Technical Report

UCAM-CL-TR-842
ISSN 1476-2986

Number 842

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Automated assessment of English-learner writing

Helen Yannakoudakis

October 2013

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted December 2012 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Wolfson College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

For my dad, who has been my role model, influence and inspiration...

Abstract

In this thesis, we investigate automated assessment (AA) systems of free text that automatically analyse and score the quality of writing of learners of English as a second (or other) language. Previous research has employed techniques that measure, in addition to writing competence, the semantic relevance of a text written in response to a given prompt. We argue that an approach which does not rely on task-dependent components or data, and directly assesses learner English, can produce results as good as prompt-specific models. Furthermore, it has the advantage that it may not require re-training or tuning for new prompts or assessment tasks. We evaluate the performance of our models against human scores, manually annotated in the Cambridge Learner Corpus, a subset of which we have released in the public domain to facilitate further research on the task.

We address AA as a supervised discriminative machine learning problem, investigate methods for assessing different aspects of writing prose, examine their generalisation to different corpora, and present state-of-the-art models. We focus on scoring general linguistic competence and discourse coherence and cohesion, and report experiments on detailed analysis of appropriate techniques and feature types derived automatically from generic text processing tools, on their relative importance and contribution to performance, and on comparison with different discriminative models, whilst also experimentally motivating novel feature types for the task. Using outlier texts, we examine and address validity issues of AA systems and, more specifically, their robustness to subversion by writers who understand something of their workings. Finally, we present a user interface that visualises and uncovers the 'marking criteria' represented in AA models, that is, textual features identified as highly predictive of a learner's level of attainment. We demonstrate how the tool can support their linguistic interpretation and enhance hypothesis formation about learner grammars, in addition to informing the development of AA systems and further improving their performance.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Ted Briscoe, for giving me the opportunity to pursue a doctoral degree in Cambridge. His valuable guidance, insightful discussions, unwavering support and encouragement have been of paramount importance for the realisation of this research. I would also like to thank my colleagues at the University of Cambridge, as well as conference attendees for their valuable feedback and suggestions. Special acknowledgement goes to Marek Rei, Øistein Andersen, Ekaterina Kochmar, Ben Medlock, Paula Buttery, Stephen Pulman, Simone Teufel and Anna Korhonen for offering their expertise and advice. My warmest thanks to Marek and Øistein for reading this thesis and for their constructive comments, technical guidance and help from the beginning of this doctorate to the end. It has been a great pleasure to collaborate with Dora Alexopoulou for parts of this work, who has offered the benefit of her expertise and helped improve its quality. I would also like to acknowledge Angeliki Salamoura, Fiona Barker, Laura Cope and Rebecca Stevens who voluntarily participated in our user studies, Nikiforos Karamanis for advice on these experiments and Tim Parish for making our system available on the web, as well as Cambridge Assessment and Cambridge University Press for permission to distribute data and use their tools respectively. This research was supported by the University of Cambridge ESOL Examinations, for which I am very grateful. Many thanks to the Computer Laboratory, Wolfson college, and my tutors, Dr. Martin Wolf and Dr. Martin Vestergaard, for supporting my conference attendance.

I would also like to thank my friends for always being there for me, even when thousands of miles set us apart. Special thanks goes to Marek, for your faith in me and for always being positive and supportive in so many different ways. Profound thanks are due to my family, for your love and support, and everything you have done for me. Without you, none of this would have been possible.

Contents

CHAPTER 1

Introduction

1.1 Language acquisition

Language acquisition is the process by which humans acquire a language, giving them the ability to perceive, comprehend, and employ a complex communication system. The term usually refers to the cognitive mechanism of acquiring a first language, otherwise known as native language, primary language, or *L1* – the first language a child learns. Over the past years, language acquisition has received considerable attention and is perhaps one of the more controversial topics in cognitive science. Language researchers, linguists, psycholinguists, cognitive scientists and others have studied first-language acquisition and developed a range of theories regarding the principles behind *how* it is acquired. The spectrum of language acquisition theories is marked by two conflicting positions at opposite ends: the *Nativism* one, which supports the existence of some innate language acquisition device which is biologically determined (Chomsky, 1965), and the *Empiricism* one, which believes that language is acquired through observation and learning from examples. In either case, successful language acquisition involves acquiring diverse capacities, including phonology, syntax, morphology, semantics, pragmatics, and so on.

The process of learning languages in addition to the native one is referred to as second-language acquisition, otherwise known as *L2* acquisition. The research on how humans acquire a second language is relatively young, having emerged around the second half of the twentieth century (Ellis, 1997). There are distinct differences between L2 and L1 acquisition. Research suggests that there is a 'critical period' in childhood during which a child should be exposed to their first language to achieve mastery (Lenneberg, 1967). On the other hand, second-language learning occurs when the L1 has been established, and may start during childhood or adulthood. Native-language acquisition involves learning through sufficient exposure to an L1-speaking environment, without explicitly being taught, and requires no conscious effort or control. Conversely, second-language learning may involve systematic learning strategies and intentional attempts to develop linguistic competence, such as cognitive strategies (for example, memorisation) or ones related to social activity and communication (O'Malley and Chamot, 1990). We should, however, note that literacy is acquired intentionally for both L1 and L2 learners.

Another divergence between the two is that it is hard for a second-language learner to achieve a proficiency level that approaches the native one, though age at which learning begins does play an important role: 'Early' bilinguals, people who learn a second language

early in life, are governed by the ability to acquire language rapidly and effortlessly, and therefore may reach a higher proficiency in that language compared to 'late' bilinguals, for whom language learning (after puberty) may involve a more variable, slow and laborious process (Johnson and Newport, 1989).

Furthermore, a learner's L2 may be influenced by their L1 (L1 transfer effects), and native-language structures may be used when a learner has not yet acquired enough of the second language (Krashen, 1982).

1.2 Language assessment

Estimates vary drastically as to the total number of spoken languages to date, something that is attributed to the difficulty of distinguishing between dialects and languages, and lies between five and eight thousand (Evans and Levinson, 2009).[1] A language *family* consists of a set of languages that are related by descent from a common ancestor. Nichols (1992) identifies around four hundred different families, including isolates. English, a language belonging to the Indo-European family and the Germanic branch, is one of the most widely used.

English is a common language in international commerce, science and technology, while many English-speaking countries and universities are the target of prospective employees and students from around the world. English is used as a *lingua franca*, and English proficiency is an essential skill for today's international employment market. It is therefore important, in an increasingly globalised environment, to be able to demonstrate one's English-language skills via objectively assessed qualifications.

Language assessment provides the means to identifying and measuring an individual's language skills, abilities, and proficiency level. There is a wide range of assessments available varying in format, rigour, and requirements, administered on paper or on computer. Questioning is one of the most common assessment instruments and may employ a number of assessment strategies; for example, certain types of questions require a specific predetermined answer, such as multiple-choice questions, true-or-false questions, fill-in-the-blank questions and constructed short responses. Others may focus on extended written responses, including prompts eliciting free-text answers, such as essays and reports. Each is designed to reflect various learning targets; these may be low-order cognitive skills, such as memorisation, or high-order ones, such as reasoning, organising ideas, synthesis and argument skills, and analytical thinking.

Assessment instruments, used in combination with standardised measurements of varying performance levels, provide strong evidence of someone's language abilities. Grades and scores are primary measurements adopted and used for purposes such as certification and self-assessment. They are assigned on the basis of specific marking criteria that serve as templates for assessment, devised to describe analytically key features in one's abilities.

1.3 Automated assessment

Automated assessment focuses on automatically analysing and assessing someone's competence. The field of automated assessment can be traced back to the early 1960s and

[1] To date, Ethnologue identifies 6,909 living languages: http://www.ethnologue.com/

emerged as a means to overcome issues arising with standardised assessment. For example, it supports a faster assessment and distribution of results, an advantage for several reasons, such as instant feedback, not only at the level of an individual, but also to institutions wishing to address educational shortfalls promptly. Further advantages become more pronounced when it comes to marking extended texts, a task prone to an element of subjectivity. Automated systems guarantee the application of constant marking criteria, thus reducing inconsistency, which may arise in particular when more than one human examiner is employed. Often, implementations include more detailed feedback on the writers' writing abilities, thus facilitating self-assessment and self-tutoring. Moreover, the potential of a reduced workload is becoming more attractive, especially in large-scale assessments. Standardised assessment entails an expensive and major logistical effort; automated assessment has the potential to drastically reduce time and costs for training and employing human scorers.

Although it is fairly easy to construct a model that assesses closed-class types of questions quite accurately, automated text assessment faces many challenges. One of the most important considerations is the possibility of building a system that emulates human behaviour in reading and making value judgements about someone's writing. This is largely dictated by the ability to evaluate not only vocabulary, grammar and syntax, but also various other aspects; different writing genres – such as essays, stories, letters, poetry, fiction, and so on – as well as cognitive aspects – such as language maturity, intellectual content, the logic behind an argument, discourse structure, clarity and fluency – are only a small part of the spectrum that needs to be considered. Additionally, it is equally important to be able to identify and automatically extract from texts measures of writing quality that are also a true reflection of the intrinsic qualities that form the basis of human judgements. The methodology and assessment criteria adopted by such systems should be transparent, understandable and meaningful. As the practical utility of automated systems depends strongly on their robustness to subversion, threats to their validity should also be identified and addressed. For example, writers who understand something of a system's workings may attempt to exploit this to maximise their scores, independently of their underlying ability. Several other challenges arise, such as their further development to function as learning tools, giving feedback on someone's writing skills and progress in similar ways and as usefully as humans typically do.

In this thesis, we will investigate automated assessment systems of free text that automatically analyse and score the quality of writing of L2 English learners. Automated text assessment systems exploit textual features chosen in an attempt to balance evidence of writing competence against evidence of performance errors in order to measure the overall quality and assign a score to a text. The earliest systems used superficial features, such as word and sentence length, as proxies for understanding the text. More recent systems have used more sophisticated automated text processing techniques to measure grammaticality, textual coherence, prespecified errors, and so forth. In the next section, we provide an overview of this thesis, followed by our research goals.

1.4 Overview

This thesis focuses on two main research directions. The first one aims at building robust state-of-the-art automated assessment models of English-learner text, while the second one investigates and analyses their internal characteristics. In the following sections, we

provide an overview of each one of them.

1.4.1 Automated assessment of learner texts

Implicitly or explicitly, previous work has mostly treated automated assessment as a supervised text classification task, that is, predicting a label for a text that is representative of its quality (e.g., a grade), based on a set of examples labelled with the classes or grades the system is trying to predict. Different techniques have been used, for instance cosine similarity of vectors representing text in various ways (Attali and Burstein, 2006), often combined with dimensionality reduction techniques such as Latent Semantic Analysis (LSA) (Landauer et al., 2003), generative and discriminative machine learning models (Briscoe et al., 2010; Rudner and Liang, 2002), domain-specific feature extraction (Attali and Burstein, 2006), and modified syntactic parsers (Lonsdale and Strong-Krause, 2003), all of which will be discussed in detail in Chapter 2.

We approach automated assessment as a supervised discriminative machine learning problem, which enables us to take advantage of annotated data. Our work investigates methods for assessing different aspects of writing prose, looks into the importance of a variety of writing quality features, and addresses validity issues related to their deployment. Further, we identify new techniques that outperform previously developed ones, and address generalisation issues.

Techniques such as LSA can be used to measure, in addition to writing competence, the semantic relevance of a text written in response to a given prompt. In contrast to previous work, we argue that an approach which does not rely on (manually developed) task-dependent components or data, and directly assesses learner English, can produce results as good as prompt-specific models. Further, it has the additional advantage that it may not require re-training or tuning for new prompts or assessment tasks. Systems that measure English competence directly are easier and faster to deploy, since they are more likely to be re-usable and generalise better across different genres compared to topic-specific ones; the latter becomes a pressing issue when attempting new tasks, since the model cannot be applied until a substantial amount of manually annotated response texts are collected for a specific prompt. A generic approach has the advantage of requiring smaller sample sizes, while its formulation represents truly consistent 'marking criteria' regardless of the prompt delivered. We should, however, note that human scoring rubrics also play an important role in the development of automated systems; this will be discussed in more detail in the next chapter.

1.4.2 The English Profile Programme

The Common European Framework of Reference for Languages (CEFR)[2] is an international benchmark of language attainment at different stages of learning (Council of Europe, 2001). The CEFR proposes the following six language proficiency levels:

A. Basic

   A1. Breakthrough (beginner)

   A2. Waystage (elementary)

[2]http://www.coe.int/t/dg4/linguistic/cadre_en.asp

B. Intermediate

B1. Threshold (intermediate)

B2. Vantage (upper intermediate)

C. Advanced

C1. Effective operational proficiency (advanced)

C2. Mastery (proficient)

A large number of Reference Level Descriptions (RLDs) have been devised to distinguish between the different levels and describe various functions that L2 learners can perform as they gradually master a language. For example, at level B2, a learner "can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options".

The English Profile (EP) research programme[3] aims to enhance the learning, teaching and assessment of English as an additional language by creating more detailed RLDs of the English-language abilities expected at each level (Saville and Hawkey, 2010). More specifically, the EP objective is to establish a set of English RLDs covering all CEFR levels and being indicative of L2 English proficiency at each level in terms of lexis, grammar, syntax, discourse, phonology, and so on, as well as address issues such as the extent to which they may vary depending on one's L1.[4]

Common methodologies in second language acquisition (SLA) research involve theory-driven approaches for formulating hypotheses on learner grammars, which are typically based on linguistic intuition and the extant literature on learner English. On the one hand, theory-driven approaches allow us to identify learner-language properties that are well understood and can inform learning theory. For example, learners whose native language lacks an article system (such as Russian, Turkish, Chinese, and Japanese) find articles challenging and tend to omit them in English, in contrast to learners from L1s with articles (such as Greek and German) whose mistakes in this area tend to be subtler and correspond to differences in article use between English and their L1. On the other hand, however, theory-driven methodologies may emphasise self-evident hypotheses and overlook properties about learner grammars that may not have been discussed in the linguistic literature.

As part of our research within the EP framework, we exploit automated assessment systems to support a novel and more empirical perspective on CEFR levels. Automated assessment models identify explicit cues in texts that can be highly predictive of specific attainment levels, grades or scores. Using visualisation techniques, we shed light on automated assessment models' internal characteristics and inspect the features they yield as the most predictive of a learner's level of attainment. We argue that investigation of those features provides an alternative route to learner grammars and offers insights into assessment and into the linguistic properties characterising each CEFR level. Machine learning, and data-driven techniques in general, are quantitatively very powerful, and thus allow us to explore a much wider hypothesis space. Effective exploitation enables us

[3] http://www.englishprofile.org/

[4] For more details on this research programme the reader is referred to Hawkins and Buttery (2009, 2010); Hawkins and Filipović (2012); Saville and Hawkey (2010).

to partially automate the process of hypothesis formation; thus, they can also serve as a useful adjunct to theory-driven approaches.

We demonstrate how effective inspection of those features can enhance hypothesis formation on developmental aspects of learner grammars – an important aspect of automated models that, to the best of our knowledge, has not been previously investigated. Finally, preliminary experiments also demonstrate that machine learning, used in tandem with visualisation techniques, effectively contributes towards identifying patterns that can help us further inform development of automated assessment systems and improve their performance.

1.5 Research goals

We investigate a number of key research directions related to systems that automatically assess (L2) writing quality. More specifically, the goals of this thesis are to:

1. Replicate, develop and extend automated assessment models that directly measure linguistic competence, and, more specifically, focus on lexical and grammatical properties, errors committed, and language complexity.

2. Replicate, develop and extend models that measure higher-order language skills, and, more specifically, discourse coherence and cohesion.

   (a) Perform the first systematic analysis of several methods for assessing discourse coherence and cohesion in learner texts.

   (b) Identify techniques suitable for assessing coherence in the noisy domain of learner texts and improve on previously-developed ones.

3. Develop generic models that produce competitive results without relying on prompt-specific data or components.

4. Identify good predictors of text quality and examine their relative importance and contribution to performance through ablation studies and significance tests.

5. Investigate performance of different machine learning algorithms on the task.

6. Examine and address validity issues of automated assessment systems, and, more specifically, their robustness to subversion by writers who understand something of their workings. Surprisingly, there is very little published data on the robustness of existing systems, although this is critical for their deployment.

7. Examine model generalisation to different learner corpora.

8. Develop and extend visualisation techniques to uncover the 'marking criteria' represented in automated assessment models. More specifically:

   (a) Build a tool that visualises textual features identified by automated assessment models as highly predictive of a learner's level of attainment.

   (b) Exploit visualisation to further improve performance of automated models.

(c) Demonstrate how the tool can support linguistic interpretation of those highly-predictive features and enhance hypothesis formation on learner grammars.

i. Evaluate the usability of the tool via user studies, the primary goals being to assess the ease with which information can be exploited by the target population, such as SLA researchers, teachers and assessors, and inform future development.

9. Anonymise and release a corpus of texts produced by learners of English as a second (or other) language, suitable for addressing automated assessment as a supervised machine learning task. Our principle aim is to facilitate further research on the task, in particular by making it possible to compare different systems directly.

1.6 Thesis structure

The content structure of the thesis is as follows: Chapter 2 begins with an overview of corpora consisting of text produced by L2 learners of English, highlights the publically available ones, and presents the texts used throughout our experiments and released in the public domain. It continues with an introduction to a variety of machine learning techniques that can or have been applied to automated assessment, discusses a number of the more influential and/or better described approaches to the automated assessment task, and compares and contrasts previous work to our own. Additionally, it gives an overview on common evaluation strategies used for automated assessment systems, including those we employ in our study. It concludes with an introduction to visualisation and a discussion of its proposed application to automated assessment models, presents visualisation research related to natural language, and points out our contributions. Further, it gives an overview of evaluation practices for visual presentations, and generally, of computer-based interfaces.

Chapters 3 and 4 tackle the automated writing assessment task from two different perspectives; general linguistic competence and discourse coherence and cohesion, and present state-of-the-art models and results, as well as address generalisability and validity issues. Chapter 5 describes a visual user interface developed to support linguistic interpretation of model-derived textual features, demonstrates its usefulness through a case study, and evaluates its usability. Additionally, preliminary results illustrate how the tool can be used to further improve performance of automated assessment systems. Finally, Chapter 6 assesses the contributions of this thesis and highlights avenues for future research.

CHAPTER 2

Background and literature review

2.1 Learner corpora

In the context of modern linguistics, McEnery and Wilson (2001) provide a 'prototypical' definition of a 'corpus' as a collection of texts that conforms to four main criteria:

1. Sampling and representativeness
2. Finite size
3. Machine-readable form
4. A standard reference

The first criterion refers to the compilation of an unbiased corpus that is representative of the population (and its range of variability) under examination. As McEnery and Wilson, p. 30, note, "We would not, for example, want to use only the novels of Charles Dickens or Charlotte Brontë as a basis for analysing the written English language of the mid-nineteenth century". The second criterion, 'finite size', is closely related to 'representativeness' and refers to collections that do not change continuously, in contrast to open-ended ones (otherwise known as 'monitor' corpora). Electronic corpora bear considerable advantages compared to other formats in that they allow for various levels of (linguistic) annotation and a wide range of analyses using (semi-)automatic techniques. As Leech (1993, p. 275) states,

> Corpus annotation is the practice of adding interpretative (especially linguistic) information to an existing corpus of spoken and/or written language, by some kind of coding attached to, or interspersed with, the electronic representation of the language material itself.

Finally, a corpus should serve as a 'standard reference' for the population it represents and be accessible to a wide community to facilitate comparisons between various studies.

Granger (2003a, p. 465) defines learner corpora as follows:

> Learner corpora, also called inter-language (IL) or L2 corpora, are electronic collections of authentic foreign or second language data.[1]

[1]For discussions on the definition of a learner corpus, see Granger (2003a), Nesselhauf (2004) and Schiftner (2008), among others.

In recent years, various L2 corpora have emerged and are becoming an increasingly important empirical resource in Applied Linguistics (Granger, 1994, 2003a). Together with L1 counterparts, they are widely established indispensable collections used to inform SLA research, language assessment, language pedagogy, lexicography, and so on. In addition, learner corpora can be diagnostic of learner language properties and give insights into the difficulties learners typically face. Their systematic design and compilation may provide a valuable longitudinal source of language development, as well as allow for analyses of various functions L2 learners can perform and the extent to which they vary depending on their L1, age, English proficiency level, the task setting, and so on (Granger, 2009; Hawkins and Buttery, 2010; Nesselhauf, 2004). As Granger (2002, p. 9) emphasises, "The usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables".

English is the language for which most learner corpora have been designed. Two of the earliest ones compiled are the International Corpus of Learner English (ICLE) (Granger, 2003b) and the Longman Learners' Corpus (LLC) (see Gillard and Gadsby, 1998). The former consists of argumentative and literary essays produced by learners whose proficiency level lies between upper intermediate and advanced. Its second version, released in 2009, contains more than three million words and around six thousand texts written by learners from sixteen different L1 backgrounds.[2] LLC comprises around ten million words, representative of various first languages, from examinations scripts and essays produced by English learners. Two of the largest learner corpora to date are the Hong Kong University of Science and Technology (HKUST) corpus of learner English (see Milton and Chowdhury, 1994) and the Cambridge Learner Corpus (CLC) (see Nicholls, 2003). HKUST is the largest collection of Chinese learner English, containing around thirty million words of assignments and exam scripts. CLC, a database of around fifty million words of written English, comprises texts produced by over 200,000 learners at various levels, from over 200 different countries and 140 different L1s. The texts include extended responses to various tasks, elicited by learners sitting English for Speakers of Other Languages (ESOL) examinations (see next section for more details).

