions in wine over the span of 14 years from the perspectives adjective antonym pairs that represent opposites of expressing attributes. A visualization system is proposed for analyzing differing properties, and examples where the change in use are examined in detail. Chapter 10 presents a detailed analysis of the adjective antonym meaning dimension of THICK and THIN. We proposed techniques that were used to examine and extract features of complex emotion expressions. The contribution of these researches is the ability to automatically analyze expressions of emotion, and provide methods to understand complex emotive expressions through visualization. These methods can be applied to a wide range of problems: from producing

materials from which language learners can study what can be used to express their emotion, to analyzing a range of complex emotions of products and services.

There still remains several topics to explore. We will leave the following questions open to future work.

On the topic of expressing intention, some syntactic errors can change the intended meaning of a sentence drastically, while other errors might have little effect. In this work, we addressed all syntactic errors equally without differentiating between the severity of errors. A method of estimating the amount of change in meaning could enable foreign language learners to focus on the correction of errors that have greater severity.

On the topic of expressing emotion, in this work we extract and visualize the features of specific sensory emotional expressions. This method could be applied to other emotional expression viewpoints by generalizing the process of analysis, and detecting the shape of the meaning dimensional to examine if it belongs to a general positive/negative relation, or that which is more complex, such as the adjective antonym meaning dimension of THICK and THIN which was analyzed in this work.

Bibliography

- [1] Abe, M., Frequency Change Patterns across Proficiency Levels in Japanese EFL Learner Speech, *Apples: Journal of Applied Language Studies*, Vol. 8, No. 3, pp. 85–96, 2014.
- [2] Bailey, S., Meurers, D., Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions, *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 107–115, 2008.
- [3] Bestgen, Y., Granger, S., Thewissen, J., Error Patterns and Automatic L1 Identification, *Approaching Language Transfer Through Text Classification*, pp. 127–153, 2012.
- [4] Brockett, C., Dolan, W.B., Gamon, M., Correcting ESL Errors Using Phrasal SMT Techniques, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-44)*, pp. 249–256, 2006.
- [5] Brooke, J., Hirst, G., Native Language Detection with ‘cheap’ Learner Corpora, In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the 1st Learner Corpus Research Conference*. Presses universitaires de Louvain, pp. 37–57, 2013.
- [6] Bykh, S., Meurers, D., Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization, *Proceedings of the International Conference on Computational Linguistics (COLING2014)*, pp. 1962–1973, 2014.
- [7] Caballero, R., Manner-of-motion Verbs in Wine Description, *Journal of Pragmatics*, Vol. 39, No. 12, pp. 2095–2114, 2007.

- [8] Chodorow, M., Tetreault, J.R., Na-Rae, H., Detection of Grammatical Errors Involving Prepositions, Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions (SigSem07), pp. 25–30, 2007.
- [9] Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press, 2001.
- [10] Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S., Intelligent Computer Classification of English Writing Errors, Intelligent Interactive Multimedia Systems and Services, Vol. 254 , pp. 174–183, 2013.
- [11] Flanagan, B., Yin, C., Hashimoto, K., Hirokawa, S., Clustering English Writing Errors Based on Error Category Prediction, Proceedings of the 3rd International Symposium on Engineering, Energy and Environments (ISEEE2013), pp. 733–738, 2013.
- [12] Flanagan, B., Yin, C., Inokuchi, Y., Hirokawa, S., Supporting Foreign Language Learning Using Mind-Maps, Information And Systems In Education (JSiSE), Vol. 12, No. 1, pp. 13–18, 2013.
- [13] Flanagan, B., Yin, C., Hirokawa, S., Hashimoto, K., Tabata, Y., An Automated Method to Generate e-Learning Quizzes from Online Language Learner Writing, International Journal of Distance Education Technologies (IJDET), Vol. 11, No. 4, pp. 63–80, 2013.
- [14] Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S., Classification of English Language Learner Writing Errors Using a Parallel corpus with SVM, International Journal of Knowledge and Web Intelligence (IJKWI), Vol. 5, No. 1, pp. 21–35, 2014.
- [15] Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S., Classification and Clustering English Writing Errors Based on Native language, Proceedings of the 3rd IIAI International Conference on Advanced Applied Informatics (LTLE2014), pp. 318–323, 2014.
- [16] Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S., Prediction of Learner Native Language by Writing Error Pattern, Learning and Collaboration Technologies (Springer LNCS 9192), pp. 87–96, 2015.