A recent review (Schiftner, 2008) identifies a list of 26 different corpora of learner English; the majority contains untimed written productions of 300–500 words, while the most common L1s are Chinese and Japanese. Most of the learner corpora are not publically available or can only be accessed on-line; the latter limits their usability and the possibility for further exploitation.[3] ICLE is available for purchase, LLC is commercially available for research (Tono, 2003), HKUST may be used by researchers who are interested in collaboration (Pravec, 2002), and the Uppsala Student English (USE) corpus (Axelsson, 2000), containing essays written by Swedish students, is available for research and educational purposes. Additionally, the National University of Singapore Corpus of Learner English (NUCLE) (Dahlmeier and Ng, 2011), consisting of essays produced at NUS, is available for research purposes under a licence agreement, and the Lancaster Corpus of Academic Written English (LANCAWE) is freely available for use in research and teaching, though detailed information is rather limited.[4] Recently, Kaggle,[5] sponsored by

[2]https://www.uclouvain.be/en-277586.html

[3]A comprehensive list of learner corpora and their availability can be found in the Université catholique de Louvain website: http://www.uclouvain.be/en-cecl-lcworld.html

[4]Please note that publically available information regarding the corpora varies largely; the reader is advised to also contact the co-ordinators of the corpus of interest.

[5]http://www.kaggle.com/

the Hewlett Foundation, hosted the Automated Student Assessment Prize (ASAP) 2012 contest, aiming to demonstrate the capabilities of automated text scoring systems. The dataset released consists of around twenty thousand texts, produced by middle-school English-speaking students. Further details about learner corpora can be found in Nesi (2008); Nesselhauf (2004); Pravec (2002); Schiftner (2008); Tono (2003).

Learner corpora have been used in various natural language processing (NLP) applications; indicatively, L1 identification (Brooke and Hirst, 2012; Kochmar, 2011; Koppel et al., 2005; Swanson and Charniak, 2012; Wong and Dras, 2011), error detection and correction (Andersen, 2011; Boyd et al., 2012; De Felice and Pulman, 2008a,b; Kochmar et al., 2012; Nitin et al., 2012; Rozovskaya and Roth, 2011; Rozovskaya et al., 2012; Swanson and Yamangil, 2012; West et al., 2011), assessment of learner level and various writing dimensions (Attali and Burstein, 2006; Briscoe et al., 2010; Dickinson et al., 2012; Higgins et al., 2006; Landauer et al., 2003; Persing et al., 2010; Rudner and Liang, 2002), as well as in analysing learner speech (Barker et al., 2011; Chen and Yoon, 2011; Galaczi et al., 2011; Osborne, 2011; Peng et al., 2012; Yoon and Higgins, 2011).

Herein, we use the CLC to address automated assessment of English learner writing as a supervised machine learning problem; our reasons are multifold. A key characteristic of the CLC is that texts are assigned marks under a strict quality control mechanism, that is, examiners are monitored and their marking is reviewed and evaluated, while sometimes second marking is applied (see for example, Ffrench et al., 2012). Quality of the marks is one of the most important considerations in supervised machine learning. Moreover, the CLC is compiled so that it comprises a large multitudinal collection of texts elicited in response to various tasks by learners from diverse backgrounds and L1s; this is a critical component in building robust models, as a criterion of success is the ability to generalise well.

Several exams represented in the CLC are intended to demonstrate skills and knowledge relevant to language proficiency, rather than specific disciplines. The marking schemes for ESOL writing tasks typically emphasise the use of varied and effective language appropriate for the genre, exhibiting a range and complexity consonant with the level of attainment required. Thus, the marking criteria are not primarily prompt-specific but linguistic, which further supports the goals of this research (see Chapter 1, Section 1.5). This makes automated assessment for ESOL text a distinct subcase of the general problem of marking essays. The exam scores are also mapped onto CEFR levels, which supports the analysis of automated models in relation to benchmarks of language proficiency. Furthermore, the texts are manually annotated with the errors committed by the learners, which allows us to identify upper bounds of error detection systems incorporated in automated assessment models. Finally, an advanced database search tool has been developed for the CLC (Gram and Buttery, 2009); it provides the opportunity for a wide range of specialised searches, and it is an essential component of our system described in Chapter 5.

Unfortunately, the CLC is not publically available, and is used by authors and writers working for Cambridge University Press (CUP) and by members of staff at Cambridge Assessment (CA). The full potential of learner corpora, however, crucially depends on their availability (see Nesselhauf, 2004, for a detailed discussion). Although there are many published analyses of individual automated assessment systems that have been shown to correlate well with examiners' marks in many experimental contexts, no cross-system comparisons are available because of the lack of a shared dataset. As it is likely

that the deployment of such systems will further increase, standardised and independent evaluation methods are important. CA gave us permission to release a subset of the CLC (see Section 2.1.1.1 below), which we use in our experiments and hope will facilitate further research, not only in automated assessment, but in related application areas too. In the next sections, we provide more details about the CLC and the experimental datasets used throughout this thesis.

2.1.1 Cambridge Learner Corpus

The Cambridge Learner Corpus[6] (CLC), developed as a collaborative project between CUP and Cambridge ESOL, is a large collection of texts produced by English language learners from around the world, sitting CA's ESOL examinations.[7] The texts include extended responses to various tasks, which have been transcribed verbatim from candidates' handwritten answers. More than half have been manually annotated with information about the linguistic errors committed, using a taxonomy of approximately 80 error codes (Nicholls, 2003) (see Appendix A), devised by CUP, that specify the error type and (usually) its part-of-speech. Production examples are given below, exemplifying different error tags:

1. *In the morning, you are* <TV>*waken*|*woken*</TV> *up by a singing puppy.*
2. [...] *the people there* <AGV>*is*|*are*</AGV> *very kind and generous.*
3. *I will give you all* <MD>|*the*</MD> *information you need.*
4. [...] *which caused me* <FN><RN>*trouble*|*problem*</RN>|*problems*</FN>.

In the first sentence, TV denotes an incorrect verb tense error, where *waken* can be corrected to *woken*, whereas in the following there is a verb agreement error (AGV), where *is* is corrected to *are*. The third sentence contains a missing determiner error (MD), while the final one contains a nested error, where *trouble* is first corrected to *problem* through a replace noun error (RN) and subsequently replaced with *problems* to correct the wrong noun form (FN).[8]

The texts are further linked to meta-data about the learners and the exam, including native language, nationality, age, sex, level of English, reason for taking the exam, which examination was taken, exam date, question prompts and the candidate's grades and marks, including those for the other exam components, for example, reading, listening, and speaking.

There are three main examination types represented in the CLC that cover the CEFR levels from A2 to C2: the Main Suite, which consists of the Certificate of Proficiency in English (CPE), Certificate of Advanced English (CAE), First Certificate of English (FCE), Preliminary English Test (PET), and Key English Test (KET); the Business Suite, including the Business English Certificate (BEC) Higher, Vantage, and Preliminary; and the International English Language Teaching System (IELTS) (Williams, 2008). Throughout

[6] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB

[7] http://www.cambridgeesol.org/

[8] Further details on the CLC error taxonomies, as well as those on other corpora can be found in Andersen (2011); Díaz-Negrillo and Fernández-Domínguez (2006).

this thesis, we will use FCE and IELTS examination scripts, described in detail in the next sections.

2.1.1.1 FCE examination scripts

We begin our experiments using scripts produced by learners taking the FCE exam, which assesses English at an upper-intermediate level (CEFR level B2). The FCE writing component consists of two tasks eliciting free-text answers, asking learners to write either a letter, a report, an article, a composition or a short story, each between 120 and 180 words. Answers to each task are annotated with scores in the range between 1 and 20. In addition, an overall mark is assigned to both tasks (in the range 1–40), which is the one we use in our experiments (example FCE scripts can be found in Appendix B). As mentioned in Chapter 1, Section 1.4.1, we make no use of prompt information and do not make any attempt to check that the text answer is appropriate to the prompt. Our focus is on developing an accurate generic system for ESOL text that does not require prompt-specific or topic-specific training.

Our data consists of 1,141 scripts from the exam year 2000, produced by 1,141 distinct learners, and 103 scripts from the year 2001, written by 103 distinct learners. The age of the learners follows a bimodal distribution with peaks at approximately 16–20 and 26–30 years of age. The data contains sixteen different L1 backgrounds, the most frequent ones being Spanish and French. A typical prompt taken from the 2000 training dataset is shown below:

> *Your teacher has asked you to write a story for the school's English language magazine. The story must begin with the following words: "Unfortunately, Pat wasn't very good at keeping secrets".*

The FCE marking criteria are primarily based on the accurate use of a range of different linguistic constructions relevant to specific communicative goals (Williams, 2008). For this reason, we believe that an approach which directly measures linguistic competence will be better suited to ESOL text assessment and will have the additional advantage that it may not require re-training for new prompts or tasks.

Anonymisation and release

CA gave us permission to release in the public domain the FCE texts used in our experiments (1,244 scripts in total) (Yannakoudakis et al., 2011). The texts are not linked to their authors, but key meta-data for each candidate has been retained in the CLC (see previous section). We only make available the native language and age group of the candidate, including their grades. The prompts eliciting the free text are also provided with the dataset. Prior to publication, we manually anonymised the responses to remove personally identifying information. We identified personal names, locations (e.g., cities), organisations (e.g., universities), numbers (e.g., phone numbers), dates and birthdays and replaced them with linguistically similar and plausible entities which do not occur in the sample to maintain the original format of the texts. We then ran analyses of the part-of-speech (POS) tags assigned to tokens prior to and after anonymisation to examine possible divergences. We tagged the texts using the POS tagger in the Robust Accurate

Statistical Parsing (RASP) system (Briscoe et al., 2006),[9] which is based on the CLAWS tagset.[10] The results are presented in Tables 2.1 and 2.2:

original POS → new POS	count	original POS → new POS	count
NN1→NP1	27	JJ→RR	1
JJ→NP1	12	JJ→VVZ	1
NP1→NN1	11	MC→NNU	1
VV0→NN1	9	NN→NP1	1
NNL1→NP1	8	NN1→JB	1
NP1→JJ	8	NN1→NNL1	1
VV0→JJ	6	NN1→NNSB1	1
NP1→NN2	5	NN1→NNU	1
PPIO2→$	5	NN1→RR	1
NN2→NP1	4	NN1→VV0	1
NP1→VV0	4	NN2→JJ	1
NPM1→NP1	4	NNJ1→NP1	1
NP1→NNU	3	NNSB1→NP1	1
VVD→NP1	3	NNU→NP1	1
&FW→NP1	2	NP1→AT1	1
ICS→RR	2	NP1→NNL1	1
II→RP	2	NP1→NNSB1	1
NN1→JJ	2	NP1→PPHS1	1
NN1→NN2	2	NP1→RR	1
NP1→&FW	2	NP1→VVD	1
TO→II	2	NP1→VVN	1
VBZ→$	2	NP1→VVZ	1
VVN→VVD	2	NP1→ZZ1	1
&FO→NP1	1	PPIS1→ZZ1	1
&FW→UH	1	VBN→NP1	1
II→TO	1	VV0→NP1	1
JB→NN1	1	VVD→JJ	1
VVG→NP1	1	VVZ→NP1	1

Table 2.1: POS substitutions and counts after anonymising the FCE texts.

Total number of POS tags in the original documents	531,796
Total number of sentences	29,692
Number of words anonymised	1835
Number of POS tags that have changed (calculated over sentences where the number of words has not changed)	160
Number of sentences containing changed POS tags	142
Number of sentences where POS tags removed	3
Number of sentences where POS tags added	10

[9] http://ilexir.co.uk/applications/rasp/
[10] http://ucrel.lancs.ac.uk/claws/

Number of POS tags removed	3
Number of POS tags added	17

Table 2.2: Statistics describing the FCE texts after anonymisation.

Some of the added POS tags may be due to our correction of character encoding errors (e.g., ôSunshine 60ö changed to 'Sunshine 60'), as well as possible keyboard accidents. However, the discrepancies are infrequent enough to assume that the linguistic information in the data has been preserved.[11] We hope that the release of the dataset described here will facilitate further research and more informed system development; to date, it has been used in the Helping Our Own 2012 Shared task on detecting and correcting preposition and determiner errors, hosted by the Building Educational Applications Workshop at NAACL (Dale et al., 2012).

2.1.1.2 IELTS examination scripts

In the second half of the thesis, we describe experiments based on scripts produced by learners taking the IELTS Academic exam. Unlike FCE, IELTS is not a level-based test but is designed to stretch across a broader proficiency continuum, and its marking scale covers the whole range of CEFR levels. Similarly to FCE, candidates are asked to provide answers to two tasks, each with a minimum number of words varying between 150 and 250. According to their performance, learners are given a score (both per answer and per script) ranging from 0 to 9 on four different, equally weighted marking criteria: task achievement, coherence and cohesion, lexical resource, and grammatical range and accuracy. Scripts are then assigned an aggregate score based on the four detailed scores (an example IELTS script can be found in Appendix C).

Our data consists of 851 texts from the examination year 2008, and 100 from year 2010, and, again, we use the script-level scores; however, in these experiments, we focus on assessing higher-order language skills and, more specifically, discourse coherence and cohesion, which, in this exam, relates to the organisation of ideas. More specifically, in task 1 – in which candidates are presented with a graph, table, chart or diagram and asked to provide descriptions in their own words – writers are assessed based on their ability to describe and compare data or objects, sequences of events, and so on. In task 2, they are asked to write an essay, the quality of which is judged upon their ability to present a subject, argue, compare and contrast evidence, and so forth. Task 2 contributes to the overall score twice as much as the first one.[12]

2.2 Machine learning

In this section we give an introduction to machine learning, as well as to various techniques that have been applied to automated assessment. Mitchell (1997) defines machine learning as follows:

[11] At this point, we would like to thank Øistein E. Andersen and Ted Briscoe for their valuable help in anonymising the FCE texts.

[12] Further information on the tasks is available in the 'IELTS Information for candidates' document: http://www.ielts.org/pdf/Information_for_Candidates_2007.pdf

Definition

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

In our case, we require a computer program whose task is to learn how to automatically score (or assign a label to) a text, and whose performance, as measured by its ability to predict the correct score, is improved using its experience, gained from examples annotated with a score. The process of learning through experience E is called *training*. The type of training experience that uses examples annotated with scores, or, otherwise, with the target value we are trying to predict, is referred to as *supervised* learning, and the examples used as a *training set*. The training set is represented by a set of *features* and their target variables (scores in our case). Using as input the training set, the learning algorithm tries to learn the function, called *hypothesis*, that best describes the relationship between the features and the target values. The notation that we will use is defined as follows:

x: denotes the input variable or features.
y: denotes the output variable or target value.
$h(x)$: denotes the input–output mapping hypothesis function, or, in other words, the hypothesis function that uses the input to estimate the output.

Once the algorithm is trained, we can apply it to unseen *test* examples, whose target values are unknown to the model, and measure its performance by comparing the gold and the predicted scores.

In the course of this project, we approach automated assessment as a supervised machine learning task. In this section, we give a short introduction in a simplified way to various, primarily supervised, machine learning algorithms that have been applied (or their variations) to the task, and we discuss their advantages and disadvantages. Further details on these techniques can be found in Bishop (2006); Jurafsky and Martin (2009); Mitchell (1997); Ng (2012); Shawe-Taylor and Cristianini (2004).

2.2.1 Linear regression

Linear regression is a machine learning algorithm in which the target variable is real-valued. In its simplest form, univariate linear regression, the hypothesis function $h(x)$ can be represented as follows:

$$h(x) = \theta_0 x_0 + \theta_1 x_1 \tag{2.1}$$

or, more generally:

$$h(\mathbf{x}) = \boldsymbol{\theta}^{\mathrm{T}} \mathbf{x} \tag{2.2}$$

The above predicts a linear function, where the θ's represent the model parameters and essentially define the hypothesis. The parameters are chosen so that we get the best possible fit to the data, and thus predict y as accurately as possible. This is equivalent to solving the following optimisation problem:

$$\min_{\boldsymbol{\theta}} \mathrm{Cost}(\boldsymbol{\theta}) \tag{2.3}$$

$$\text{Cost}(\boldsymbol{\theta}) = \left(\sum_{\mathbf{x}} (h(\mathbf{x}) - y)^2 \right) + \lambda \|\boldsymbol{\theta}\|^2 \tag{2.4}$$

The above identifies the θ's that minimise the sum of the squared difference between the hypothesis and the target value, which is referred to as the *cost function*,[13] where the sum is taken over the training examples. The last term, $\lambda\|\boldsymbol{\theta}\|^2$, is the *regularisation term*, used to avoid overfitting the data and enhancing the model's generalisation to unseen examples, and λ is the parameter that controls the balance between underfitting and overfitting. We should, however, note that the use of a regularisation term becomes more important as the model's parameters increase. Various optimisation algorithms can be used to minimise the cost function. A popular approach is *gradient descent*[14] in which the parameters of the model are simultaneously updated until convergence, using the following formula:

$$\theta_i = \theta_i - \alpha \frac{\partial}{\partial \theta_i} \text{Cost}(\theta_i) \tag{2.5}$$

where α is the *learning rate*, and each θ_i gets updated based on the partial derivative of the cost function for linear regression (with respect to each θ_i).

2.2.2 Logistic regression

Logistic regression is an algorithm used for classification tasks, in which the focus is on assigning discrete target values or *classes*. The simplest form of logistic regression is binary classification, where y takes only two possible discrete values, in our case represented by 0 and 1. Then, the hypothesis $h(x)$ can be expressed by the following sigmoid function:

$$h(\mathbf{x}) = \frac{1}{1 + \exp\left(-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}\right)} \tag{2.6}$$

The hypothesis outputs the probability of an example belonging to a particular class, and, given a threshold, the most appropriate one is selected. More specifically:

$$h'(\mathbf{x}) = \begin{cases} 1 & \text{if } h(\mathbf{x}) \geq 0.5 \quad \text{or, equivalently} \quad \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} \geq 0 \\ 0 & \text{if } h(\mathbf{x}) < 0.5 \quad \text{or, equivalently} \quad \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} < 0 \end{cases} \tag{2.7}$$

A cost function that can be used to automatically choose the parameters θ of this model is represented below, again using a regularisation term:

$$\text{Cost}(\boldsymbol{\theta}) = -\sum_{\mathbf{x}} \left(y \log h(\mathbf{x}) + (1 - y) \log\left(1 - h(\mathbf{x})\right) \right) + \lambda \|\boldsymbol{\theta}\|^2 \tag{2.8}$$

Minimising the above minimises the classification error. Whenever the model's decision diverges from the correct one, the cost value increases, penalising for the wrong decision, otherwise, it is zero. Minimisation of the cost function can again be approached using gradient descent, applied to the case of logistic regression.

[13] Alternative cost functions can also be used; however, a discussion of this is beyond the scope of this thesis.

[14] Other faster advanced optimisation techniques can also be used, but they are also beyond our scope.

2.2.3 Support Vector Machines

The support vector machine (SVM) (Vapnik, 1995) is one of the most popular and powerful statistical learning algorithms and can be used for different learning tasks, including classification and regression. Again, in its standard form, it performs binary classification. An SVM is not a probabilistic model, in contrast to logistic regression, and the hypothesis function for classification can be defined as follows:

$$h(\mathbf{x}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} \geq 0 \\ 0 & \text{if } \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} < 0 \end{cases} \tag{2.9}$$

The optimisation objective can be defined as minimising (2.10), subject to specific constraints, (2.11):

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\theta}\|^2 \tag{2.10}$$

$$\text{subject to } \begin{cases} \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} \geq 1 & \text{if } y = 1 \\ \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} \leq -1 & \text{if } y = 0 \end{cases} \tag{2.11}$$

The constraints give SVMs one of their most important properties; they ensure that the model selects the hypothesis that has the largest distance from the closest data points, hence they are *large-margin* classifiers, while at the same time the algorithm minimises the classification error when applied to unseen data. This is a quadratic optimisation problem, and a popular solution is implemented in SVM$^{\text{light}}$ (Joachims, 1999) using decomposition algorithms.

In order to convert the model to a non-linear classifier, high-order features may be used, a case which can also apply to linear and logistic regression. However, a better way to approach this would be to make use of 'similarity' functions, called *kernels* (see Scholkopf and Smola, 2001 for more details). Such functions allow us to map the features to a higher-dimensional space and solve non-linear problems through linear optimisation techniques, and the choice of these functions essentially controls the hypothesis h.

2.2.4 Artificial Neural Networks

Artificial Neural Networks (ANNs), which try to mimic the human brain, are one of the oldest machine learning techniques (Pomerleau, 1989). ANNs consist of an input and output layer, as well as a number of hidden layers. Each layer consists of units, while the units between layers are interconnected. The perceptron is a type of an ANN, which, in its standard (linear) formulation, corresponds to a single hidden layer ANN with one unit that performs a linear transformation of the input and produces an output. More specifically, it is a classification algorithm, and the hypothesis function can be defined similarly to logistic regression and SVMs:

$$h(\mathbf{x}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} \geq 0 \\ 0 & \text{if } \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} < 0 \end{cases} \tag{2.12}$$

Learning the parameters of a perceptron that produce the target for the training set involves minimising the squared difference between the target and the predicted output, in a similar way to linear regression (though the hypotheses are defined differently), while gradient descent can again be used to solve the minimisation problem. Further variations

exist for building more complex models (in the case of non-linearly separable data), which involve multi-layer networks consisting of multiple hidden layers, as well as techniques such as *back-propagation*[15] and *forward-propagation* to learn the model parameters and make a prediction respectively.

2.2.5 Naive Bayes

Naive Bayes is a probabilistic classifier that uses Bayes' rule to assign a class to a text. The objective is to find the target that maximises the conditional probability:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|\mathbf{x}) \tag{2.13}$$

Using Bayes' rule, this can be formulated as follows:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, \frac{P(\mathbf{x}|y)\, P(y)}{P(\mathbf{x})} = \underset{y}{\operatorname{argmax}}\, P(\mathbf{x}|y)\, P(y) \tag{2.14}$$

The denominator is invariant to the target and thus can be ignored. Data sparsity issues make $P(\mathbf{x}|y)$ hard to estimate directly, so we can instead assume conditional independence between the features given a class (Naive Bayes hypothesis), an approach which tends to work well in practice:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left(\prod_{x} P(x|y) \right) P(y) \tag{2.15}$$

A simple way to learn the parameters of the model is to use *Maximum Likelihood Estimation* and calculate the frequencies based on the training data (usually in combination with smoothing techniques). The algorithm belongs to the family of models referred to as *generative*, in contrast to the ones presented above, known as *discriminative*. Discriminative classifiers directly learn a mapping of input–output variables, or directly model the posterior $P(y|\mathbf{x})$. On the other hand, generative ones learn a model of the joint probability $P(\mathbf{x}, y)$ and can make predictions using Bayes' rule to find $P(y|\mathbf{x})$ and pick the most likely target (Ng and Jordan, 2001).

2.2.6 Clustering

In contrast to the supervised techniques described previously, clustering is an unsupervised machine learning algorithm, that is, there is no target function and the training set thus does not contain any target variables. Clustering methods try to find a structure in the data, where similar objects are grouped together into *clusters*. In addition to the training set, most algorithms require the number of clusters to be given as input. A popular *centroid-based* clustering algorithm is K-means (Lloyd, 1982), in which clusters are represented by a central vector, called the cluster *centroid*, and essentially controls the assignment of objects to clusters. K-means can be defined as an iterative optimisation problem consisting of two steps repeated until convergence:

1. Assign each object to its nearest centroid.

[15] The idea of back-propagation was first invented by Arthur E. Bryson and Yu-Chi Ho in 1969.

2. Re-define the centroids by averaging the objects assigned to them.

Given the above, the optimisation objective can be defined as follows, which identifies the centroids $\mathbf{c} \in C$ that minimise the cost function, in our case their squared distance between the objects:

$$\min_{C} \text{Cost}(C) \tag{2.16}$$

$$\text{Cost}(C) = \sum_{\mathbf{c}} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}\|^2 \tag{2.17}$$

2.2.7 Discussion

The choice of machine learning algorithms largely depends on the task at hand, as well as on data availability. For instance, unsupervised methods, such as clustering, can be employed to discover structures in unlabelled training data. The amount of data also plays a key role, as experiments have shown various machine learning methods to converge in performance when very large corpora are used during training (e.g., Banko and Brill, 2001). However, some algorithms tend to be more computationally expensive (such as ANNs and SVMs) than others (e.g., Naive Bayes). On the other hand, existence of large amounts of (correctly) labelled data is often the exception rather than the rule, as manual annotation by experts can be expensive and time-consuming. With small amounts of data, some algorithms may be more prone to overfitting (e.g., SVMs) compared to others that may generalise better (e.g., Naive Bayes) (Ng and Jordan, 2001). However, when training with a sufficient amount of data, discriminative methods have been shown to generally outperform generative ones (Joachims, 1998), though this may also depend on the learning problem (Long and Servedio, 2007). Typically, generative models allow for more flexibility in modelling data dependencies, and they can more easily be framed to an unsupervised setting.

Finally, further criteria can dictate the suitability of different techniques, in addition to the data and application, such as data representation, and, more specifically, the size of the feature space. For instance, SVMs with kernels or ANNs may be better suited to small feature spaces and large samples, compared to logistic regression (Ng, 2012). Herein, we address automated assessment as a supervised discriminative machine learning problem. This is largely motivated by a previous study, closely following our methodology and reporting the superiority of discriminative methods to generative ones in ESOL texts (Briscoe et al., 2010). We discuss this in more detail in the following section and chapters.

2.3 Machine learning in automated assessment

There is an impressive body of literature with regards to the development, performance, usability and evaluation of automated text assessment and scoring systems (Attali et al., 2008; Burstein et al., 2003, 1998a,b; Callear et al., 2001; Coniam, 2009; Dickinson et al., 2012; Ericsson and Haswell, 2006; Higgins and Burstein, 2007; Higgins et al., 2006; Kakkonen et al., 2004; Kakkonen and Sutinen, 2008; Larkey, 1998; Leacock and Chodorow, 2003; Miller, 2003; Mitchell et al., 2002; Mohler et al., 2011; Phillips et al., 2007; Preston and Goodman, 2012; Pulman and Sukkarieh, 2005; Rosé et al., 2003; Rudner et al., 2006;

Shermis and Burstein, 2003; Sukkarieh et al., 2003; Tandalla, 2012; Williamson et al., 2012). Most recently, Shermis and Hammer (2012) report a comprehensive comparison of the capabilities of eight existing commercial essay scoring systems, evaluated as part of the ASAP contest organised by Kaggle.[16]

Extant approaches to automated assessment (hereafter AA) deploy a wide range of techniques from dimensionality reduction over matrices of terms through to extraction of linguistically deeper features such as types of syntactic constructions and specific error types (e.g., non-agreement of subject and main verb). In this section, we discuss a number of the more influential and/or better described approaches since the incipience of automated text assessment and give an overview of the various methodologies adopted; further systems will be compared and contrasted to our work in the following chapters. Detailed overviews of existing AA systems have been published in various studies (Dikli, 2006; Pérez-Marín et al., 2009; Shermis and Hammer, 2012; Valenti et al., 2003; Williamson, 2009).

Project Essay Grade (PEG) (Page, 1967, 1968) is one of the earliest systems, largely motivated by the potential to reduce labour-intensive marking activities. The system uses a number of manually identified mostly shallow textual features, which are considered to be proxies for intrinsic qualities of writing competence. Examples of such features include the essay length, number of pronouns and other POS tags, number of punctuation marks, the presence of a title, number of paragraphs, and so on. Linear regression is used to assign optimal feature weights to maximise the correlation with the examiner scores. The main issue with this system is that features such as word length and script length are easy to manipulate independently of genuine writing ability, potentially undermining the validity of the system (Kukich, 2000). Later versions were modified to include more sophisticated modules, such as ones incorporating the use of parsers (Page, 2003).

e-Rater (Attali and Burstein, 2006; Burstein, 2003), an automated essay scoring system developed by Educational Testing Service (ETS), was the first one to be deployed for operational scoring of high-stakes assessments in 1999. In e-Rater texts are represented using vectors of weighted features. Each feature corresponds to a different property of texts, such as an aspect of grammar (e.g., pronoun errors, missing words, subject–verb agreement), style (e.g., word repetition, passive voice, sentence length), mechanics (e.g., capitalisation of proper nouns, missing punctuation, spelling), organisation and discourse (e.g., number of discourse elements, subordinating clauses), semantic coherence and topic similarity (e.g., similarity between words in a text and those found in manually graded examples for each grade). Some features representing stereotypical grammatical errors, for example, are extracted using manually coded task-specific detectors based, in part, on typical marking criteria. An unmarked text is scored based on the cosine similarity between its weighted feature vector and the ones derived from the training set. Feature weights and/or scores can be fitted to a marking scheme by linear regression to produce a holistic score. However, the system contains some manually developed task-specific components and may require re-training or tuning for new prompts and assessment tasks.

Larkey (1998) and Rudner and Liang (2002) are among the first studies to explicitly model AA as a text classification task. The former demonstrates high results on five different datasets using Naive Bayes trained on vectors of stemmed words. Later, Rudner and Liang describe the Bayesian Essay Test Scoring sYstem (BETSY) (Coniam, 2009; Rudner and Liang, 2002), a system which is freely available for research purposes. BETSY

[16] http://www.kaggle.com/c/asap-aes

uses multinomial or Bernoulli Naive Bayes models to classify texts into different classes (e.g., pass and fail, or grades between A and F) based on content and style features such as word unigrams and bigrams, sentence length, number of verbs, noun–verb pairs, and so on. Classification decisions are based on the conditional probability of a class given a set of features, which is calculated under the assumption that each feature is independent of the others (see Section 2.2.5 above). Regression is used to optimise the fit between the classifier's confidence and the grade-point scales used. These systems show that treating AA as a text classification problem is viable; however, the feature types used are all fairly shallow, and the approach does not make efficient use of the training data, as a separate classifier is trained for each grade point.

Chen et al. (2010) propose an unsupervised clustering approach to AA of texts addressing the same topic, based on a voting algorithm. The underlying idea behind the algorithm is similar to e-Rater's hypothesis: good texts should resemble other good ones. Texts are clustered according to their grade and given an initial Z-score. A model is trained where the initial score of a text changes iteratively based on its similarity with the rest of the texts as well as their Z-scores. The approach might be better described as weakly supervised as the distribution of text grades in the training data is used to fit the final Z-scores to grades. The system uses a bag-of-words representation of text, which is prone to subversion and can potentially undermine its validity (more details on techniques that lead automatic systems astray are discussed in Chapter 3, Section 3.4). Nevertheless, exploration of the trade-offs between the degree of supervision required in training and grading accuracy is an important area for future research.

Recently, Briscoe et al. (2010) pointed out a paucity of studies investigating the application of discriminative machine learning to AA. Generative models often employ incorrect assumptions about the underlying properties of texts, for example, that the probability of a feature given a class is conditionally independent of the remaining features. Discriminative learning techniques make weaker assumptions, directly optimise performance on the training data, and often outperform non-discriminative ones in the context of text classification (Joachims, 1998). Briscoe et al. present a novel discriminative model, a variant of the batch perceptron algorithm (Bös and Opper, 1998) and report superior results compared to probabilistic classifiers, such as Naive Bayes and Maximum Entropy, as well as to dimensionality reduction techniques that have been successfully used in earlier AA studies (see next section). They experimentally show that their model, trained on CLC texts and employing a variety of lexical and grammatical features (e.g., POS ngrams and phrase-structure rules) performs very close to the upper bound as defined by the agreement between human examiners. Our research closely resembles their methodology and extends their work in relation to our research goals. Further details are discussed in the following chapters.

2.4 Other approaches to automated assessment

Intelligent Essay Assessor (IEA) (Landauer et al., 2003) uses Latent Semantic Analysis (LSA) (Landauer et al., 1998) to compute the semantic similarity between texts, at a specific grade point, and a test text. Contrary to other techniques, LSA can be construed as both a model of human knowledge representation and acquisition and as a method for capturing semantic content in texts (Landauer and Dumais, 1997; Landauer et al., 1997; Wolfe et al., 1998). In LSA, text is represented by a matrix, where rows correspond to

words and columns to context (texts). Singular Value Decomposition (SVD) is used to obtain a reduced dimension matrix clustering words and contexts (see Chapter 4, Section 4.2.2 for more details). The system is trained on topic and/or prompt specific texts while test texts are assigned a score based on the ones in the training set that are most similar. The overall score, which is calculated using regression techniques, is based on the content score as well as on other properties of texts, such as style, grammar, and so forth, though the methods used to assess these are not described in any detail in published work. However, the system requires re-training or tuning for new prompts and assessment tasks.

A rather different methodology is adopted by Lonsdale and Strong-Krause (2003), who use a modified syntactic parser to analyse and score texts. Their method is based on a modified version of the Link Grammar parser (Sleator and Templerley, 1995) where the overall score of a text is calculated as the average of the scores assigned to each sentence. Sentences are scored on a five-point scale based on the parser's cost metric, which roughly measures the complexity and deviation of a sentence from the parser's grammatical model. This approach bears some similarities to the representation of our feature space; however, grammatical features depict only one component of our overall system and of the task (see next chapter).

2.5 Evaluation strategies

The evaluation of automated assessment systems has been based on various criteria. Typically, it involves comparisons against human scoring and measurements of consistency. Human scores are used as the basis for optimising computational models of text quality; their association to system evaluation is thus an accepted indicator of the quality of the predicted scores. Traditional metrics include the correlation between predicted scores and human scores (e.g., Pearson's product–moment correlation coefficient) and the percentage of exact or adjacent agreement (e.g., agreement within one point), as well as kappa statistics (Cohen, 1960) (such as the quadratic-weighted kappa metric) that are designed to calculate the agreement between raters whilst at the same time excluding agreement by chance. Williamson (2009), in his thorough discussion on frameworks for evaluating and implementing automated scoring for high-stakes assessment, notes that percentage of agreement is scale-dependent, as, for example, higher performance may be observed by chance with a scale with few distinct points compared to one with more.

Throughout this thesis, we evaluate performance of our models against human scores using Pearson's product–moment correlation and Spearman's rank correlation coefficient. Human–human and human–machine correlation has been widely used in AA studies, and, at the same time, one of our principal aims is to facilitate comparison with previous research that closely resembles our own (Briscoe et al., 2010). However, we do recognise the inherent reliability problems related to human scoring, as well as the need to identify and evaluate against further criteria, including correlations with extrinsic metrics such as state assessment scores and course grades (Shermis and Hammer, 2012; Williamson, 2009).

2.6 Information visualisation

Recent advances in machine learning has led to self-contained out-of-the-box machine learning solutions that more often than not are viewed as 'black boxes', that is, they are primarily inspected in terms of their input and output, and their internal workings and characteristics are often ignored and not examined. This lack of knowledge may lead to difficulties in output interpretation, as well as result in undiscovered important patterns that could potentially be used to make the model more powerful and effective.

Generic approaches to AA have the advantage of modelling truly consistent 'marking criteria' regardless of the prompt delivered. AA models identify explicit cues in text that determine its quality and a learner's assessment. Visualisation techniques can help us shed light on AA 'black boxes' and let us inspect the features they yield as the most predictive of a learner's level of attainment. As Noah Iliinsky remarked during his talk at the European Bioinformatics Institute (2012), "visualisation makes data *accessible*". Given the quantitatively powerful nature of the models' internal characteristics, visualisation can help us gain a deeper understanding of important phenomena represented in large databases (Card et al., 1999).

We demonstrate how visual presentations of machine-learned features can point to a range of interesting patterns in learner data. More specifically, we integrate highly-weighted discriminative linguistic features into a graph-based visualiser to support SLA research. We present a coordinated approach, using search tools and graph visualisation combined with easy access to the underlying raw data, create an analysis scenario to demonstrate its usefulness, and evaluate its usability; the primary goal of the latter is to assess the ease with which information can be accessed by target users. This is the first attempt to visually analyse, as well as perform a linguistic interpretation of automatically-determined features that characterise learner English. Though several approaches have been proposed for linguistic visualisation (see below), our work differs in, and contributes towards, the following: using visualisation as a search tool for hypothesis generation. In addition, we illustrate how visualisation can facilitate the identification of new discriminative features that can further improve performance of automated assessment systems.

2.6.1 Visualisation approaches

The number of possible visualisation techniques that can be applied to different applications is big. However, there are several visualisation methods that have been well investigated and applied successfully to a wide variety of tasks. These methods include graphs, histograms, circle graphs, self-organising maps, hyperbolic trees, treemaps, fish-eye graphs and menus, scatterplots, as well as hybrid forms (Card et al., 1999; Feldman and Sanger, 2007; Heer et al., 2005). Card et al. (1999) is a useful resource on information visualisation research. Further, Noah Iliinsky (2012) provides valuable guidelines on best uses of visual encodings and their properties given the nature of the data (see Appendix D) and on the design of data visualisations (Iliinsky and Steele, 2011), as well as examines various case studies and their approaches to projects from a variety of perspectives (Steele and Iliinsky, 2010). Below, we briefly discuss several visualisation practices on different tasks.

In recent years, several studies have emerged that involve visualisation of natural language, and are perhaps more related to our research than others; for example, Lyding

et al. (2012) use *Structured Parallel Coordinates* (SPCs) (Culy et al., 2011) to visualise diachronic changes in academic discourse, in terms of lexicogrammatical features of registers. SPCs involve the use of axes, each representing different data dimensions, while data points represented on the axes are connected with lines to visualise their relationships. Van Ham et al. (2009) introduce *Phrase Net*, a system that analyses unstructured text by taking as input a predefined pattern and displaying a graph whose nodes are words and whose edges link the words that are found as matches. Users can interactively specify a pattern or choose from a list of default ones. Patterns can take the simple form of $X and Y$ or $X or Y$, or be defined using regular expressions. Another visualisation technique which is popular for representing information in texts is "word clouds". Viégas et al. (2009) investigate Wordle, a tool for making "word clouds", that is, graphic statements in which words are packed tightly and can be placed vertically, horizontally or diagonally. Further, the colour and size of the words can be used to represent different types of information, for example, frequent words may be given more prominence via using larger fonts.

Collins (2010) in his dissertation addresses different visualisation techniques for natural language processing (NLP) research. The *Bubble Sets* visualisation draws secondary set relations around arbitrary collections of items, such as a linguistic parse tree. *VisLink* provides a general platform within which multiple visualisations of language (e.g., a force-directed graph and a radial graph) can be connected, cross-queried and compared. Moreover, he explores the space of content analysis using *DocuBurst*, an interactive visualisation of document content, which spatially organises words using an expert-created ontology (e.g., WordNet). *Parallel Tag Clouds* combine keyword extraction and coordinated visualisations to provide comparative overviews across subsets of a faceted text corpus. Recently, Rohrdantz et al. (2011) proposed a new approach to detecting and investigating changes in word senses by visually modelling and plotting aggregated views about the diachronic development in word contexts.

Visualisation techniques have been successfully used to support humanities research (e.g., Plaisant et al., 2006 and Don et al., 2007), as well as genomics (e.g., Meyer et al., 2010a and Meyer et al., 2010b). For example, Don et al. (2007) have developed a system that visualises the distribution of frequent patterns found in text collections, displays their context and supports analysis of their correlations. Plaisant et al. (2006) have built a user interface which aids the interpretation of literary work by integrating text mining algorithms. Their system allows visual exploration of documents, preparation of training sets and reviewing of classification algorithm results. Meyer et al. (2010a) present a system that supports the inspection and curation of data sets showing gene expression over time, in conjunction with the spatial location of the cells where the genes are expressed.

Graph-based visualisations, which we adopt in our work, have been used effectively in various areas. As Herman et al. (2000) note, "The area of graph visualization has reached a level of maturity in which large applications and application frameworks are being developed. However, it is difficult to enumerate all the systems because of the sheer quantity". An overview on graph visualisation methods and systems is beyond the scope of this thesis. However, recent examples that are similar to ours, design-wise, include the analysis of domains such as social networks to allow for a systematic exploration of a variety of Social Network Analysis measures (SNA). Gao et al. (2009) present *MixVis* and Perer and Shneiderman (2006) *SocialAction*, two systems designed to help structural analysts examine social networks (e.g., a terrorism network). Both tools allow

systematic exploration of SNA measures[17] by linking together the statistical and visual components of a network. Heer and Boyd (2005) have implemented *Vizster*, a visualisation system for the exploration of on-line social networks (e.g., Facebook) designed to facilitate the discovery of people, promote awareness of community structure, and so on. `VisualComplexity.com` is a unified resource space of projects regarding a variety of graph/network visualisation methods across different domains (Lima, 2011). Examples include visualising the Bible, Wikipedia, computer systems, food webs, semantic networks, topic shifts, and so forth. Useful resources on more technical details on graph drawing, visualisation and layout algorithms include Battista et al. (1994, 1998); Eades and Sugiyama (1990); Gansner et al. (1993); Herman et al. (2000).

2.6.2 Evaluation

Evaluation of visual presentations, visualisation systems and, more generally, of computer-based interfaces is a key component to ensuring their quality and success. For example, poor system usability may lead to low user effectiveness, increased errors in completing tasks, and consequently low adoption rates. The foci of evaluation may relate to various development stages, such as evaluation of a prototype with respect to state-of-the-art techniques, or deployment-level evaluation in order to assess system effectiveness and usage as part of the users' real-world workflow. In addition, they may relate to the visualisation itself, or to assessment of a more holistic view of the user experience (Lam et al., 2011). There is a rich flora of evaluation methodologies available, varying in complexity and typically involving representative users, whose choice and settings largely depend on the evaluation goals and the underlying application context. Popular techniques include informal usability testing, formal studies and controlled experiments, longitudinal studies and large-scale log-based usability testing (Hearst, 2009, Ch. 2).

Informal usability testing is common during early stages of development and includes iterative stages during which a usually small number of target users are given successive prototypes with the goal to identify major problems or users' preferences, or to test candidate system-features and designs. Evaluation and revision based on user feedback in a cyclical fashion is typical until required characteristics are attained, and low-fidelity designs are transformed into high-fidelity ones. Early stages of design may also include heuristic evaluations (Mack and Nielsen, 1995; Nielsen, 1992; Tory and Möller, 2005; Zuk et al., 2006), where a set of predefined guidelines or heuristics form the basis for evaluation, or field studies, focused on observing and documenting usage or completion of evaluator-defined tasks as part of the users' everyday workflow, rather than being laboratory-based, and thus emphasising the element of realism. Additionally, observational studies may often be combined with interviews and (self-reporting) questionnaires.

Formal studies are typically artificially constrained in order to focus on key points of interest, are commonly conducted in a laboratory, and involve a large number of users. They are usually preceded by pilot testing to check the experimental design and/or by user training to increase system and experiment familiarity. Controlled experiments (Blandford et al., 2008; Keppel et al., 1992; Kohavi et al., 2007, 2009), a form of formal testing, are used to test hypotheses, while the focus is on quantitative analyses (Blandford et al.,

[17]Examples of SNA measures include those representing the *betweenness centrality* of a node, which refers to how frequently it appears on the shortest path between other nodes, having thus a control over the network flow (Freeman, 1979).

2008). They are commonly used methodologies for rigorously comparing and benchmarking novel techniques with existing state-of-the-art counterparts, otherwise known as head-to-head comparisons, as participants can perform identical tasks across different systems (Lam et al., 2011), and the tasks tend to be simple and specific. Objective evaluation measures include overall task completion time, errors made, number of keystrokes, number of correct answers per specific time intervals, or may involve experts to evaluate user results. The experimental design may be conducted between or within subjects, while special care should be taken to ensure minimisation of confounding variables, that is, variables that unintentionally vary between experimental conditions and can affect the results; for example, comparing user speed using two different systems on different hardware.