- [17] Flanagan B., Wariishi N., Suzuki T., Hirokawa S., Predicting and Visualizing Wine Characteristics Through Analysis of Tasting Notes From Viewpoints, *HCI International 2015-Posters' Extended Abstracts*, pp. 613–619, 2015.
- [18] Flanagan, B., Hirokawa, S., The Relationship of English Foreign Language Learner Proficiency and an Entropy Based Measure, *Information Engineering Express (IEE)*, Vol. 1, No. 3, pp. 29–38, 2015.
- [19] Flanagan, B., Paradis, C., Hashimoto, K., Hirokawa, S., Analysis of The Diachronic Relations of Adjective Antonym Pairs in Wine Tasting Notes, *Proceedings of the 21st International Symposium on Artificial Life and Robotics 2016 (AROB2016)*, pp. 749–754, 2016.
- [20] Flanagan, B., Hashimoto, K., Hirokawa, S., Analysis of Antonymic Adjective Meaning Dimensions in Winespeak, *Proceedings of the 11th International Symposium on Natural Language Processing (SNLP2016)*, CD, 2016.
- [21] Flanagan, B., Hirokawa, S., Automatic Extraction and Prediction of Word Order Errors From Language Learning SNS, *Proceedings of the 5th IIAI International Congress on Advanced Applied Informatics (LTLE2016)*, pp. 292–295, 2016.
- [22] Flanagan, B., Hirokawa, S., Kaneko, E., Izumi, E., Classification of Speaking Proficiency Level by Machine Learning and Feature Selection, *Proceedings of the 1st International Symposium on Emerging Technologies for Education, (CD)*, 2016.
- [23] Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W.B., Belenko, D., Vanderwende, L., Using Contextual Speller Techniques and Language Modeling for ESL Error Correction, *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 449–456, 2008.
- [24] Han, N.R., Chodorow, M., Leacock, C., Detecting Errors in English Article Usage with a Maximum Entropy Classifier Trained on a Large, Diverse Corpus, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, pp. 1625–1628, 2004.