A common problem in formal studies is the order in which users are assigned to experimental conditions. Order effects can influence the users and bias the results. Popular techniques used to counterbalance these effects are the 'blocked design' and 'latin-squares design', which ensure a systematic approach to variation. For example, with two experimental conditions, $C1$ and $C2$, we can create $2! = 1 \times 2 = 2$ different orderings, '$C1$ $C2$' and '$C2$ $C1$', and randomly assign participants to each one of them. Last but not least, the majority of such studies is usually followed by questionnaire-based assessments to solicit user opinions and ratings, with the use of five- or seven-point Likert scales (Likert, 1932) being quite common.

Longitudinal studies are useful for revealing long-term usage and application patterns in everyday environments; observations of dozens of users over months or years contributes towards the reliability, validity, and generalisability of the results (Shneiderman and Plaisant, 2006). This study differs from the previous ones in that it goes beyond first-time user experience and examines participant behaviour as system familiarity increases. Shneiderman and Plaisant (2006) propose assessment of information visualisation tools through observation, interviews, surveys, automated logging of user activities and component frequency usage, difficulty in learning a tool and system-adoption rates, as well as success in achieving one's goals.

Large-scale log-based testing is another form of evaluation that is typically adopted in Web-based systems, whose application context has the advantage of allowing for a large number of users. New features and designs can be tested by recording user behaviour and comparing it to other versions, and these experiments can be followed by laboratory studies. In contrast to formal studies, users are not required to undertake specific tasks, and they are neither explicitly asked to opt-in to the study, nor is feedback explicitly elicited (Hearst, 2009).

Several resources provide valuable details and give guidance on the use of appropriate evaluation methodologies and practices (Blandford et al., 2008; Dumas and Redish, 1999; Hearst, 2009; Horsky et al., 2010; Käki and Aula, 2008; Kohavi et al., 2007, 2009; Lam et al., 2011; Mack and Nielsen, 1995; Munzner, 2009; Plaisant, 2004; Shneiderman and Plaisant, 2006; Tory and Möller, 2005).

CHAPTER 3

LINGUISTIC COMPETENCE

In this chapter, we demonstrate how supervised discriminative machine learning techniques can be used to automate the assessment of ESOL examination scripts. In particular, we report experiments on rank preference SVMs trained on FCE data, on detailed analysis of appropriate feature types derived automatically from generic text processing tools, and on comparison with different discriminative models. Experimental results on the publically available FCE dataset show that the system can achieve levels of performance close to the upper bound – as defined by the agreement between human examiners on the same corpora – for directly measuring linguistic competence. We report a consistent, comparable and replicable set of results based entirely on the FCE dataset and on public-domain tools and data, whilst also experimentally motivating some novel feature types for the automated assessment (AA) task, thus extending the work described in Briscoe et al. (2010). Finally, using a set of outlier texts, we test the validity of the model and identify cases where the model's scores diverge from that of a human examiner.

Work presented in this chapter was submitted and accepted as a full paper in the 49th meeting of the Association for Computational Linguistics: Human-Language Technologies (Yannakoudakis et al., 2011).

3.1 Extending a baseline model

As described in Chapter 2, Section 2.3, Briscoe et al. (2010) were the first to apply discriminative machine learning methods to the AA task, which often outperform non-discriminative ones in the context of text classification (Joachims, 1998). They present a novel variant of the batch perceptron algorithm (Bös and Opper, 1998), the Timed Aggregate Perceptron (TAP), that efficiently learns preference ranking models (see next section for details). They experimentally show that their model, employing a variety of (linguistic) features and trained on around 3,000 FCE ESOL texts, performs very close to the upper bound, as well as outperforms generative counterparts. Our contribution within this framework is fivefold:

1. We focus on reporting a replicable set of results based entirely on public domain tools and (training/test) data.

2. We motivate the use of novel feature types and extend their model.

3. We study the contribution of different feature types to the AA task.

4. We present a comparison of different machine learning models.

5. We test the validity of our best model on outlier texts.

3.1.1 Feature space

We report results on the publically available FCE dataset (see Chapter 2, Section 2.1.1.1) and, following Briscoe et al. (2010), we parse the training and test data using the Robust Accurate Statistical Parsing (RASP) system (Briscoe et al., 2006) with the standard tokenisation and sentence boundary detection modules in order to broaden the space of candidate features suitable for the task. RASP, an open-source system, includes an unlexicalised parser, which is expected to perform well in the noisy domain of learner text, where misspellings and grammatical errors are common, though this is evaluated implicitly through the usefulness of the features extracted from the parser's analyses. As in Briscoe et al. (2010), our focus is on developing an accurate AA system for ESOL text that does not require prompt-specific or topic-specific training. Although the FCE corpus of manually-marked texts was produced by learners of English in response to prompts eliciting free-text answers, the marking criteria are primarily based on the accurate use of a range of different linguistic constructions. For this reason, it is plausible to assume that an approach which directly measures linguistic competence will be better suited to ESOL text assessment, and will have the additional advantage that it may not require re-training or tuning for new prompts or assessment tasks.

We extract the features used by Briscoe et al. (2010), which are mainly motivated by the fact that lexical and grammatical properties should be highly discriminative for automatically assessing linguistic competence in learner writing, and replicate their model. Their full feature set is as follows:

1. Lexical ngrams
   (a) Word unigrams
   (b) Word bigrams
2. Part-of-speech (POS) ngrams
   (a) POS unigrams
   (b) POS bigrams
   (c) POS trigrams
3. Features representing syntax
   (a) Phrase structure (PS) rules
4. Other features
   (a) Script length
   (b) Error rate

Word unigrams and bigrams are lower-cased and used in their inflected forms. POS unigrams, bigrams and trigrams are extracted using the RASP tagger, which uses the CLAWS tagset. The most probable posterior tag per word is used to construct POS ngram features; however, given the large number of misspellings in learner data, we use the RASP parser's option to analyse words assigned multiple tags when the posterior probability of the highest ranked tag is less than 0.9, and the next n tags have probability greater than $\frac{1}{50}$ of it.

Based on the most likely parse for each identified sentence, the rule names from the phrase structure (PS) tree are extracted. RASP's rule names are semi-automatically generated and encode detailed information about the grammatical constructions found (e.g., 'V1/modal_bse/+-', a VP consisting of a modal auxiliary head followed by an (optional) adverbial phrase, followed by a VP headed by a verb with base inflection). Moreover, rule names explicitly represent information about peripheral or rare constructions (e.g., 'S/pp-ap_s-r', a S with preposed PP with adjectival complement, e.g., *for better or worse, he left*), as well as about fragmentary and likely extra-grammatical sequences (e.g., 'T/txt-frag', a text unit consisting of two or more subanalyses that cannot be combined using any rule in the grammar). Therefore, many (longer-distance) grammatical constructions and errors found in texts can be (implicitly) captured by this feature type.

Although FCE contains information about the linguistic errors committed (see Chapter 2, Section 2.1.1.1), Briscoe et al. (2010) try to estimate an error rate in a way that doesn't require manually tagged data. They build a trigram language model (LM) using ukWaC (ukWaC LM) (Ferraresi et al., 2008), a large corpus of English containing more than 2 billion tokens. A word trigram is counted as an error if it is not found in the language model. They compute presence/absence efficiently using a Bloom filter encoding of the language models (Bloom, 1970). However, they also use an error rate calculated from the FCE error tags to obtain an upper bound for the performance of an automated error estimator (true FCE error rate).

Feature instances of types 1 and 2 are weighted using $tf*idf$ and their vectors are normalised by the L2 norm, that is, the square root of the sum of squares. Feature type 3 is weighted using frequency counts, while 3 and 4 are scaled so that their final value has approximately the same order of magnitude as 1 and 2. The script length is based on the number of words and is mainly added to balance the effect the length of a script has on other features. Finally, features whose overall frequency is lower than four are discarded from the model.

In extending Briscoe et al.'s AA model, we hypothesise that features capturing the syntactic complexity of sentences should also be indicative of a learner's writing competence. More specifically, we investigate the impact of complexity measures representing the distance between a head and a dependent (in word tokens) in a grammatical relation (GR). GRs represent syntactic dependencies between constituents in a clause, and are automatically identified by RASP. An example is illustrated in Figure 3.1 using an FCE excerpt, which shows the different types of relations between words represented as lemmas and POS tags.[1] For example, 'ncsubj' represents binary relations between non-clausal subjects (NPs, PPs) and their verbal heads, as in *have*_VH0 *you*_PPY. The distance in word tokens between *have*_VH0 and *you*_PPY is 1, while the distance between *If*_CS and *have*_VH0 is 2. The direction of the relation, or equivalently, the position of

[1]The dependency graph was produced using the SemGraph tool: http://www.marekrei.com/projects/semgraph/

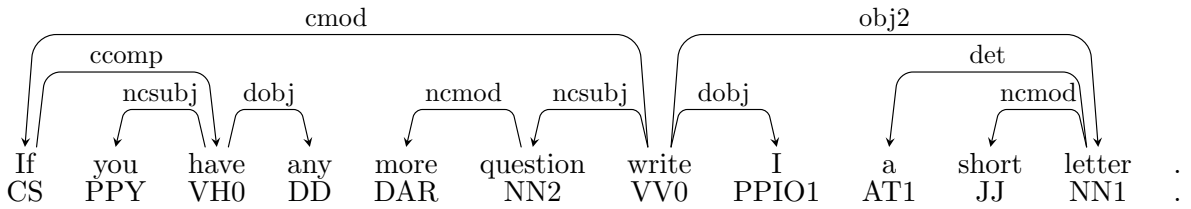


Figure 3.1: Example GR types and dependencies.

the head compared to the dependent distinguishes positive dependencies from negative ones. For example, *have*_VH0 and *you*_PPY have a positive dependency, while *have*_VH0 and *any*_DD a negative one (as the head precedes the dependent) (for more details see Briscoe, 2006).

We extract a number of complexity measures representing GR distance in various ways from RASP and explore their impact on performance. In particular, we experiment with 24 different numerical features, grouped for positive and negative dependencies and presented below:

1. GR-LONGEST-TOP-P/N: longest distance in word tokens between a head and dependent in a grammatical relation (GR) over the top ranked derivation for positive and negative dependencies (P/N) separately.

2. GR-TOTAL-TOP-P/N: sum of the distances between a head and dependent over the top ranked derivation for P/N dependencies separately.

3. GR-MEAN-TOP-P/N: the means for P/N dependencies calculated by dividing GR-TOTAL-TOP-P/N by the number of GRs in the set for the top parse only.

4. GR-LONGEST-NBEST-P/N: longest distance for P/N over the top 100 parses.[2]

5. GR-TOTAL-NBEST-P/N: sum of the distances for all GR sets over the top 100 parses for P/N separately.

6. GR-MEAN-NBEST-P/N: the means for P/N dependencies calculated by dividing GR-TOTAL-NBEST-P/N by the number of GRs in the top 100 parses.

7. NBEST-MED-GR-TOTAL-P/N: median for GR-TOTAL-NBEST-P/N (calculated over the top 100 parses).

8. NBEST-STD-GR-TOTAL-P/N: standard deviation for GR-TOTAL-NBEST-P/N.

9. NBEST-AVG-GR-TOTAL-P/N: average for GR-TOTAL-NBEST-P/N.

10. NBEST-MED-GR-LONGEST-P/N: median for GR-LONGEST-NBEST-P/N.

11. NBEST-STD-GR-LONGEST-P/N: standard deviation for GR-LONGEST-NBEST-P/N.

12. NBEST-AVG-GR-LONGEST-P/N: average for GR-LONGEST-NBEST-P/N.

[2]We chose the top 100 parses mostly for convenience, as various statistics are easily available from RASP for the top 100 derivations.

Intuitively these complexity measures capture aspects of the grammatical sophistication of the writer through the representation of the distance between heads and dependents in various forms (e.g., longest or mean distance), and we hypothesise they can be used to assess linguistic competence. However, they may also be confounded in cases where sentence boundaries are not identified, for example, due to poor punctuation. In the experiments presented here, we evaluate performance of individual measures as well as their combinations for the AA task. We identify a set of discriminative complexity measures and use their values as features in the document vectors. Although these features bear some similarities to Lonsdale and Strong-Krause (2003)'s method, who roughly measure the complexity and deviation of a sentence from the parser's grammatical model in order to assign a score to a text, this is the first study on the application of these complexity features on learner language assessment and their evaluation under a data-driven methodology.

Next, in order to get a better estimate of the error rate, we extend the ukWaC language model with trigrams extracted from FCE texts (ukWaC+FCE LM). As FCE contains texts produced by second language learners, we only extract frequently occurring trigrams from highly ranked scripts to avoid introducing erroneous ones to our language model. We hypothesise that by adapting the LM to the FCE vocabulary will further improve performance of the AA system, as it will allow us to calculate an error rate that will directly capture (correct) learner word-usage patterns.

3.2 Approach

Briscoe et al. (2010) present a novel discriminative model, TAP, a variant of the batch perceptron algorithm (Bös and Opper, 1998), and report superior results compared to generative models. The batch perceptron learning procedure updates the weight vector $\boldsymbol{\theta}$ for all misclassified samples simultaneously, as opposed to updating $\boldsymbol{\theta}$ for every instance. TAP, a wide margin algorithm, uses an *aggregate vector* containing the sum of all misclassified instances, and iteratively updates $\boldsymbol{\theta}$ in the direction of the (normalised) aggregate vector. The aggregate vector is normalised according to a *timing variable*, which is analogous to α, the learning rate in the standard perceptron (see Section 2.2). The timing variable also controls the termination of the learning process, and therefore the extent to which TAP fits the data. For example, early stopping leads to a less complex model and a more approximate fit (for more details see Briscoe et al., 2010).

Briscoe et al. address AA as a TAP rank preference learning problem and achieve results close to the upper bound. One of the advantages of TAP is its linear training complexity, which makes it less computationally expensive, especially with large amounts of data, though the implementation is not publically available. We also treat automated assessment of FCE texts as a discriminative ranking learning problem, and, more specifically, we use ranking SVMs (Joachims, 2002) through the SVM$^{\text{light}}$ package (Joachims, 1999), a publically-available efficient implementation of the SVM framework (Vapnik, 1995), which has been shown to achieve state-of-the-art performance in various natural language processing tasks. Although, as discussed in Chapter 2, Section 2.2.7, SVMs tend to be slow when data size increases, we expect this to not have a large effect on our experiments given our sample sizes. Further, among the aims of this research is to facilitate replicability and cross-system comparisons. SVM rank preference optimisation is described in more detail below.

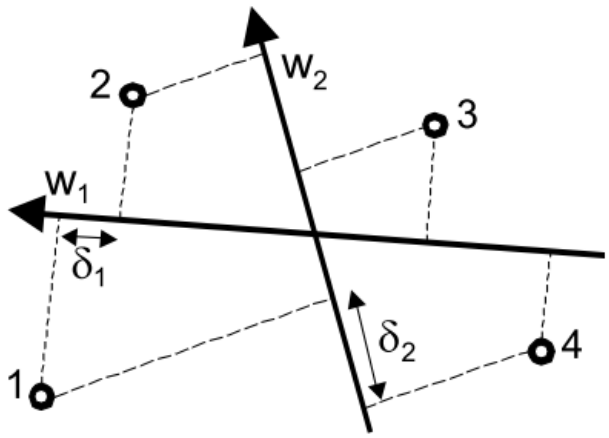


Figure 3.2: Example weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ producing rankings (1, 2, 3, 4) and (2, 3, 1, 4) respectively (Joachims, 2002).

SVMs have been extensively used for learning classification, regression and ranking functions. In its basic form, a binary SVM classifier learns a linear threshold function that discriminates data points of two categories (see Chapter 2, Section 2.2.3). By using a different loss function, the ε-insensitive loss function (Smola, 1996), SVMs can also perform regression, while maintaining all the main maximal margin properties (Cristianini and Shawe-Taylor, 2000). SVMs in regression mode estimate a function that outputs a real number based on the training data. In both cases, the model generalises by computing a hyperplane that has the largest (soft-)margin.

In rank preference SVMs, the goal is to learn a ranking function which outputs a rank/score for each data point, from which a global ordering of the data is constructed. In contrast to regression, a ranking model seeks to identify an optimal ordering of the data directly, rather than to fit a model to a specific score range. The datapoints are ordered by their projection onto the hyperplane; an example is presented in Figure 3.2 to illustrate this point.[3] More specifically, given $\mathbf{w}_1$ the ordering of the four datapoints is (1, 2, 3, 4), while $\mathbf{w}_2$ produces the ordering (2, 3, 1, 4). The rank preference optimisation procedure only considers the difference between pairs of data as evidence (pair-wise difference vectors), and seeks to identify the hyperplane that has the the minimum number of discordant pairs. The intuition behind the use of a ranking model in AA is that high-scoring scripts should be ranked higher than low-scoring ones. Similarly to regression, the script scores can be used as the target values; however, a ranking model will interpret these as the target ranks, and the algorithm will directly model the relationships between scripts, defined by the ordering imposed by their text quality scores.

More formally, this procedure requires a set R consisting of training samples $\mathbf{x}_n$ and their target rankings r_n:

$$R = \{(\mathbf{x}_1, r_1), (\mathbf{x}_2, r_2), ..., (\mathbf{x}_n, r_n)\} \tag{3.1}$$

such that $\mathbf{x}_i \succ_R \mathbf{x}_j$ when $r_i < r_j$, where $1 \leq i, j \leq n$ and $i \neq j$. As mentioned earlier, a rank preference model is not trained directly on this set of data objects and their labels; rather a set of pair-wise difference vectors is created. The goal of a linear ranking model

[3]This example is taken from Joachims (2002).

is to compute a weight vector $\boldsymbol{\theta}$ that maximises the number of correctly ranked pairs:

$$\forall(\mathbf{x}_i \succ_R \mathbf{x}_j) : \boldsymbol{\theta}(\mathbf{x}_i - \mathbf{x}_j) > 0 \tag{3.2}$$

This is equivalent to solving the following optimisation problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum \xi_{ij} \tag{3.3}$$

$$\text{subject to } \begin{cases} \forall(\mathbf{x}_i \succ_R \mathbf{x}_j) : \boldsymbol{\theta}(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij} \\ \xi_{ij} \geq 0 \end{cases} \tag{3.4}$$

The factor C allows a trade-off between the training error and the margin size, while ξ_{ij} are non-negative slack variables that measure the degree of misclassification. The optimisation problem is equivalent to that for the classification model on pair-wise difference vectors. In this case, generalisation is achieved by maximising the differences between closely-ranked data pairs.

Briscoe et al. (2010) outline the key properties of ranking methods: the principal advantage of applying rank preference learning to the AA task is that it allows us to explicitly represent the grade relationships between scripts and learn an optimal ranking model of text quality, across an arbitrary grade range, without having to specify numerical scores or introduce an arbitrary pass/fail boundary. Learning a ranking directly, rather than fitting a classifier score to a grade point scale after training, is both a more generic approach to the task and one which exploits the labelling information in the training data efficiently and directly.

3.3 Evaluation

In order to evaluate the AA system, we follow Briscoe et al. (2010) and use two correlation measures, Pearson's product-moment correlation coefficient (r) and Spearman's rank correlation coefficient (ρ). Pearson's correlation determines the degree to which two linearly dependent variables are related. As Pearson's correlation is sensitive to the distribution of data and, due to outliers, its value can be misleading, Spearman's correlation is also reported. The latter is a non-parametric robust measure of association which is sensitive only to the ordinal arrangement of values. As the data contains some tied values, Spearman's correlation is calculated by using Pearson's correlation on the ranks.

The experimental setup involves building discriminative machine learning models trained on pairs of answers (full scripts), using as label the overall script quality score, again following Briscoe et al. (2010). Compared to training on individual answers and their corresponding score, or training different models for different types of answers/tasks, this setup consistently produces better models. More specifically, analysis of the results showed that correlation between examiners improves when measured at the overall script level,[4] which, in turn, allows us to learn better and more consistent ranking functions. Moreover, as previously mentioned, our focus is on developing an FCE AA model that does not require prompt or topic-specific training.

[4]This might be due to the fact that combining the two scores together hides the effect possible outliers might have.

Features	r	ρ
Baseline	0.664	0.580
GR-LONGEST-TOP-P	0.665	0.579
GR-TOTAL-TOP-P	0.658	0.570
GR-MEAN-TOP-P	0.666	0.580
GR-LONGEST-NBEST-P	0.663	0.573
GR-TOTAL-NBEST-P	**0.670**	**0.585**
GR-MEAN-NBEST-P	0.664	0.577
NBEST-MED-GR-TOTAL-P	0.664	0.580
NBEST-STD-GR-TOTAL-P	0.663	0.572
NBEST-AVG-GR-TOTAL-P	0.666	0.578
NBEST-MED-GR-LONGEST-P	0.661	0.569
NBEST-STD-GR-LONGEST-P	0.665	0.578
NBEST-AVG-GR-LONGEST-P	0.663	0.567

Table 3.1: Correlation between the FCE scores and the AA system predicted values on the development set when adding different complexity features for positive dependencies on top of the baseline AA system.

In this section, we start by examining the predictive power of each of the complexity measures described in Section 3.1.1. In particular, we measure the effect on performance when they are combined with the AA system features described in Briscoe et al. (2010), which we use as our baseline. Tables 3.1 and 3.2 give results on the development set for features relating to positive and negative dependencies respectively. We randomly selected 92 texts from the examination year 2000 as our development data, and used the remaining 1,049 from the same examination year as our training data (see Chapter 2, Section 2.1.1.1). Given the large number of feature instances (approximately 35,000) used by the AA model and the relatively small amount of training data, throughout the experiments we learn a linear ranking function to avoid overfitting.

Most of the complexity measures have a minimal effect on performance, whereas quite a few slightly decrease correlation. Among measures calculated on positive dependencies, the best identified one on this dataset is GR-TOTAL-NBEST, which increases performance by 0.006 and 0.005. The same measure also gives the highest r among negative dependencies, while GR-LONGEST-NBEST-N gives the highest ρ. All measures, however, have small differences compared to each other and the baseline. To examine whether these measures can contribute to a higher improvement in performance, we evaluated a variety of their combinations, and identified a set that achieves a higher increase. In particular, we use as features the maximum values of NBEST-AVG-GR-TOTAL-P/N and NBEST-MED-GR-TOTAL-P/N per script, which represent the mean and median values of GR-TOTAL-NBEST-P/N, that is, the sum of the longest distance in word tokens between a head and dependent in a GR from the RASP GR output, calculated for each GR graph from the top 100 parses per sentence.

The second part of our modifications to the model involves building a better estimate of the error rate. In particular, we extend the ukWaC LM with frequently occurring trigrams extracted from FCE texts (ukWaC+FCE LM), using a threshold of at least eight occurrences. Table 3.3 presents the results when adding either the complexity measures or the new error rate on top of the baseline. Each feature type improves performance on the development data by approximately one percent. We also run tests to examine whether

Features	r	ρ
Baseline	0.664	0.580
GR-LONGEST-TOP-N	0.665	0.579
GR-TOTAL-TOP-N	0.665	0.578
GR-MEAN-TOP-N	0.667	0.582
GR-LONGEST-NBEST-N	0.667	**0.585**
GR-TOTAL-NBEST-N	**0.668**	0.578
GR-MEAN-NBEST-N	0.667	0.582
NBEST-MED-GR-TOTAL-N	0.664	0.580
NBEST-STD-GR-TOTAL-N	0.662	0.570
NBEST-AVG-GR-TOTAL-N	0.665	0.578
NBEST-MED-GR-LONGEST-N	0.663	0.578
NBEST-STD-GR-LONGEST-N	0.664	0.576
NBEST-AVG-GR-LONGEST-N	0.666	0.582

Table 3.2: Correlation between the FCE scores and the AA system predicted values on the development set when adding different complexity features for negative dependencies on top of the baseline AA system.

Features	r	ρ
Baseline	0.664	0.580
Complexity measures	0.670	0.590
Error rate feature		
ukWaC+FCE LM	0.673*	0.588

Table 3.3: Correlation between the FCE scores and the AA system predicted values on the development data when adding the complexity measures or the extended LM on top of the baseline. * indicates a significant improvement in performance at $\alpha = 0.05$.

improvement in correlation is significant. More specifically, we use one-tailed tests for the difference between dependent correlations (Steiger, 1980; Williams, 1959). Complexity measures improve performance significantly on the development data at $\alpha = 0.08$,[5] while the new LM significantly improves Pearson's at $\alpha = 0.05$.

In order to examine the extent to which the new model generalises on unseen data, we validate it on a set of 97 test texts from the exam year 2001, again following Briscoe et al. (2010), whose experiments also demonstrate that marking rubrics[6] evolve over time and thus it is important to have a small temporal distance between training and test data. In addition to testing the new features, we wanted to examine the behaviour of the previously-identified discriminative features. A detailed analysis of the full set of features is presented in Table 3.4. Pearson's and Spearman's correlation between the FCE scores and the AA system predicted values are reported when incrementally adding to the model all feature types of the extended AA model. Each feature type improves the model's performance, including the complexity measures. Extending the LM with FCE

[5]A 0.08 level of significance is not necessarily a too relaxed one; null hypothesis significance testing is biased by sample size, and with small sample sizes, which is our case, we are more susceptible to committing Type II errors, which occur when we fail to reject a false null hypothesis. Thus, increasing the significance level will allow us to reduce the probability of committing them (Rubin, 2009).

[6]By marking rubrics we refer to a standard, based on which someone's performance is evaluated. This consists of the marking criteria, their definitions and examples, as well as the rating scales.

Features	r	ρ
Word ngrams	0.601	0.598
+POS ngrams	0.682	0.687
+Script length	0.692	0.689
+PS rules	0.707	0.708
+Complexity measures	0.714	0.712
Error rate features		
+ukWaC LM	0.735	0.758
+FCE LM	**0.741**	**0.773**
+True FCE error-rate	0.751	0.789

Table 3.4: Incremental correlation between the FCE scores and the AA system predicted values on the test data.

trigrams improves Pearson's and Spearman's correlation by 0.006 and 0.015 respectively. We further experiment with the manually annotated FCE error tags, in order to obtain an upper bound for the performance of an automated error estimator (true FCE error-rate). The addition of the error rate calculated from the FCE error tags on top of all the features further improves performance by 0.01 and 0.016, which shows that the extended LM contributes to further closing this gap. An evaluation of our best error detection method shows a Pearson correlation of 0.611 between the estimated and the true FCE error counts. This suggests that there is room for further improvement in the language model developed. In the experiments reported hereafter, we use the ukWaC+FCE LM to calculate the error rate.

In order to assess the independent as opposed to the order-dependent additive contribution of each feature type to the overall performance of the system, we run a number of ablation tests. An ablation test consists of removing one feature type of the system at a time and re-evaluating the model on the test set.[7] Table 3.5 presents Pearson's and Spearman's correlation between the FCE and the system-predicted values under this evaluation setup. All features have a positive effect on performance, while the error rate has a big impact, as its absence is responsible for a 0.061 decrease of Spearman's correlation. In addition, the removal of either the word ngrams, the PS rules, or the error rate estimate contributes to a large decrease in Pearson's correlation. Again, we test the significance of the improved correlations. The results showed that POS ngrams, PS rules, the complexity measures, and the estimated error rate contribute significantly to the improvement of Spearman's correlation, while PS rules also contribute significantly to the improvement of Pearson's correlation at $\alpha = 0.05$.

As mentioned earlier in Section 3.2, the main advantage of rank preference learning is that it explicitly models the grade relationships between scripts and learns an optimal ranking model. A different way of approaching this problem is to train a binary SVM classifier instead, that discriminates passing from failing FCE texts, and use the confidence margin value generated per text by the decision function of the model as an estimate of the extent to which it has passed or failed. As expected, the results – presented in Table 3.6 – are worse compared to the ranking model when using classification (with significant differences), since the latter does not explicitly model degrees of text quality, but rather directly optimises a pass/fail boundary.

[7] Of course, removing combinations of different feature types may also give rise to useful insights.

Ablated feature	r	ρ
None	**0.741**	**0.773**
Word ngrams	0.713	0.762
POS ngrams	0.724	0.737*
Script length	0.734	0.772
PS rules	0.712*	0.731*
Complexity measures	0.738	0.760*
ukWaC+FCE LM	0.714	0.712*

Table 3.5: Ablation tests showing the correlation between the FCE and the AA system on the test data. * indicates a significant improvement in performance at $\alpha = 0.05$ when these features are added to the AA model.

Model	r	ρ
SVM classification	0.621	0.703
SVM regression	0.697	0.706
TAP rank preference	0.740	0.765
SVM rank preference	**0.741***	**0.773***
Upper bound	0.796	0.792

Table 3.6: Comparison between different discriminative models on the test data. * indicates there is a significant difference in performance at $\alpha = 0.05$ compared to SVM classification and regression.

One of the main approaches adopted by previous systems involves the identification of features that measure writing skill, and then the application of linear or stepwise regression to find optimal feature weights so that the correlation with manually assigned scores is maximised. We trained a SVM regression model with the full set of feature types and compared it to SVM rank preference. The results are given in Table 3.6. The rank preference model improves Pearson's and Spearman's correlation by 0.044 and 0.067 respectively, and these differences are significant, suggesting that rank preference is a more appropriate model for the AA task. It is interesting to note that ρ is approximately the same when using either classification or regression. Our final comparison involved training using TAP, used by Briscoe et al. (2010) to report their best performing system. TAP in ranking mode produces competitive results that are close to the SVM model, while the resulting differences are not significant.

Upper bound

In Briscoe et al. (2010), four senior and experienced ESOL examiners re-marked the 97 FCE test scripts drawn from 2001 exams, using the marking scheme from that year. In order to obtain a ceiling for the performance of the AA system, the average correlation between the FCE and the examiners' scores is calculated. Table 3.6 presents the upper bound – 0.796 and 0.792 Pearson's and Spearman's correlation respectively. These results show that the final AA system is close to the ceiling for the task on this dataset.

In order to evaluate the overall performance of the system, we also calculate its correlation with the four senior examiners in addition to the FCE scores. Tables 3.7 and 3.8 present the results obtained. The average correlation of the AA system with the FCE and the examiner scores again shows that it is close to the upper bound for the task.

	FCE	E1	E2	E3	E4	AA
FCE	–	0.820	0.787	0.767	0.810	0.741
E1	0.820	–	0.851	0.845	0.878	0.721
E2	0.787	0.851	–	0.775	0.788	0.730
E3	0.767	0.845	0.775	–	0.779	0.747
E4	0.810	0.878	0.788	0.779	–	0.679
AA	0.741	0.721	0.730	0.747	0.679	–
Avg	0.785	0.823	0.786	0.782	0.786	0.723

Table 3.7: Pearson's correlation of the AA system predicted values with the FCE and the examiners' scores, where E1 refers to the first examiner, E2 to the second etc.

	FCE	E1	E2	E3	E4	AA
FCE	–	0.801	0.799	0.788	0.782	0.773
E1	0.801	–	0.809	0.806	0.850	0.675
E2	0.799	0.809	–	0.744	0.787	0.724
E3	0.788	0.806	0.744	–	0.794	0.738
E4	0.782	0.850	0.787	0.794	–	0.697
AA	0.773	0.675	0.724	0.738	0.697	–
Avg	0.788	0.788	0.772	0.774	0.782	0.721

Table 3.8: Spearman's correlation of the AA system predicted values with the FCE and the examiners' scores, where E1 refers to the first examiner, E2 to the second etc.

Human–machine correlation is comparable to that of human–human, with the exception of Pearson's correlation with examiner E4, and Spearman's correlation with examiners E1 and E4, where the discrepancies are higher. It is likely that a larger training set and/or more consistent grading of the existing training data would help to close this gap. However, we should note that the final system is not measuring some properties of the scripts, such as discourse coherence and cohesion or relevance to the prompt eliciting the text, that examiners will take into account (the former is discussed and addressed in detail in Chapter 4).

Discussion

We would like to note at this point the difference in the model's performance between development and test data. The two sets are drawn from a different examination year, so we would expect a (slight) drop in correlation on the 2001 test data, as variation between prompts and thus scripts used for training/development and testing should be larger. However, having a closer look at the data, we found that similar prompts (though not identical) tend to be repeated within and between exam years, which explains why performance on the test data does not necessarily decrease. On the other hand, the test set contains texts elicited by prompts of a specific examination period only, whereas the development set (which was randomly selected) contains answers elicited by three different sets of prompts, which increases the variation of texts to be evaluated and partly explains the lower performance in this set. To investigate this further, we also plotted the distribution of scores in each dataset, though clear differences are hard to observe (Figures 3.3 and 3.4). The mean score in the training set is 27.85, in the test is 27.46, and in the development set is 27.35, which suggests that the scores in the test set are slightly closer

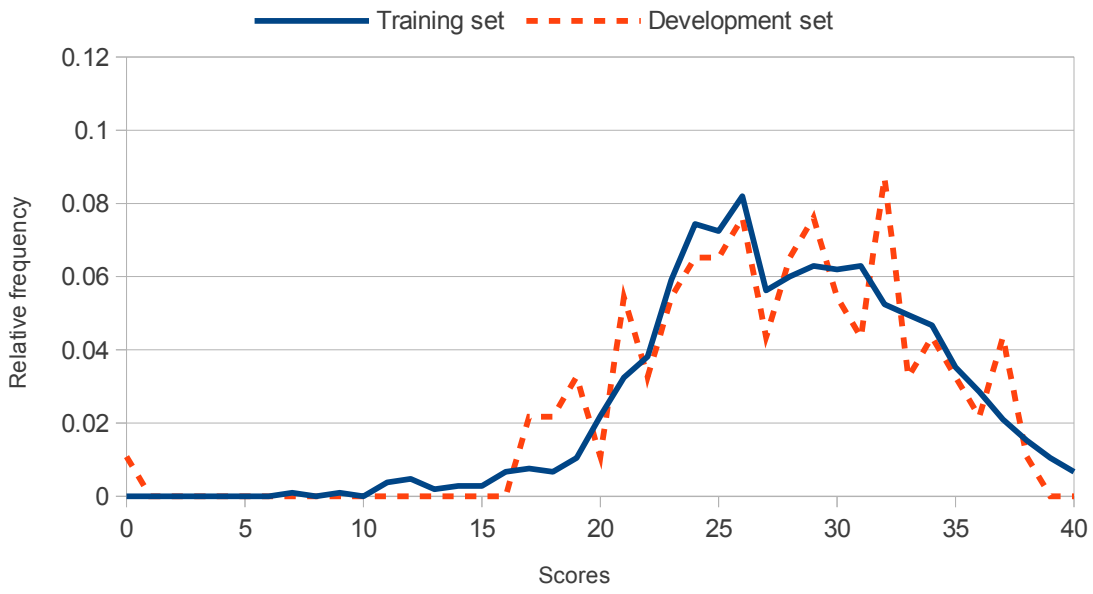


Figure 3.3: Score distribution in training and development data.

to the training data, compared to the development set. Also, skewness in the training and test set is closer to zero, -0.23 and 0.16 respectively, while in the development set is much larger, -0.94, which may explain the lower correlation. We should also note that sampling of different development data gave higher correlations, particularly in ρ by at least 5 percentage points, when using the baseline AA system only.

In further investigations, we ran an experiment to test examiner consistency. We randomly selected and removed from our training set 32 texts elicited by prompts also found in the development data – but not in the test data – and re-evaluated the baseline model on the development set. It is interesting to note that performance improved by at least 0.01 and 0.02 r and ρ respectively.[8] The latter suggests that examiner inconsistency, or perhaps errors in manually transcribing scores, might be further factors that affect performance. Having examiners re-mark the development set would also give us more clear evidence with respect to how close we are or can get to the upper bound on this dataset. Last but not least, experiments on the test data involve a larger training set, which includes the development texts; given our relatively small sample sizes, this is bound to have an effect on performance. Removing the development data from the training set decreases performance on the test set to 0.727 and 0.742 r and ρ respectively.

3.4 Validity tests

The practical utility of an AA system will depend strongly on its robustness to subversion by writers who understand something of its workings and attempt to exploit this to maximise their scores (independently of their underlying ability). Surprisingly, there is very little published data on the robustness of existing systems. However, Powers et al. (2002) invited writing experts to trick the scoring capabilities of an earlier version of e-Rater (Burstein et al., 1998c). e-Rater (see Chapter 2, Section 2.3 for more details) assigns

[8]Differences are significant at $\alpha = 0.07$.

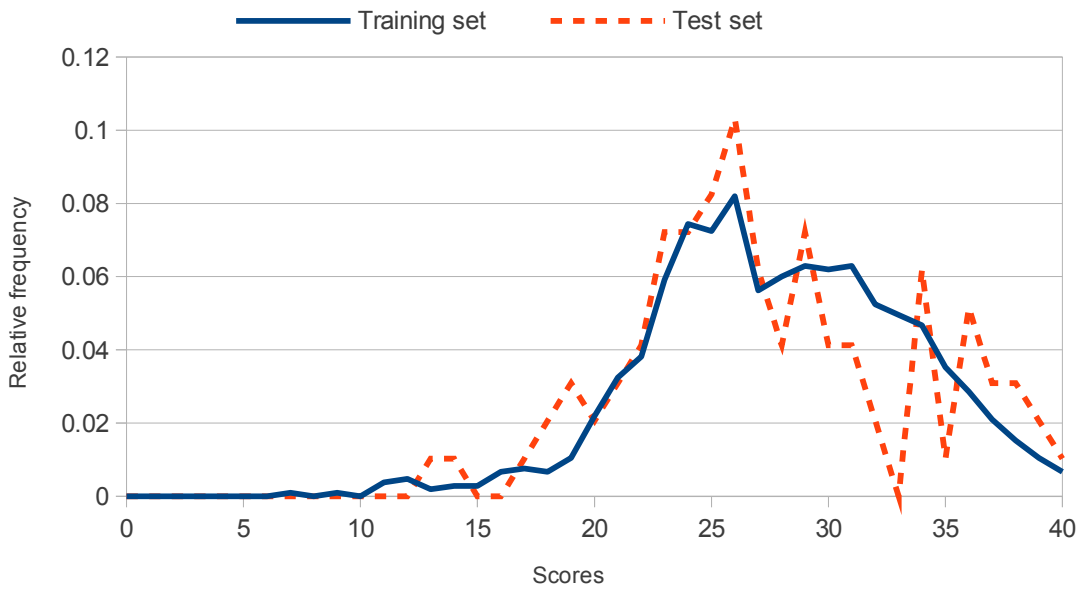


Figure 3.4: Score distribution in training and test data.

a score to a text based on linguistic feature types extracted using relatively domain-specific techniques. Participants were given a description of these techniques as well as of the cue words that the system uses. The results showed that it was easier to fool the system into assigning higher than lower scores.

Regardless of the likelihood of someone gaming the system, it is important that the methodologies utilised guard against threats to its validity. Our goal here is to determine the extent to which knowledge of the feature types deployed poses a threat to the validity of our AA system, where certain text generation strategies may give rise to large positive discrepancies. As mentioned in Chapter 2, Section 2.1.1.1, the marking criteria for FCE scripts are primarily based on the accurate use of a range of different grammatical constructions relevant to specific communicative goals, but the system assesses this indirectly.

We extracted six high-scoring FCE scripts from the CLC that do not overlap with our training, development and test data. Based on the features used by the model and without bias towards any modification, we modified each script in one of the following ways:

1. Randomly order:
   (a) word unigrams within a sentence
   (b) word bigrams within a sentence
   (c) word trigrams within a sentence
   (d) sentences within a script
2. Swap words that have the same POS within a sentence

Although the above modifications do not exhaust the potential challenges a deployed AA system might face, they represent a threat to the validity of the system described

Modification	r	ρ
1(a)	0.960	0.912
1(b)	0.938	0.914
1(c)	0.801	0.867
1(d)	0.08	0.163
2	0.634	0.761

Table 3.9: Correlation between the predicted values and the examiner's scores on outlier texts.

here since it is using a highly related feature set. In total, we created 30 such outlier texts (examples are presented in Appendix E), which were given to an ESOL examiner for marking.[9] Using the outlier scripts as well as their original/unmodified versions, we ran our system on each modification separately and calculated the correlation between the predicted values and the examiner's scores. Table 3.9 presents the results.

The predicted values of the system have a high correlation with the examiner's scores when tested on outlier texts of modification types 1(a), 1(b) and 1(c). However, as 1(c) has a lower correlation compared to 1(a) and 1(b), it is likely that a random ordering of ngrams with $N > 3$ will further decrease performance. A modification of type 2, where words with the same POS within a sentence are swapped, results in a relatively high Pearson and Spearman correlation of 0.634 and 0.761 respectively. Analysis of the results showed that the system predicted higher scores than the ones assigned by the examiner. This can be explained by the fact that texts produced using modification type 2 contain a small portion of correct sentences. However, the marking criteria are based on the overall writing quality. The final case, where correct sentences are randomly ordered, receives the lowest correlation. As the system is not measuring discourse coherence and cohesion, discrepancies are much higher; the system's predicted scores are high whilst the ones assigned by the examiner are very low.[10] However, for a writer to be able to generate text of this type already requires significant linguistic competence. On the other hand, however, an examinee might learn by rote a set of well-formed sentences and re-produce these in an exam in the knowledge that an AA system is not checking for coherence. Additionally, not checking for prompt relevance is another factor that may undermine its validity. A number of off-prompt detection models as well as generic methods for assessing text and/or discourse cohesion have been developed and could be deployed in an extended version of the system. We discuss this in detail in the next chapter.

At the other end of the spectrum, it is also likely that highly creative outlier essays may give rise to large negative discrepancies. Recent comments in the British media have focussed on this issue, reporting that, for example, one deployed essay marking system assigned Winston Churchill's speech 'We Shall Fight on the Beaches' a low score because of excessive repetition.[11] Our model predicted a high passing mark for this text, but not the highest one possible, that some journalists clearly feel it deserves.

[9]Please note that these are also included in the released FCE dataset (see Chapter 2, Section 2.1.1.1).

[10]Similarly, repeating one (or more) high-scoring sentence(s) multiple times within a text is also expected to lead the system astray and give rise to large positive divergences between the gold and predicted scores.

[11]http://news.bbc.co.uk/1/hi/education/8356572.stm

3.5 Conclusions

We have shown experimentally how SVM rank preference models can be effectively deployed for automated assessment of FCE ESOL free-text answers. The principal advantage of applying ranking methods to the AA task is that we explicitly model the grade relationships between scripts, across an arbitrary grade range, without having to specify numerical scores or introduce an arbitrary pass/fail boundary, and do not need to apply a further regression step to fit the classifier output to the scoring scheme. Based on a range of previously-used and novel feature types automatically extracted using generic text processing techniques, the final system achieves performance close to the upper bound for the task. Ablation tests highlight the contribution of each feature type to the overall performance, while significance of the resulting improvements in correlation with human scores has been calculated. None of the published work of which we are aware has systematically compared the contribution of different feature types to the AA task, and only a few assess the ease with which the system can be subverted given some knowledge of the features deployed (Chen et al., 2010; Powers et al., 2002). Preliminary experiments based on a set of automatically generated outlier texts have shown the types of texts for which the system's scoring capability can be undermined. A comparison between classification, regression and rank preference models further supports use of the latter.

An area for further research is to experiment with better error detection techniques, since the overall error-rate of a script is one of the most discriminant features, as well as to integrate with the AA system an automatic off-prompt detection model, such as the one described in Briscoe et al. (2010), which does not require re-training for each new question prompt. It is clear from the outlier experiments reported here that the AA system would benefit from features assessing discourse coherence, and to a lesser extent from features assessing semantic (selectional) coherence over longer bounds than those captured by ngrams. The addition of an incoherence metric to the feature set of an AA system has been shown to improve performance significantly (Miltsakaki and Kukich, 2000, 2004), and we address this in detail in the next chapter.