- [25] Han, N.R., Chodorow, M., Leacock, C., Detecting Errors in English Article Usage by Non-native Speakers, *Natural Language Engineering*, Vol. 12, No. 2, pp. 115–129, 2006.
- [26] Hirano, T., Hirate, Y., Yamana, H., Detecting Article Errors in English using Search Engines, *DBSJ Letters* 6, No. 3, pp. 13–16, 2007. (in Japanese)
- [27] Hirokawa, S., Flanagan, B., Suzuki, T., Yin, C., Learning Winespeak from Mind Map of Wine Blogs, In S. Yamamoto (Ed.): *Human Interface and the Management of Information Part II (Springer LNCS 8522)*, pp. 383–393, 2014.
- [28] Ionescu, R., Popescu, M., Cahill, A., Can Characters Reveal your Native Language? A Language-independent Approach to Native Language Identification, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2014)*, pp. 1363–1373, 2014.
- [29] Izumi, E., Uchimoto, K., Isahara, H., *The NICT JLE corpus*, ACL Publishing, 2004. (in Japanese)
- [30] Izumi, E., Uchimoto, K., Isahara, H., *The NICT JLE Corpus: Exploiting the Language Learner’s Speech Database for Research and Education*, *International Journal of the Computer, the Internet and Management*, Vol. 12, No. 2, pp. 119–125, 2004.
- [31] Izumi, E., Uchimoto, K., Isahara, H., *The Overview of the SST Speech Corpus of Japanese Learner English and Evaluation through the Experiment on Automatic Detection of Learners’ Errors*, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, pp. 1435–1438, 2004.
- [32] Jarvis, S., Castaneda-Jiménez, G., Nielsen, R., *Investigating L1 Lexical Transfer Through Learners’ Wordprints*, *Second Language Research Forum (SLRF)*, 2004.
- [33] Jarvis, S., Crossley, S.A., *Approaching Language Transfer Through Text Classification: Explorations in the Detection based Approach*, *Multilingual Matters*, Vol. 64, 2012.
- [34] Jarvis, S., Bestgen, Y., Pepper, S., *Maximizing Classification Accuracy in Native Language Identification*, *Proceedings of the 15th Annual Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT2013), pp. 111–118, 2013.
- [35] Joachims, T., Training Linear SVMs in Linear Time, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–226, 2006.
- [36] Joachims, T., Learning to Classify Text using Support Vector Machines: Methods, Theory and Algorithms, Kluwer Academic Publishers, 2002.
- [37] Karypis, G., CLUTO - A Clustering Toolkit, No. TR-02-017. Minnesota University, Minneapolis, Dept of Computer Science, 2002.
- [38] Kerren, A., Prangova, M., Paradis, C., Visualization of Sensory Perception Descriptions, Proceedings of the International Conference on Information Visualization, pp. 135–144, 2011.
- [39] Kochmar, E., Identification of a Writer’s Native Language by Error Analysis, Master’s thesis, University of Cambridge, 2011.
- [40] Koppel, M., Schler, J., Zigdon, K., Determining an Author’s Native Language by Mining a Text for Errors, Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 624–628, 2005.
- [41] Kroll, B., What Does Time Buy? ESL Student Performance on Home Versus Class Compositions, In B. Kroll (Ed.), Second language writing: Research insights for the classroom, Cambridge: Cambridge University Press, pp. 140–154, 1990.
- [42] Leacock, C., Chodorow, M., Gamon, M., Tetreault, J., Automated Grammatical Error Detection for Language Learners, Synthesis lectures on human language technologies, Vol. 7, No. 1, pp. 1–170, 2014.
- [43] Levenshtein, V., Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, Vol. 10, No. 8, pp. 707–710, 1966.
- [44] Lo, J.J., Wang, Y.C., Yeh, S.W., WRITE: Writing Revision Instrument for Teaching English, Proceedings of the International Conference on Technologies for E-Learning and Digital Entertainment, pp. 1–8, 2008.

- [45] Medhat, W., Hassan, A., Korashy, H., Sentiment Analysis Algorithms and Applications: A Survey, *Ain Shams Engineering Journal*, Vol. 5, No. 4, pp. 1093–1113, 2014.
- [46] Miki, N., A New Parallel Corpus Approach to Japanese learners' English, Using Their Corrected Essays, *Themes in Science and Technology Education*, Vol. 3, No. 1–2, pp. 159–180, 2011.
- [47] Miyake, H., Tsushima, T., On The Features of There Constructions Used by Japanese Speakers of English, *The Journal of Humanities & Natural Sciences*, Vol. 132, pp. 55–79, 2012.
- [48] Mizumoto, T., Komachi, M., Nagata, M., Matsumoto, Y., Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 147–155, 2011.
- [49] Odlin, T., Cross-linguistic Influence, *The Handbook of Second Language Acquisition*, pp. 436–486, 2003.
- [50] Page, E.B., The Use of the Computer in Analyzing Student Essays, *International Review of Education*, Vol. 14, No. 2, pp. 210–225, 1968.
- [51] Paradis, C., Eeg-Olofsson, M., Describing Sensory Experience: The Genre of Wine Reviews, *Metaphor and Symbol*, Vol. 28, No. 1, pp. 22–40, 2013.
- [52] Paradis, C., Löhndorf, S., van de Weijer, J., Willners, C., Semantic Profiles of Antonymic Adjectives in Discourse, *Linguistics*, Vol. 53, No. 1, pp. 153–191, 2015.
- [53] Petersen, S.E., Ostendorf, M., A Machine Learning Approach to Reading Level Assessment, *Computer Speech and Language*, Vol. 23, No. 1, pp. 89–106, 2009.
- [54] Polio, C., Fleck, C., “If I only had more time:” ESL Learners' Changes in Linguistic Accuracy on Essay Revisions, *Journal of Second Language Writing*, Vol. 7, No. 1, pp. 43–68, 1998.
- [55] Sakai, T., Hirokawa, S., Feature Words that Classify Problem Sentence in Scientific Article, *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, pp. 360–367, 2012.