CHAPTER 4

Discourse coherence and cohesion

To date, few attempts have been made to develop new methods and validate existing ones for automatic evaluation of discourse coherence in the noisy domain of learner texts. In this chapter, we present the first systematic analysis and examine the predictive power of several methods for assessing discourse coherence and cohesion, which is also a strong indicator of a learner's level of attainment, under the framework of AA of learner free-text responses. Discourse features also serve to make it harder to subvert AA systems by submitting globally-incoherent but individually high-quality sequences of sentences, which poses a threat to their validity. Additionally, we identify new techniques that outperform previously developed ones and improve on the best published result for AA on the publically-available FCE dataset of English learner free-text examination scripts.

The results presented in Section 4.4 (FCE experiments) were submitted and accepted as a full paper in the 7th Workshop on the Innovative Use of NLP for Building Educational Applications, North American Chapter of the Association for Computational Linguistics: Human-Language Technologies (Yannakoudakis and Briscoe, 2012). Additionally, the work described in Section 4.5 (IELTS experiments) was presented in the Cambridge Assessment English Profile seminars 2011.

4.1 Introduction

As discussed in Chapters 1 and 2, AA systems of English learner text assign grades based on textual features which attempt to balance evidence of writing competence against evidence of performance errors. Previous work has mostly treated AA as a supervised text classification or regression task. As multiple factors influence the linguistic quality of texts, such systems exploit features that correspond to different properties of texts, such as grammar, style, vocabulary usage, topic similarity, and discourse coherence and cohesion.

Cohesion refers to the use of explicit linguistic cohesive devices (e.g., anaphora, lexical semantic relatedness, discourse markers, etc.) within a text that can signal primarily suprasentential discourse relations between textual units (Halliday and Hasan, 1976). Cohesion is not the only mechanism of discourse coherence, which may also be inferred from meaning without presence of explicit linguistic cues. Coherence can be assessed locally in terms of transitions between adjacent clauses, parentheticals, and other textual units capable of standing in discourse relations, or more globally in terms of the overall

topical coherence of text passages.

There is a large body of work that has investigated a number of different coherence models on news texts (e.g., Lin et al., 2011, Elsner and Charniak, 2008, and Soricut and Marcu, 2006). Recently, Pitler et al. (2010) presented a detailed survey of current techniques in coherence analysis of extractive summaries. To date, however, few attempts have been made to develop new methods and validate existing ones for automatic evaluation of discourse coherence and cohesion in texts produced by non-native speakers of English, which are typically noisy and spelling and grammatical errors are common. Moreover, previous work has mostly formulated coherence as a pair-wise ranking problem, in which a set of random permutations is generated per document and then performance is evaluated by measuring how many times a permutation is ranked higher than its original version (see Section 4.7 for more details). The advantage of this is that we can automatically generate and make use of large incoherent text samples. However, it is unrealistic to assume that such patterns are representative of incoherence properties (in learner texts). In an educational setting, we are also interested in classifying a learner text as coherent or not, or otherwise determine an overall ranking (or score) of texts based on their coherence, rather than comparing random permutations of the same document. We thus expect that (some) previously developed coherence models may not generalise well when used in our AA evaluation framework.

Coherence quality is typically present in marking criteria for evaluating learner texts, and it is identified by examiners as a determinant of the overall score. Thus we expect that adding a coherence metric to the feature set of an AA system would better reflect the evaluation performed by examiners and improve performance. Additionally, as demonstrated later in this chapter, the presence of such features also makes it harder to undermine the system's validity. The goal of the experiments presented in this chapter is to measure the effect a number of (previously-developed and new) coherence models have on performance of AA systems. Our contribution is fivefold:

1. We present the first systematic analysis of several methods for assessing discourse coherence and cohesion in the framework of AA of learner free-text responses.

2. We identify new discourse features that serve as proxies for the level of (in)coherence in texts and outperform previously developed techniques.

3. We examine AA model generalisation to different learner corpora, and, in particular, we investigate the extent to which feature spaces are exam-(in)dependent.

4. We improve the best publically-available results, presented in Chapter 3, on the released FCE corpus of learner texts.

5. We explore the utility of our best model for assessing the incoherent outlier texts used in Chapter 3, Section 3.4, and re-examine validity issues of AA.

In the next sections, we start by describing a number of different models for assessing local and global text coherence properties, and then continue with their systematic assessment and evaluation on two different learner corpora. Most of the methods we investigate require syntactic analysis. Again, we analyse all texts using the RASP toolkit (Briscoe et al., 2006).

4.2 Local coherence

4.2.1 'Superficial' proxies

In this section we introduce diverse classes of 'superficial' cohesive features that serve as proxies for coherence. Surface text properties have been assessed in the framework of automatic summary evaluation (Pitler et al., 2010), have been shown to significantly correlate with the fluency of machine-translated sentences (Chae and Nenkova, 2009), and are part of tools developed to provide measures of cohesion and text difficulty in human-written texts (Graesser et al., 2004).

4.2.1.1 Part-of-Speech distribution

The AA system described in Chapter 3 exploited features based on POS tag sequences, but did not consider the distribution of POS types across grades. In coherent texts, textual units depend on each other for their interpretation. Anaphors such as pronouns relate sentences to those where the entities were previously introduced, the recovery of which is essential in coherence. Pronouns can be directly related to (lack of) coherence and make intuitive sense as cohesive devices. We compute the number of pronouns in a text and use it as a shallow feature for capturing coherence. The underlying idea is that the extent to which pronouns are used within a text should have an impact on coherence; too many pronouns may contribute to difficulty in processing the information in the text, whereas too few may impoverish its continuity.

4.2.1.2 Discourse connectives

Discourse connectives are linguistic devices that link units of discourse (such as clauses or sentences) and support their interpretation (for example, *because*, *however*). The use of such connectives in a text should be indicative of (better) coherence. We experimented with a number of different shallow cohesive features as proxies for coherence, and identified a good subset based on fixed lists of words belonging to the following categories:

1. Addition (e.g., additionally)
2. Comparison (e.g., likewise)
3. Contrast (e.g., whereas)
4. Conclusion (e.g., therefore)

The frequencies of these four categories are used as features in our feature vectors. Details of the word lists can be found in Appendix F.

4.2.1.3 Word length

The previous FCE AA system treated script length as a normalising feature, but otherwise avoided such 'superficial' proxies of text quality. However, many cohesive words (though not all) are longer than average, especially for the closed-class functional component of English vocabulary. For example, *furthermore* consists of eleven letters, which is roughly

twice as long as the average English word length (around five letters per word).[1] We thus assess the minimum, maximum and average word length as a superficial proxy for coherence. The intuition behind the use of these features is that the extent to which cohesive words are used in the text should have an influence on word length statistics. On the other hand, however, such features can also measure other aspects of text, such as lexical complexity and vocabulary.

4.2.2 Semantic similarity

Among the features used in Chapter 3, none explicitly captures coherence and none models inter-sentential relationships. In this section, we explore the utility of inter-sentential feature types for assessing discourse coherence using word-level distributional models. Such models induce a semantic space from input texts using vectorial representations that capture word co-occurrence patterns. The underlying idea is that if two words often co-occur in similar contexts, then they are semantically related (Charles, 2000; Rubenstein and Goodenough, 1965).

To date, there is a rich flora of semantic space models developed. The typical process of inducing word space models involves the construction of a full high-dimensional co-occurrence matrix – where the rows represent words and columns represent contexts (whose dimensions depend on the size of the data) – which is then transformed to a new, low-dimensional one – by employing dimensionality reduction techniques, such as Singular Value Decomposition (SVD) (Golub and Reinsch, 1970) or Principal Component Analysis (PCA) (Pearson, 1901) that approximate the original matrix – to account for efficiency and scalability problems. A variety of other methods avoid using dimensionality reduction techniques for several reasons, among which is the high computational cost involved in the process (computation time and memory usage), as well as non-flexibility in updating the model with further data, since this requires re-creation and re-transformation of the co-occurrence matrix. Incremental Semantic analysis (ISA) (Baroni et al., 2007) and Random Indexing (RI) (Sahlgren, 2005) are two such word space models.

In this section, we employ ISA, an efficient technique which directly constructs a low-dimensional co-occurrence matrix without using dimensionality reduction techniques. The steps used to construct an ISA word-level semantic space model are the following: each word is assigned an arbitrary vector of fixed dimensionality δ containing mostly zeros and a small number of randomly distributed $+1$ and -1 values, called a *signature* vector s. The fixed dimensionality is used to reduce the number of dimensions required to represent the full high-dimensional co-occurrence matrix, while allowing a trade-off between accuracy and efficiency. Additionally, a word is assigned a *history* vector h, which records the contexts in which the word occurred. More specifically, given a target word t and a context word c, the context-dependent representation of t, h_t, is obtained by adding a weighted sum of the signature, s_c, of the word with which it co-occurs and its history vector, h_c. In particular, h_t is calculated as follows:

$$h_t \mathrel{+}= i\left(m_c h_c + (1 - m_c)s_c\right) \tag{4.1}$$

where i is a small constant, called *impact rate*, which typically improves performance. The weighting factor m_c represents the extent to which the history of h_c influences the history of h_t. The underlying idea is that the semantics of frequently occurring words

[1]The average word length in the Wall Street Journal is 5.03 letters.

have less informative histories and thus have a small impact on other words' semantics. In particular, the m_c factor depends on the frequency of the context word c as follows:

$$m_c = \frac{1}{\exp(\frac{\text{count}(c)}{k_m})} \tag{4.2}$$

where k_m controls how fast the decrease will be. A word's meaning is captured by this high dimensional vector h representing its co-occurrence with other words. Similarity among words is measured by comparing their history vectors using vector similarity measures, such as cosine similarity.

The main difference between ISA and RI is the way in which the history of a word is constructed. In particular, whenever we observe a target word t and a context word c, RI will update h_t without taking into account the history of c:

$$h_t \mathrel{+}= i\, s_c \tag{4.3}$$

Both models are incremental in the sense that the history vectors can be used to find word similarities at any stage of data processing. However, ISA is fully incremental, as the history vectors of the words evolve based on the current semantic information encoded in their context representation, and, contrary to RI, does not rely on stoplists or global statistics for weighting purposes – instead it uses formula (4.2) as a weighting scheme that depends on the current frequency of the context word – something which also makes it efficient to compute. Moreover, in contrast to RI, ISA can efficiently capture second-order effects in common with other dimensionality-reduction methods based on SVD that account for their effectiveness (Manning and Schütze, 1999). In their noun evaluation task, Baroni et al. (2007) found that ISA outperformed both RI and an SVD-based method on the Lara dataset (Rowland et al., 2005), a longitudinal corpus of transcripts of natural conversation collected from a single child.

Utilising the S-Space package (Jurgens and Stevens, 2010), we trained an ISA model with fairly standard parameters – 1800 dimensions, a context window of 3 words, impact rate $i = 0.0003$ and decay rate $k_m = 50$ – using a subset of ukWaC (Ferraresi et al., 2008), a large corpus of English containing more than 2 billion tokens. We used the POS tagger lexicon provided with the RASP system to discard documents whose proportion of valid English words to total words is less than 0.4; 78,000 documents were extracted in total and were then preprocessed replacing URLs, email addresses, IP addresses, numbers and emoticons with special markers. To measure local coherence we define the similarity between two sentences s_i and s_{i+1} as the maximum cosine similarity between the history vectors h of the words they contain. We exclude articles, conjunctions, prepositions and auxiliary verbs from the calculation of sentence similarity. The overall coherence of a text T is then measured by taking the mean of all sentence-pair scores:

$$\text{coherence}(T) = \frac{\sum_{i=1}^{n-1} \max_{k,j} \text{sim}(s_i^k, s_{i+1}^j)}{n-1} \tag{4.4}$$

where $\text{sim}(s_i^k, s_{i+1}^j)$ is the cosine similarity between the history vectors of the k^{th} word in s_i and the j^{th} word in s_{i+1}, and n is the total number of sentences. We investigate the efficacy of ISA by adding this coherence score, as well as the maximum sim() value found over the entire text, to the vectors of features associated with a text. The hypothesis is that the degree of semantic relatedness between adjoining sentences serves as a proxy for local discourse coherence; that is, coherent text units contain semantically-related words.

Higgins et al. (2004) and Higgins and Burstein (2007) use RI to determine the semantic similarity between sentences of same/different discourse segments (e.g., from the essay thesis and conclusion, or between sentences and the essay prompt), and assess the percentage of sentences that are correctly classified as related or unrelated. The main differences from our approach are that we assess the utility of semantic space models for predicting the overall grade for a text, in contrast to binary classification at the sentence-level, and we use ISA rather than RI. However, we also experimented with RI in addition to ISA, and found that it did not yield significantly different results. In particular, we trained a RI model with 2,000 dimensions and a context window of 3 on the same ukWaC data. Below we only report results for the fully-incremental ISA model, mainly because it does not rely on stoplists or global statistics for weighting purposes for its computation.

4.2.3 Entity-based coherence

The entity-based coherence model, proposed by Barzilay and Lapata (2008), is one of the most popular statistical models of inter-sentential coherence, and learns coherence properties similar to those employed by Centering Theory (Grosz et al., 1995). Local coherence is modelled on the basis of sequences of entity mentions that are labelled with their syntactic roles (e.g., subject, object). More specifically, each text is represented by a grid; the rows of the grid represent the sentences, while the columns represent discourse entities (noun phrases are represented by their head nouns). If an entity is present in a sentence, the cells of the grid represent its syntactic role in that sentence: S for subject, O for object, X for neither, '$-$' if absent.[2] When an entity has more than one role in a given sentence, the one with the highest ranking is chosen. In our case, the hierarchy is $S > O > X$. Below we can see an excerpt from a highly marked FCE text and its corresponding entity grid in Table 4.1.

1. *[Money], Money, Money*
2. *Young [people] always need money, specially [students] who don't earn their [life].*
3. *They need money to pay their [studies] or for going out, but how to find a [job] which is good and not under-paid?*
4. *One of the best is working in a big [supermarket].*
5. *Some can find it boring but it is well-paid and not too tiring.*
6. *Furthermore departement [stores] are always looking for students who would like to work.*
7. *Another [solution] could be working as a [barman], the [problem] is that you go to [bed] very late, but it's very exciting because you meet a lot of people and enjoy your [night].*
8. *Unfortunately it's not possible unless you can rest in the [morning].*

The entity grid has eight rows since the excerpt consists of eight sentences. The total number of entities found is fourteen, therefore we also have fourteen columns. If we look at the first column, we can see that the entity *money* is present in the first sentence, neither as a subject nor as an object, in the second as an object, and in the third again

[2]Note that other representations are possible as well. For example, grid cells may only contain information about whether an entity is present or not, in which case we have two possible labels instead of four.

	money	people	students	life	studies	job	supermarket	stores	solution	barman	problem	bed	night	morning
1	X	–	–	–	–	–	–	–	–	–	–	–	–	–
2	O	S	X	O	–	–	–	–	–	–	–	–	–	–
3	O	–	–	–	O	O	–	–	–	–	–	–	–	–
4	–	–	–	–	–	–	X	–	–	–	–	–	–	–
5	–	–	–	–	–	–	–	–	–	–	–	–	–	–
6	–	–	X	–	–	–	–	S	–	–	–	–	–	–
7	–	X	–	–	–	–	–	–	S	X	S	X	O	–
8	–	–	–	–	–	–	–	–	–	–	–	–	–	X

Table 4.1: Example entity grid, where each cell represents the syntactic role of an entity in a specific sentence.

as an object. In the rest of the sentences, the noun is absent. The underlying assumption of the entity-grid is that coherent texts will contain a small portion of dense columns, containing mostly S's and O's, and many sparse ones, which will mostly consist of '$-$'. On the other hand, such properties will be less pronounced in incoherent texts.

Using the entity grid, we can extract subsequences of the grid columns, which represent syntactic-role transitions of entities between sentences, and calculate entity transition probabilities per document. These probabilities can then be used as features in our feature vectors. For example, the transition probability for the sequence '$-X$' in Table 4.1, which is of length two, is $7/98 = 0.07$ and is calculated as follows:

$$\frac{\text{frequency of transition '} - X\text{'}}{\text{frequency of transitions of length two}} \tag{4.5}$$

We construct the entity grids using the Brown Coherence Toolkit[3] (Elsner and Charniak, 2011b). The tool does not perform full coreference resolution; instead, coreference is approximated by linking entities that share a head noun.[4] Although coreference resolution systems have been shown to perform well, they are usually trained on grammatical texts and their performance is expected to deteriorate when applied to learner data, where misspellings and grammatical errors are common. We use as features the probabilities of different entity transition types, defined in terms of their role in adjacent sentences. In particular, we represent entities with specified roles (S, O, X, $-$) and use transition probabilities of length 2, 3 and 4. Burstein et al. (2010) show how the entity-grid can be used to discriminate high-coherence from low-coherence learner texts. The main difference with our approach is that we evaluate the entity-grid model in the context of AA text grading, rather than binary classification.

4.2.4 Pronoun coreference model

Pronominal anaphora is another important aspect of coherence. Charniak and Elsner (2009) present an unsupervised generative model of pronominal anaphora for coherence

[3]https://bitbucket.org/melsner/browncoherence

[4]Details regarding this heuristic are given in Poesio et al. (2005) and Elsner and Charniak (2010).

modelling, where all the parameters are learned using Expectation Maximisation (EM). Given anaphoric pronouns, they start by training a simple generative model that selects a possible antecedent using P(antecedent|context). This model uses a total of six different multivariate features – corresponding to the position of the sentence relative to the pronoun, position of the head of the antecedent, position and type of the pronoun, syntactic position and type of the candidate antecedent – which result in 2,592 parameters to be learned. They use four EM iterations to learn parameters, and then gradually learn more complex models. More specifically, given the antecedent they generate the pronoun's person, P(person|antecedent), gender, P(gender|antecedent), number, P(number|antecedent) and head/relation-to-head, P(head/relation|antecedent).

In their implementation, they hypothesise that each pronoun is generated by an antecedent around the previous two sentences. The underlying idea is that if the probability of the pronoun given the antecedent(s) is low, this is an indication of low coherence, as it is hard to resolve it correctly. The overall probability of a text is then calculated as the probability of its pronoun assignments. In our experiments, we use the pre-trained model distributed by Charniak and Elsner (2009) for news text (North-American News Corpus, McClosky et al., 2008) to estimate the probability of a text and include it as a feature. However, this model is trained on high-quality texts, so performance may deteriorate when applied to learner data. It is not obvious how to train such a model on learner texts and we leave this for future research.

4.2.5 Discourse-new model

Elsner and Charniak (2008) apply a discourse-new classifier to model coherence. They train a maximum-entropy classifier that distinguishes noun phrases (NPs) that have not been introduced in the discourse (new) from those that have (old), using a number of features inspired by Uryupina (2004), who employs a total of 32 different syntactic and context features, such as the POS tags of head words, the types of determiners, appositions etc. To model coherence, they first assign each NP in a text a label $L_{np} \in \{new, old\}$ using the same-head heuristic – in which NPs with the same head are considered to be coreferent – and then calculate the probability of a text as $\Pi_{np:NPs} P(L_{np}|np)$. Again, following Elsner and Charniak (2008), we use the same model trained on news text (Wall Street Journal) to find the probability of a text and include it as a feature.

4.2.6 IBM coherence model

Soricut and Marcu (2006), inspired by Kevin Knight after a personal communication in 2003, adapted the IBM model 1 (Brown et al., 1993) used in machine translation (MT) to model local discourse coherence. The intuition behind this model in MT is that the use of certain words in a source language is likely to trigger the use of certain words in a target language. Instead, they hypothesised that the use of certain words in a sentence tends to trigger the use of certain words in surrounding sentences. In contrast to semantic space models such as ISA or RI (discussed above) – in which word similarity calculations are symmetric – this method models the intuition that local coherence is signalled by the identification of recurring word patterns across adjacent sentences, thus also adding asymmetry to word associations, which should, in principle, be a better predictor of coherence.

Using the principles behind IBM model 1, the probability of a document can be calculated as follows:

$$P_{\text{IBM}_{\text{dir}}}(T) = \prod_{i=1}^{n-1} \prod_{j=1}^{|s_{i+1}|} \frac{\varepsilon}{|s_i|+1} \sum_{k=0}^{|s_i|} p(s_{i+1}^j | s_i^k) \tag{4.6}$$

where $|s_i|$ and $|s_{i+1}|$ is the total number of words in sentences s_i and s_{i+1} respectively, n is the total number of sentences, and $p(s_{i+1}^j | s_i^k)$ denotes the probability that the j^{th} word in s_{i+1} (s_{i+1}^j) is being triggered by the k^{th} word in s_i (s_i^k). Additionally, the model uses a hidden variable to align words in adjoining sentences and identify which word triggered the use of another one in an adjacent sentence. Thus, all the parameters are learned through EM. The calculations also include the NULL word (s_i^0), which indicates that a word may not be triggered by any other one. The above model is referred to as the direct IBM model 1. Soricut and Marcu (2006) also define its inverse variation, in which the likelihood of observing the words in a sentence is now conditioned on the words in the subsequent sentence.

We extract three million adjacent sentences from ukWaC,[5] and use the GIZA++ (Och and Ney, 2000, 2003) implementation of IBM model 1, which outputs word-alignment probability tables, to obtain the probabilities of recurring word patterns. We then calculate the direct and inverse probabilities per text, and use their values as features in our feature vectors. Pitler et al. (2010) have also investigated this model to measure text quality in automatically-generated texts, but its performance was relatively poor compared to other models, such as the entity-grid.

We further extend the above model and incorporate syntactic aspects of text coherence by training on POS tags instead of lexical items. We try to model the intuition that local coherence is signalled by the identification of POS co-occurrence patterns across adjacent sentences, where the use of certain POS tags in a sentence tends to trigger the use of other POS tags in an adjacent sentence (for example, nouns might trigger the use of pronouns in a subsequent sentence). We analyse the same three million adjacent sentences using the RASP POS tagger and now train the models to obtain the probabilities of recurring POS patterns. Text probabilities are calculated in the same way.

4.2.7 Lemma/POS cosine similarity

A simple method of incorporating (syntactic) aspects of text coherence is to use cosine similarity between vectors of lemma and/or POS-tag counts in adjacent sentences. We experiment with both: each sentence is represented by a vector whose dimension depends on the total number of lemmas/POS-types. The sentence vectors are weighted using lemma/POS frequency, and the cosine similarity between adjacent sentences is calculated. The coherence of a text T is then calculated as the average value of cosine similarity over the entire text:

$$\text{coherence}(T) = \frac{\sum_{i=1}^{n-1} \text{sim}(s_i, s_{i+1})}{n-1} \tag{4.7}$$

Pitler et al. (2010) use word cosine similarity to measure continuity in automatically-generated texts, and they identify it as one of their best models.

[5] We use the same subset of documents as the ones used to train our ISA model in Section 4.2.2.

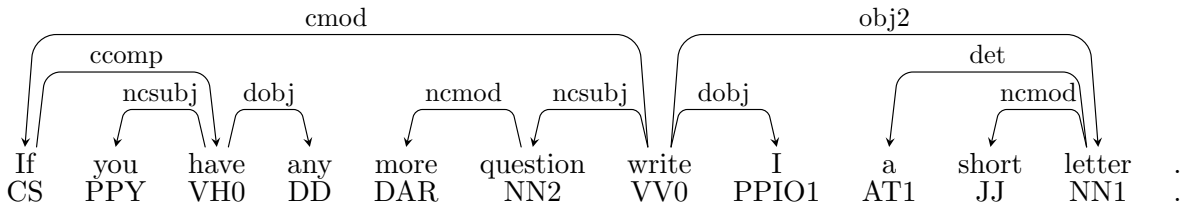


Figure 4.1: Example GR types and dependencies.

4.2.8 GR cosine similarity

Based on the most likely RASP parse for each identified sentence, we extract the list of grammatical relation (GR) types and calculate the GR cosine similarity between adjacent sentences. GR types are automatically identified by RASP, and represent syntactic dependencies between constituents in clauses. An example is illustrated in Figure 4.1 using an FCE excerpt, which shows the different types of relations between words represented as lemmas and POS tags. For example, 'ncsubj' represents binary relations between non-clausal subjects (NPs, PPs) and their verbal heads, as in *have*_VH0 *you*_PPY (for more details see Briscoe, 2006).

The role of grammatical function plays a key part in many theoretical and computational entity-based approaches to (local) coherence (e.g., Grosz et al., 1995 and Barzilay and Lapata, 2008). The entity-grid (see Section 4.2.3), for example, measures local coherence on the basis of sequences of syntactic roles (e.g., subject, object) in adjoining sentences. Also, measures of similarity between adjacent sentences (e.g., word cosine similarity) directly encode aspects of continuity within a text, and have been previously used to assess local coherence (Pitler et al., 2010). We combine the two, and calculate the cosine similarity between adjoining sentences (similarly to the previous section) on the basis of their GRs, and use as features the average and maximum values across a text. We therefore explore the utility of this inter-sentential feature type and investigate whether the degree to which two adjacent sentences share the same GR types serves as a proxy for local coherence.

4.3 Global coherence

4.3.1 Locally-weighted bag-of-words

In the popular bag-of-words (BOW) model a text is represented by a histogram of word occurrences. While this representation is computationally efficient, it is unable to model patterns of sequential information within a text (it can, of course, model short patterns to the extent allowed by the use of contiguous n-grams). The locally-weighted bag-of-words (LoWBOW) framework, introduced by Lebanon et al. (2007), is a sequentially-sensitive alternative to BOW. In BOW, we represent a text as a word histogram, whose dimension depends on the vocabulary used to generate that text. In LoWBOW, a text is represented by a set of local word histograms instead, each one of them calculated over successive regions of text, but smoothed by kernels concentrated around specific positions in the text.

More specifically, a smoothed characterisation of the local histogram is obtained by integrating a length-normalised document with respect to a non-uniform measure that is

concentrated around a particular location $\mu \in [0, 1]$. According to the statistical literature on non-parametric smoothing, such a measure is called a *smoothing kernel*. A smoothing kernel has two parameters, μ and σ, which specify sequential positions in the text, and the extent of smoothing applied over the surrounding region respectively. In contrast to BOW or n-grams, which keep track of frequently occurring patterns independent of their positions, this representation is able to effectively model medium and long range sequential trends in text by keeping track of changes in the histograms from its beginning to end.

More formally, given a set of all the words in our vocabulary V, a document d containing p word positions, and a matrix D which represents document d – with dimensions $|V| \times p$ – and contains 1's whenever a word is found in a specific location and 0 otherwise, a local histogram for d centred at location μ is obtained as follows:

$$\text{hist}_{d,\mu} = \text{sum}((D\, K_\mu)^{\text{T}}) \tag{4.8}$$

where K_μ is a $p \times p$ matrix which contains in its diagonal the position weights obtained from a kernel smoothing function centred at location μ, and sum() sums up the values of the columns of the matrix $(D\, K_\mu)^{\text{T}}$ – which is the transpose of the $(D\, K_\mu)$ matrix multiplication. In other words, it sums the weights for the different positions of the same word. This calculation returns a $1 \times |V|$ vector which contains the summed-up weights per word, which can then be normalised using, for example, the L1 norm, that is, the sum of the absolute values.

Geometrically, LoWBOW uses local smoothing to embed texts as smooth curves in the multinomial simplex. These curves summarise the progression of semantic and/or statistical trends through the text. By varying the amount of smoothing we obtain a family of sequential representations with different sequential resolutions or scales. Low resolution representations capture topic trends and shifts while ignoring finer details. High resolution representations capture fine sequential details, but make it difficult to identify the general trends within the text (Mao et al., 2007). For more details regarding LoWBOW and its geometric properties see Lebanon et al. (2007) and Mao et al. (2007).

Since coherence involves both cohesive lexical devices and sequential progression within a text, we believe that LoWBOW can be used to assess the sequential content and the global structure and coherence of texts. We use a publically-available LoWBOW implementation[6] (Mao et al., 2007) to create local histograms over word unigrams. For the LoWBOW kernel smoothing function, we use the Gaussian probability density function restricted to $[0, 1]$ and re-normalised. We further extend the above model and incorporate syntactic aspects of text coherence by using local histograms over POS unigrams. This representation is able to capture sequential trends abstracted into POS tags. We try to model the hypothesis that coherence is signalled by sequential, mostly inter-sentential progression of POS types. The σ values as well as the total number of local histograms vary across datasets and are presented in the relevant sections.

Since each text is represented by a set of local histograms/vectors, we need to modify standard SVM kernels to work with sets of vectors instead. The Fisher diffusion kernel (Lafferty and Lebanon, 2005) compares local histograms at the same locations only, by taking the inverse cosine of the inner product of the square root of the two vectors, and has proven to be useful for related tasks (Escalante et al., 2011; Lebanon et al., 2007). To the best of our knowledge, LoWBOW representations have not been investigated

[6] http://goo.gl/yQ0Q0

for coherence evaluation (under the AA framework). So far, they have been applied to discourse segmentation (AMIDA, 2007), text categorisation (Lebanon et al., 2007), and authorship attribution (Escalante et al., 2011).

4.4 FCE experiments

We examine the predictive power of the different coherence models described above by measuring the effect on performance when combined with an AA system that achieves state-of-the-art results, but does not use discourse coherence features. Specifically, we present a number of different experiments improving on the FCE AA system described in Chapter 3; AA is treated as a rank preference supervised learning problem and ranking SVMs are used to explicitly model the grade relationships between scripts. This system uses a number of different linguistic features that achieve good performance on the AA task. However, these features only focus on lexical and grammatical properties, as well as errors within individual sentences, ignoring discourse coherence, which is also present in marking criteria for evaluating learner texts, as well as a strong indicator of a writer's understanding of a language and language level.

Also, in Chapter 3, we presented experiments that test the validity of the system using a number of automatically-created outlier texts. The results showed that the model is particularly vulnerable to input where individually high-scoring sentences are randomly ordered within a text. Failing to identify such pathological cases makes AA systems vulnerable to subversion by writers who understand something of its workings, thus posing a threat to their validity. For example, an examinee might learn by rote a set of well-formed sentences and re-produce these in an exam in the knowledge that an AA system is not checking for coherence.

4.4.1 Evaluation

Again, we evaluate the grade predictions of our models against the gold standard grades in the dataset using Pearson's product-moment correlation coefficient (r) and Spearman's rank correlation coefficient (ρ) as is standard in AA research (Briscoe et al., 2010). As discussed in Chapter 3, the machine learning models are trained on full scripts, using the overall score as a label. Our experimental setup involves 5-fold cross-validation using all 1,141 FCE texts from the exam year 2000 (see Chapter 2, Section 2.1.1.1), with 10% used for development and another 10% for testing. This way we can identify parameters that are likely to generalise, as well as test the final models on unseen data from the same distribution/exam year. Table 4.2 presents the final cross-validation results on the test set, obtained by augmenting the baseline model with each of the coherence features described above.[7]

Most of the resulting models have minimal effect on performance.[8] A reason for this might be the fact that coherence properties are already modelled, to some extent, by the baseline features. As Higgins et al. (2004) mention, grammar, usage and mechanics errors

[7]We note that mean values of correlation coefficients should be computed by first applying the r-to-Z Fisher transformation, and then using the Fisher weighted mean correlation coefficient (Faller, 1981; Garcia, 2010).

[8]Significance tests in averaged correlations are omitted as variable estimates are produced, whose variance is hard to be estimated unbiasedly.

	Features	r	ρ
0	Baseline	0.651	0.670
1	POS distr.	0.653	0.670
2	Disc. connectives	0.648	0.668
3	Word length	**0.667**	**0.676**
4	ISA	**0.675**	**0.678**
5	EGrid	0.650	0.668
6	Pronoun	0.650	0.668
7	Disc-new	0.646	0.662
8	$\text{LoWBOW}_{\text{lex}}$	**0.663**	**0.677**
9	$\text{LoWBOW}_{\text{POS}}$	0.659	0.674
10	IBM $\text{model}_{\text{lex}_\text{f}}$	0.649	0.668
11	IBM $\text{model}_{\text{lex}_\text{b}}$	0.649	0.667
12	IBM $\text{model}_{\text{POS}_\text{f}}$	**0.661**	**0.672**
13	IBM $\text{model}_{\text{POS}_\text{b}}$	0.658	0.669
14	Lemma cosine	0.651	0.667
15	POS cosine	0.650	0.665
16	5+6+7+10+11	0.648	0.665
17	All	**0.677**	0.671

Table 4.2: 5-fold cross-validation performance on test texts from year 2000 when adding different coherence features on top of the baseline AA system.

should diminish coherence and flow of text passages. We address this in more detail in the next section, 4.5. However, word length, ISA, $\text{LoWBOW}_{\text{lex}}$, and the IBM $\text{model}_{\text{POS}_\text{f}}$ derived models all improve performance, while larger differences are observed in r. The highest performance – 0.675 and 0.678 – is obtained with ISA, while the second best feature is word length. The entity-grid, the pronoun model and the discourse-new model do not improve on the baseline. Although these models have been successfully used as components in state-of-the-art systems for discriminating coherent from incoherent news documents (Elsner and Charniak, 2011b), and the entity-grid model has also been successfully applied to learner text (Burstein et al., 2010 show that it improves on the baseline for the Criterion essay data[9]), they seem to have minimal impact on performance, while the discourse-new model decreases ρ by ~0.01. On the other hand, $\text{LoWBOW}_{\text{lex}}$ and $\text{LoWBOW}_{\text{POS}}$ give an increase in performance, which confirms our hypothesis that local histograms are useful. Also, the former seems to perform slightly better than the latter. For the LoWBOW kernel smoothing function, we use the Gaussian probability density function restricted to $[0, 1]$ and re-normalised, and a smoothing σ value of 0.02. Additionally, we consider a total number of 9 local histograms per answer.

Our adapted version of the IBM model – $\text{IBM model}_{\text{POS}}$ – performs better than its lexicalised version, which does not have an impact on performance, while larger differences are observed in r. Additionally, the increase in performance is larger than the one obtained with the entity-grid, pronoun or discourse-new model. The forward version of $\text{IBM model}_{\text{POS}}$ seems to perform slightly better than the backward one, while the results are comparable to $\text{LoWBOW}_{\text{POS}}$ and outperformed by $\text{LoWBOW}_{\text{lex}}$. The rest of the models do not perform as well; the number of pronouns or discourse connectives gives low results,

[9]Criterion (Burstein et al., 2003) is an on-line writing evaluation service that integrates e-Rater (see Chapter 2, Section 2.3) and outputs scores as well as diagnostic feedback.

Features	r	ρ
Baseline	0.741	0.773
+ISA	**0.749**	**0.790***
Upper bound	0.796	0.792

Table 4.3: Performance on the exam scripts drawn from the examination year 2001. * indicates a significant difference at $\alpha = 0.05$.

Features	r	ρ
Baseline	0.723	0.721
+ISA	**0.727**	**0.736**

Table 4.4: Average correlation between the AA model, the FCE dataset grades, and four examiners on the exam scripts from year 2001.

while lemma and POS cosine similarity between adjacent sentences are also among the weakest predictors.

Elsner and Charniak (2011b) have shown that combining the entity-grid with the pronoun, discourse-new and lexicalised IBM models gives state-of-the-art results for discriminating news documents and their random permutations. We also combine these models and assess their performance under the AA framework. Row 16 of Table 4.2 shows that the combination does not give an improvement over the individual models. Moreover, combining all feature classes together in row 17 does not yield higher results than those obtained with ISA, while ρ is no better than the baseline.

In the following experiments, we evaluate the best model identified on year 2000 on the 97 texts from the exam year 2001, previously used in Chapter 3 to report results of the final best system. Validating the model on a different exam year also shows us the extent to which it generalises between years. Table 4.3 presents the results. The previous best correlations on this dataset are 0.741 and 0.773 r and ρ respectively. Adding ISA on top of the previous system significantly improves the results on the 2001 texts, getting closer to the upper-bound. Again, significance is calculated using one-tailed tests for the difference between dependent correlations (Steiger, 1980; Williams, 1959). The upper-bound on this dataset, as mentioned in the previous chapter, is 0.796 and 0.792 r and ρ respectively, calculated by taking the average correlation between the FCE grades and the ones provided by four senior ESOL examiners. Table 4.4 also presents the average correlation between our extended AA system's predicted grades and the four examiners' grades, in addition to the original FCE dataset grades. Again, our extended model improves over the baseline.

Finally, we explore the utility of our best model for assessing the publically available outlier texts used in the previous chapter. The previous FCE AA system is unable to appropriately downgrade outlier scripts containing individually high-scoring sentences with poor overall coherence, created by randomly ordering a set of highly-marked texts. To test our best system, we train an SVM rank preference model with the ISA-derived coherence feature, which can explicitly capture such sequential trends. A generic model for flagging putative outlier texts – whose predicted score is lower than a predefined threshold – for manual checking might be used as the first stage of a deployed AA system. The ISA model improves r and ρ by 0.320 and 0.463 respectively for predicting a score on this type of outlier texts and their original version (Table 4.5). However, testing on a larger

Features	r	ρ
Baseline	0.08	0.163
ISA	**0.400**	**0.626**

Table 4.5: Performance of the ISA AA model on outliers.

and/or real-world outlier dataset would allows us to draw more reliable conclusions.

4.4.2 Discussion

In the previous section, we evaluated various cohesion and coherence features on learner data, and found different patterns of performance compared to those previously reported on news texts (see Section 4.7 for more details). Although most of the models examined gave a minimal effect on AA performance, ISA, $\text{LoWBOW}_{\text{lex}}$, IBM $\text{model}_{\text{POS}_{\text{f}}}$ and word length gave a clear improvement in correlation, with larger differences in r. Our results indicate that coherence metrics further improve the performance of a competitive AA system. More specifically, we found the ISA-derived feature to be the most effective contributor to the prediction of text quality. This suggests that incoherence in FCE texts might be due to topic discontinuities. Also, the improvement obtained by LoWBOW suggests that patterns of sequential progression within a text can be useful: coherent texts appear to use similar token distributions at similar positions across different documents.

The word length feature was successfully used as a proxy for coherence, perhaps because many cohesive words are longer than average. However, such a feature can also capture further aspects of texts, such as lexical complexity, and the extent to which it measures different properties is not clear. On the other hand, the minimal effect of the entity-grid, pronoun and discourse-new model suggests that infelicitous use of pronominal forms or sequences of entities may not be an issue in FCE texts. Preliminary investigation of the scripts showed that learners tend to repeat the same entity names or descriptions rather than use pronouns or shorter descriptions. However, the last two models are trained on correct text, so their performance is expected to degrade on learner data. Burstein et al. (2010), among their experiments, show how an augmented version of the entity-grid, containing additional features related to writing quality and word usage, can be used to improve performance on discriminating high-coherence from low-coherence learner texts (for further details see 4.7). Application of this model to FCE texts would be an interesting avenue for future research.

A possible explanation for the difference in performance between the lexicalised and POS IBM model is that the latter abstracts away from lexical information and thus avoids misspellings and reduces sparsity. Elsner (2011) provide an alternate version of IBM model 1, which is trained only on nouns and verbs. Evaluation of this version is another direction for future work. Finally, although the use of discourse connectives is part of the marking criteria, they do not seem to have predictive power. This may be because our manually-built word lists do not have sufficient coverage, or, as discussed in De Felice and Pulman (2008b) for ESOL CLC texts, L2 learners tend to rely on and overuse small sets of fixed expressions, including discourse markers, which can reduce their discriminative power.

	r					ρ				
	s	ta	cc	lr	gra	s	ta	cc	lr	gra
s	–	0.969	0.959	0.970	0.967	–	0.962	0.964	0.968	0.957
ta	–	–	0.919	0.911	0.903	–	–	0.926	0.902	0.876
cc	–	–	–	0.899	0.895	–	–	–	0.909	0.884
lr	–	–	–	–	0.947	–	–	–	–	0.943
gra	–	–	–	–	–	–	–	–	–	–

Table 4.6: Score correlation between the CLC IELTS scores measured against the overall script-level marks.

4.5 IELTS experiments

In the previous sections we hypothesised that adding coherence metrics to the feature set of an AA system would further improve performance, since they are present in the marking criteria and thus would better reflect the evaluation performed by examiners. The results confirmed the hypothesis and coherence features significantly improved performance of our FCE AA system. However, overall quality scores and coherence scores do not necessarily correlate. For example, although Miltsakaki and Kukich (2004) also significantly improved performance of e-Rater using features that directly model incoherence (see Section 4.7), they identified highly coherent texts whose overall quality score was low and vice versa. In this section we use texts annotated with discourse cohesion and coherence scores to identify and model (in)coherence properties in learner data that directly reflect such a score. Previous work (see Section 4.7) has mostly treated coherence as a ranking problem, in which a set of random permutations is generated per document and then performance is evaluated by measuring how many times a permutation is ranked higher than its original version (e.g., Elsner and Charniak, 2011b on news texts), or as a binary classification task (e.g., discriminating high from low coherence texts Burstein et al., 2010). Our goal is to exploit a coherence score to identify appropriate feature types and investigate their contribution to overall performance. Further, it would be interesting to see whether and to what extent this model diverges from the FCE one. As mentioned earlier, coherence properties should already be modelled, to some extent, by the baseline FCE features, as text coherence and flow is typically affected by errors.

IELTS scripts, also represented in the CLC (see Section 2.1.1.2 for more details), are evaluated and manually marked by examiners according to four different criteria: task achievement (ta), coherence and cohesion (cc), lexical resource (lr), and grammatical range and accuracy (gra). IELTS is not a level-based test (like FCE) but is rather designed to cover a much broader proficiency continuum. Candidates are given a bandscore from 0 to 9 on each of the four criteria according to their performance, as well as an overall bandscore – aggregate score (s) – derived from these four skill-based scores (Williams, 2008). In tables 4.6 and 4.7 we can see that there is a high correlation between different scores, which suggests that examiners do assess different aspects of linguistic quality without exclusively ignoring the rest, which further confirms the results presented previously. On the other hand, however, the marking guidelines and re-marking procedures may encourage examiners to give similar scores across the board, which may also account for the high correlations found.

	r					ρ				
	s	ta	cc	lr	gra	s	ta	cc	lr	gra
s	–	0.944	0.937	0.949	0.944	–	0.942	0.937	0.940	0.939
ta	–	–	0.865	0.859	0.836	–	–	0.878	0.852	0.827
cc	–	–	–	0.867	0.853	–	–	–	0.861	0.837
lr	–	–	–	–	0.909	–	–	–	–	0.902
gra	–	–	–	–	–	–	–	–	–	–

Table 4.7: Score correlation between the CLC IELTS scores measured against per-answer marks.

4.5.1 Feature space

Our focus is on building generic coherence-assessment models for ESOL text that do not require prompt-specific or topic-specific training. To find an optimal set of feature types for the task, we conducted a large number of experiments using held-out development data, reserving a test set for our final evaluation so that we can assess the stability of the selected features across tasks and examination years (see next section, 4.5.2). We investigated the effectiveness of all the coherence models described in Sections 4.2 and 4.3 and tested on FCE, as well as the FCE features presented in Chapter 3, Section 3.1.1, and found different patterns than those previously observed. More specifically, the final set of features identified to be discriminative for directly predicting a coherence score (cc) for the IELTS writing tasks is presented below:

1. Lexical ngrams
   (a) Lemma unigrams
   (b) Lemma bigrams
   (c) Lemma trigrams
2. Number of particular POS tags
   (a) Pronouns (P)
   (b) Cardinal numbers (MC)
   (c) Locative nouns (NNL)
   (d) Lexical verbs (VV)
3. GR features
   (a) GR complexity measures
      i. NBEST-MED-GR-TOTAL-N
   (b) GR cosine similarity
4. LoWBOW
5. Other features
   (a) Number of unique words

(b) Error rate

Lemma unigrams, bigrams and trigrams are all lower-cased, while POS tags are assigned using the RASP tagger, which uses 146 (CLAWS) POS types. The distribution of pronouns (P), cardinal numbers (MC), locative nouns (NNL) as well as lexical verbs (VV) is found to be most discriminative from the POS types. Pronouns make intuitive sense as cohesive devices, but the effectiveness of the other POS types may be due to the high correlation between the coherence score and the aggregate score (see Table 4.6).