- [56] Schmid, H., Probabilistic Part-of-speech Tagging Using Decision Trees, Proceedings of the International Conference on New Methods in Language Processing, Vol. 12, pp. 44–49, 1994.
- [57] Schwarm, S.E., Ostendorf, M., Reading Level Assessment Using Support Vector Machines and Statistical Language Models, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 523–530, 2005.
- [58] Sugiura, M., Narita, M., Ishida, T., Sakaue, T., Murao, R., Muraki, K., A Discriminant Analysis of Non-native Speakers and Native Speakers of English, Proceedings of the Corpus Linguistics Conference (CL2007), 2007.
- [59] Supnithi, T., Uchimoto, K., Saiga, T., Izumi, E., Virach, S., Isahara, H., Automatic Proficiency Level Checking Based on SST Corpus, Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP), pp. 29–33, 2003.
- [60] Swan, M., Smith, B., Learner English: A Teacher’s Guide to Interference and Other Problems, Cambridge University Press, 2nd edition, 2001.
- [61] Tanimoto, T., Ohta, M., Examination of English Error Detection Using the Number of Search Results, Proceedings of the 4th Forum on Data Engineering and Information Management (DEIM2012), 9.1, 2012. (in Japanese)
- [62] Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., Understanding and Sharing Intentions: The Origins of Cultural Cognition, Behavioral and Brain Sciences, Vol. 28, No. 5, pp. 675–691, 2005.
- [63] Tetreault, J.R., Chodorow, M., The Ups and Downs of Preposition Error Detection in ESL Writing, Proceedings of the 22nd International Conference on Computational Linguistics, Vol. 1, pp. 865–872, 2008.
- [64] Tetreault, J., Blanchard, D., Cahill, A., A Report on the First Native Language Identification Shared Task, Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 48–57, 2013.
- [65] Tetreault, J., Leacock, C., Automated Grammatical Error Correction for Language Learners, International Conference on Computational Linguistics (COLING2014), pp. 8–10, 2014.

- [66] Tomokiyo, L.M., Jones, R., You're not from 'round here, are you?: Naive Bayes Detection of Non-native Utterance Text, Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, pp. 1–8, 2001.
- [67] Tono, Y., Mochizuki, H., Toward Automatic Error Identification in Learner Corpora: A DP Matching Approach, Proceedings of the Corpus Linguistics Conference, 2009.
- [68] Tono, Y., Automatic Extraction of L2 Criterial Lexico-grammatical Features Across Pseudo-longitudinal Learner Corpora: Using Edit Distance and Variability-based Neighbour Clustering, L2 Vocabulary Acquisition, Knowledge and Use, pp. 149–176, 2013.
- [69] Tono, Y. (Ed.), The CEFR-J Handbook: A Resource Book for Using CAN-DO Descriptors for English language Teaching, Taishukan Publishing, 2013. (in Japanese)
- [70] Tsur, O., Rappoport, A., Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words, Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA '07), pp. 9–16, 2007.
- [71] Weltig, M.S., Effects of Language Errors and Importance Attributed to Language on Language and Rhetorical-level Essay Scoring, Spaan Fellow Working Papers in Second or Foreign Language Assessment, Vol. 2, pp. 52–79, 2004.
- [72] Xue, L., Orgun, M., Zhang, K., A User-centred Consistency Model in Real-time Collaborative Editing Systems, Proceedings of the International Workshop on Distributed Communities on the Web, pp. 138–150, 2002.