Based on the most likely RASP parse for each identified sentence, we extract the list of GR types and use as features the maximum and average GR cosine similarity between adjacent sentences (see Section 4.2.8 above). Additionally, various grammatical complexity measures were again extracted from parses, and their impact on performance of the system was explored, similarly to the FCE experiments described in Chapter 3. We found the average and minimum values across a script of the median of GR-TOTAL-NBEST-N to be discriminative (NBEST-MED-GR-TOTAL-N), which represents the sum of the distances for all GR sets over the top 100 parses for negative dependencies (see Chapter 3, Section 3.1.1). Intuitively, the latter is measuring the linguistic complexity of the text rather than specifically coherence.

We also investigated the use of local histograms over ngrams at the word level obtained by the LoWBOW framework. As previously mentioned, LoWBOW allows us to model abstractly sequential information together with word usage. We found that discriminative LoWBOW parameters for directly modelling coherence are different compared to the FCE experiments. More specifically, we construct two local histograms per answer and use a smoothing σ value of 0.15. In order to estimate the error rate, we follow an approach similar to the FCE AA experiments. We use a trigram LM in the same way as for FCE, which is now extended with frequently occurring trigrams extracted from high-ranked IELTS scripts (ukWaC+IELTS LM). We then count a word trigram as an error if it is not found in the language model. It is expected that a large number of errors will also impede textual coherence. Last but not least, the number of unique words captures aspects of the vocabulary used by the learner, and its discriminative power may be explained by the high correlation between the coherence and the lexical resource score, which, in turn, highly correlates with the overall aggregate score.

Feature instances of type 1 are weighted using $tf*idf$ and their vectors are normalised by the L2 norm. Feature type 2 is weighted using frequency counts, while 2, 3(a) and 5 are scaled so that their final value has approximately the same order of magnitude as 1. Features whose overall frequency is lower than three are discarded from the model.

4.5.2 Evaluation

In line with the FCE experiments, our research goals, and previous research on ESOL AA models of overall text quality (Briscoe et al., 2010), we train and evaluate our models using the overall coherence score for both answers on texts from consecutive examination years. In Tables 4.6 and 4.7 we can further see that correlation between scores is higher when using the script-level annotation. Briscoe et al. (2010) also demonstrate that marking rubrics evolve over time and thus it is important to have a small temporal distance between training and test data. Therefore, we use 728 scripts for training and 123 scripts for developing our model from the examination year 2008, and 100 texts from year 2010 for testing. The training and development scripts are the full set available from the manually

Features	r	ρ
Lemma ngrams	0.583	0.527
+POS counts	0.631	0.585
+Unique words	0.657	0.604
+GR cosine	0.683	0.641
+LoWBOW	0.714	0.674
+NBEST-MED-GR-TOTAL-N	0.725	0.683
+ukWaC+IELTS LM	**0.749**	**0.707**
+True IELTS error rate	0.759	0.723

Table 4.8: Correlation between the IELTS coherence scores and the AA system predicted values on the development set.

error-coded CLC closest in time to the 2010 test set. Again, we restricted the data to the error-coded part of the CLC so that we could compare performance of automatically-estimated error features to the ones derived from the manual error coding (as with the FCE experiments in Chapter 3, Section 3.3), in addition to facilitating future research on AA involving error-type detection.

Table 4.8 presents Pearson's and Spearman's correlation between the IELTS coherence scores and the AA system's predicted values on the development set when incrementally adding to the model the feature types presented above. Each feature type has a positive effect and improves the model's performance by at least a 0.01 increase in r. The highest correlations obtained are 0.749 and 0.707 r and ρ respectively. The addition of the error-rate obtained from the manually-annotated IELTS error tags on top of these features further improves performance by 0.01 and 0.016, similar to the effect observed on FCE, which confirms that the error rate is a good predictor. An evaluation of our best error detection method shows a Pearson correlation of 0.740 between the estimated and the true error counts. This suggests that our language model captures the true error rate to a large extent. In the experiments reported hereafter, we do not use any features based on manual annotation in the CLC.

We also trained a TAP ranking model and an SVM regression model with our selected set of feature types and compared them to our SVM ranking model. The results are given in Table 4.9. Our ranking model improves r and ρ by approximately 0.02 and 0.004 compared to the second best learning system, TAP rank preference, though the differences are not significant. Next, we divided the data into pass (mark above 5) and fail classes and trained a binary SVM classifier. The hypothesis is that the confidence margin value generated per text by the decision function of the model can be used as an estimate of the extent to which it has passed or failed. In line with the FCE results, the latter does not produce high correlations, while SVM ranking significantly outperforms SVM classification and regression. The differences in performance between ranking and regression are much larger compared to the FCE experiments in Chapter 3, Section 3.3.

Using the best feature set and machine learning method found on the development set, we run experiments on 100 test texts from the examination year 2010. The first row of Table 4.10 presents the overall performance on the test set: 0.771 and 0.785 r and ρ respectively. In order to assess the independent contribution of each feature type to the overall performance of the system on the test set, we run a number of ablation tests. This will give us more clear evidence regarding their effectiveness and generalisation. Table 4.10 also presents Pearson's and Spearman's correlation between IELTS and our system when

Model	r	ρ
SVM classification	0.522	0.447
SVM regression	0.544	0.514
TAP rank preference	0.726	0.703
SVM rank preference	0.749*	0.707*

Table 4.9: Comparison between different discriminative models on the development data. * indicates there is a significant difference in performance at $\alpha = 0.05$ compared to SVM classification and regression.

Ablated feature	r	ρ
None	0.771	0.785
Lemma ngrams	0.718*	0.719*
Unique words	0.728*	0.740*
ukWaC+IELTS LM	0.744*	0.759*
NBEST-MED-GR-TOTAL-N	0.767	0.776*
LoWBOW	0.772	0.779
GR cosine	0.777	0.788
POS counts	0.779	0.802
Upper bound	0.794	0.789

Table 4.10: Ablation tests on the test set using the best feature combination found on the development set. * indicates there is a significant difference in performance at $\alpha = 0.05$.

removing one feature type at a time. Most features have a positive effect on performance; lemma ngrams and the number of unique words have a big impact as their absence is responsible for at least a 0.04 decrease in r and ρ; the differences in performance are significant, including those for the estimated error rate and the GR complexity measure. On the other hand, GR cosine similarity does not seem to have an effect on performance, while the absence of POS counts increases correlation by 0.008 and 0.017, though the differences are not significant. Further, we tested the extent to which performance varies depending on the amount of training data. Disregarding the development texts during training decreases performance on the test set to 0.754 and 0.771.

Upper bound

In order to estimate an upper bound for the performance of any model evaluated on the test set, we asked four senior and experienced ESOL examiners to re-mark the dataset using the appropriate marking guidelines from the 2010 examination year. We then calculated the average correlation between the IELTS and the examiners' scores on the test set and found a ceiling of 0.794 and 0.789, as illustrated in Table 4.11, in addition to the upper bound for the rest of the detailed scores. From the table we can clearly see that coherence scores are among the ones that cause the highest disagreement between examiners, compared to the rest of the scores and the aggregate one (s), which displays the largest correlation. Nevertheless, our system is close to the coherence upper bound, with minimal differences in ρ and a 0.02 difference in r. Since we are using scores derived from IELTS to train our model and these are based on marks assigned by unknown examiners, who we assume do not outperform those employed to do the re-marking, we cannot expect in general to go beyond these levels of correlation. It is interesting to

Score	r	ρ
s	0.864	0.852
ta	0.810	0.728
cc	0.794	0.789
lr	0.818	0.813
gra	0.824	0.839

Table 4.11: Correlation between the CLC and the examiners' scores on the test set for each of the four separate IELTS scores as well as for the aggregate score per script. 's' represents the aggregate score; 'ta' represents the task achievement score; 'cc' the coherence and cohesion score; 'lr' the lexical resource score; 'gra' the grammatical range and accuracy score.

	IELTS	**E1**	**E2**	**E3**	**E4**	**AA**
IELTS	-	0.789	0.735	0.820	0.821	0.771
E1	0.789	-	0.703	0.825	0.800	0.745
E2	0.735	0.703	-	0.722	0.744	0.610
E3	0.820	0.825	0.722	-	0.859	0.801
E4	0.821	0.800	0.744	0.859	-	0.775
AA	0.771	0.745	0.610	0.801	0.775	-
Avg	0.790	0.776	0.706	0.810	0.803	0.747

Table 4.12: Pearson's correlation of the AA system predicted values with the CLC and the examiners' scores on the test set, where E1 refers to the first examiner, E2 to the second etc.

note, however, that ablating the feature type involving the counts of particular POS tags increases correlation to 0.779 and 0.802 r and ρ, where the latter outperforms the upper bound (see Table 4.10). This can be explained by the fact that some examiners have introduced larger discrepancies compared to the rest, and this affects the overall upper bound, as demonstrated in Tables 4.12 and 4.13. Examiner number 2 (E2) seems to have the largest disagreement with the rest.

In order to have an overall view of our system's performance, we further calculated its correlation with the four senior examiners in addition to the IELTS scores. Tables 4.12 and 4.13 present the results obtained on the test set. The average correlation of the IELTS AA system with the CLC and the examiner scores shows that it is close to the upper bound for the task. Human-machine correlation is comparable to that of human-human correlation, with the exception of correlation with examiners E1 and E2, where the discrepancies are higher. These examiners seem to have the lowest agreement with the IELTS scores compared to E3 and E4 and this is also represented in our model's performance; correlation between the CLC mark and the examiners strongly affects the correlation between the AA system and the examiners. On the other hand, the AA system's results are higher when compared against E3 and E4. It is expected that a larger training set and/or more consistent grading of the existing training data would help to close this gap. Nevertheless, our results indicate that good performance can be achieved without the need to train on scripts that use identical tasks, and these results are in line with those obtained for the FCE AA model.

	IELTS	E1	E2	E3	E4	AA
IELTS	-	0.790	0.721	0.814	0.821	0.785
E1	0.790	-	0.660	0.810	0.797	0.691
E2	0.721	0.660	-	0.695	0.715	0.571
E3	0.814	0.810	0.695	-	0.859	0.786
E4	0.821	0.797	0.715	0.859	-	0.759
AA	0.785	0.691	0.571	0.786	0.759	-
Avg	0.788	0.756	0.676	0.799	0.796	0.727

Table 4.13: Spearman's correlation of the AA system predicted values with the CLC and the examiners' scores on the test set, where E1 refers to the first examiner, E2 to the second etc.

4.5.3 Discussion

It is interesting to note at this point the similarities and differences between the feature types identified in the different datasets. Lexical and POS ngrams are prominent in both FCE and IELTS, though in different forms. PS-rules are highly discriminative for FCE, though not for IELTS; however, GR cosine similarity appeared to have a positive impact on performance during IELTS system development. This is somewhat surprising, since RASP automatically produces GRs from PS-rules. The GRs themselves, however, were not found to be discriminative in either dataset.

The GR complexity measure as well as the error rate are discriminative in both datasets, while LoWBOW has a positive effect on IELTS, though not the highest one possible when evaluated on FCE. However, in the FCE coherence experiments we focused on a systematic assessment of several (individual) models, while in IELTS our goal was to build a coherence-assessment model through identification of appropriate sets of features.

On the other hand, features based on vector space models, such as ISA, and other coherence models discussed in literature, such as the IBM model, were not found to be part of a discriminative feature set in these experiments, in contrast to the results presented on FCE and despite their use to assess coherence in previous work (see Section 4.7). Differences between IELTS and FCE can be attributed to the different marking rubrics and guidelines, in addition to the differences in the underlying text collections (see Section 4.6 below for a direct comparison).

Further investigation would be needed to understand why specific IELTS features are discriminative and to what extent they reflect text coherence (e.g., GR cosine similarity and counts of particular POS tags, such as locative nouns). Research using visualisation to examine these highly weighted features and/or their distribution across scripts would be an interesting area for future research. However, we do address this to some extent for FCE AA discriminative features in the next chapter.

The fact that models suggested in previous research did not perform as well suggests that learner data and/or framing the task as a scoring problem is a distinct subcase of coherence assessment. However, we were able to show the utility of some new feature types. It is worth mentioning at this point that during IELTS feature selection we did observe a positive effect on performance when using the discourse-new model (Elsner and Charniak, 2008).

Train set	Test set	Features	r	ρ
FCE train+dev	FCE test	FCE	0.749*	0.790*
FCE train+dev	FCE test	IELTS	0.661	0.660
FCE train	FCE test	FCE	0.735*	0.756*
FCE train	FCE test	IELTS	0.651	0.650
FCE train	FCE dev.	FCE	0.687	0.618
FCE train	FCE dev.	IELTS	0.624	0.569

Table 4.14: Correlation between the gold and the AA system predicted scores on the FCE development and test set under different feature spaces. * indicates a significant difference at $\alpha = 0.05$ compared to the IELTS counterpart.

4.6 Feature-space generalisation

Feature selection and, generally, AA system development for FCE and IELTS, was motivated by our building of generic task-independent models to assess the quality of a text itself. In the previous experiments, we were able to see the extent to which the set of features identified to be discriminative for FCE and IELTS diverges between exams and datasets. In this section, we experiment with two different conditions to assess these differences quantitatively. First, we train a model on the FCE dataset using the IELTS features, and then evaluate it on the FCE development and test set. Next, we run a similar experiment on IELTS data, training and testing the model on IELTS texts, but now using the FCE features. As discussed in the previous section, the IELTS and FCE feature spaces do share some similarities, and the overall quality and coherence scores have been shown to highly correlate. On the other hand, the models have been developed on different datasets and thus degradations in performance are expected. Nevertheless, this experiment will also allow us to quantitatively investigate the extent to which these models/features are exam-(in)dependent. Tables 4.14 and 4.15 present the results on the FCE and IELTS development and test data using the configurations and best feature combinations described in Sections 4.4.1 and 4.5.1. To increase control over feature-set comparisons and remove possible bias, we run further experiments in which the dataset used for training is the same under each condition.

In Table 4.14 we can see that the FCE features give significantly higher performance on the FCE test set compared to the IELTS ones, although the latter perform relatively well with correlation varying between 0.65 and 0.66. A similar pattern is observed on the development set, though the differences in this case are not significant. On the other hand, in Table 4.15 we observe a slightly varying effect. The IELTS features do perform better compared to the FCE counterpart on the test set, though performance differences are not significant. Correlation between gold and predicted scores using the FCE feature space are very close to the best performing IELTS model. However, on the IELTS development set there is a significant difference, while performance decreases down to 0.52 and 0.48 when using the FCE features, which is a larger degradation compared to the lowest IELTS feature-space performance on the FCE development set, that is 0.62 and 0.56 r and ρ respectively.

Overall, the FCE features produce the most variable results, with correlation varying between a low of 0.48 and a high of 0.76 on IELTS data, while performance using the IELTS features exhibits larger stability and lies between 0.56 and 0.66 on the FCE texts.

Train set	Test set	Features	r	ρ
IELTS train+dev	IELTS test	IELTS	0.771	0.785
IELTS train+dev	IELTS test	FCE	0.747	0.766
IELTS train	IELTS test	IELTS	0.754	0.771
IELTS train	IELTS test	FCE	0.720	0.743
IELTS train	IELTS dev.	IELTS	0.749*	0.707*
IELTS train	IELTS dev.	FCE	0.523	0.489

Table 4.15: Correlation between the gold and the AA system predicted scores on the FCE and IELTS test set under different conditions. * indicates that there is a significant difference at $\alpha = 0.05$ between the last two conditions.

Although correlation differences can be attributed to the different marking schemes and guidelines, and overall quality and coherence scores need not always correlate, the results suggest that development of an exam-independent model is a potential direction for future research.

4.7 Related work

Comparatively few metrics have been investigated for evaluating coherence in ESOL learner texts. Miltsakaki and Kukich (2004) manually annotated a corpus of learner texts with coreference, applied Centering Theory's algorithm (Grosz et al., 1995), and showed that the distribution of Centering transitions correlates with examiner scores. In particular, they show that Centering Theory's Rough-Shift transitions – which represent the lowest degree of coherence and, more specifically, abrupt changes of the focus across adjacent sentences – contribute significantly to the assessment of learner texts, when integrated in the feature set of e-Rater (Attali and Burstein, 2006). They find that incoherence in their corpus is due to discontinuities introduced by the use of multiple undeveloped topics within a conceptually uniform segment (i.e., a paragraph) rather than infelicitous use of pronominal forms, which closely resembles our findings on the FCE dataset.

Sentence similarity measures have guided research on aspects of coherence in learner data (e.g., Wiemer-Hastings and Graesser, 2000, Higgins et al., 2004 and Higgins and Burstein, 2007). Higgins et al. (2004) focus on four different aspects of coherence in learner texts: relatedness to the question prompt; relatedness between and within discourse segments; intra-sentential quality, where the main goal is to provide feedback to learners with respect to particular text units. To model the first two, they train a RI model and then extract a number of different features, including the semantic similarity scores between sentences and the prompt, and the semantic similarity between sentences in different discourse segments. Further features include the number of sentences in a discourse segment, the number of sentences in a segment whose similarity to other discourse segments is greater than a threshold, and the maximum semantic similarity score between a sentence and the ones in the prompt. These scores are then given as input to an SVM classifier that predicts whether a sentence is classified as related or not either to the prompt or other discourse segments. The last dimension, intra-sentential quality, is modelled using heuristic rules that look for grammar, usage and mechanics errors, whose presence should affect coherence of text passages. This is similar to our approach on the IELTS data, where we found the error rate to be a highly discriminative feature for

directly assessing coherence. Perhaps more directly comparable to our FCE experiments are their results on the third dimension, relatedness within discourse segments. Although the use of semantic similarity features makes intuitive sense in this scenario too, they found that it is hard to beat the baseline and identify sentences which are not related to other ones in the same discourse segment, as 98.1% of the sentences were annotated as highly related. Herein, we demonstrated that the related fully-incremental ISA model can be used to improve AA grading accuracy on the FCE dataset, as opposed to classifying the (non-)relatedness of sentences.

In a following paper, Higgins and Burstein (2007) further investigate the relationship of a sentence to the text prompt. They argue that identifying off-prompt content results in a "breakdown in coherence due to more global aspects of essay-based discourse structure". Although we do not explicitly measure this aspect of coherence here, we do believe that the LoWBOW model captures this to some extent. Highly marked texts should exhibit good coherence, which is reflected in word usage (including content words). Word distribution within texts is captured by LoWBOW, and thus we expect that content organisation – which reflects prompt-specific aspects – should also be modelled. On the other hand, Briscoe et al. (2010) describe an approach to automatic off-prompt detection with high performance on CLC data that does not require re-training for each new question prompt and uses an ISA model. We plan to integrate this approach with our system in the near future.

Burstein et al. (2010) examine three different sets of essay data and show how the entity-grid can be used to discriminate high-coherence from low-coherence learner texts. Entity transition features improve over the baseline on one set of texts. Augmenting this model with additional features related to writing quality and word usage shows a positive effect on performance for binary automated coherence prediction in all their data. On the texts used here, entity-grids do not improve AA grading accuracy. This may be because the texts are shorter or because grading is a more difficult task than binary classification. Application of their augmented entity-grid model to CLC texts would be an interesting avenue for future research.

There is large body of work that has investigated coherence on news texts and articles. Foltz et al. (1998) examine local coherence in textbooks and articles using Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer et al., 1998). They assess semantic relatedness using vector-based similarity between adjacent sentences. The hypothesis is that coherent texts exhibit a high degree of meaning overlap between adjoining sentences. They argue that LSA may be more appropriate for comparing the relative quality of texts; for determining the overall text coherence it may be difficult to set a criterion for the coherence value since it depends on a variety of different factors, such as the size of the text units to be compared. Nevertheless, our results show that ISA, a similar distributional semantic model with dimensionality reduction, improves FCE grading accuracy, though it is not a discriminative feature for IELTS. Moreover, and contrary to our findings, Barzilay and Lapata (2008) show that the entity-grid outperforms LSA on three different applications: text ordering using synthetic data, automatic coherence evaluation of machine-generated summaries, and readability assessment.

Barzilay and Lee (2004) implement lexicalised content models that represent global text properties on news articles and narratives using Hidden Markov Models (HMMs). In the HMM, states represent distinct topics, and transitions between states represent the probability of moving from one topic to another. This approach has the advantage

of capturing the order in which different topics appear in texts; however, the HMMs are highly domain specific and would probably need retraining for each distinct essay prompt. In the text ordering task of Barzilay and Lapata (2008), their entity-based model performs at least as well as the HMMs and in other cases significantly better. Again, however, we expect that global text properties are modelled, to some extent, by LoWBOW. Investigating prompt-specific training and evaluation using LoWBOW is a possible direction for future work.

Soricut and Marcu (2006) use a log-linear model that combines local and global models of coherence and show that it outperforms each of the individual ones on news articles and accident reports. Their global model is based on the document content model proposed by Barzilay and Lee (2004). Their local model of discourse coherence is based on the entity-grid, as well as on the lexicalised IBM model; we have experimented with both, and showed that they have a minimal effect on grading performance with the CLC dataset.

Elsner and Charniak (2008) and Elsner and Charniak (2011a) apply a discourse-new classifier and a pronoun coreference system (Charniak and Elsner, 2009) (discussed in the previous sections) to model coherence on dialogue and news texts. They found that combining these models with the entity-grid and the IBM model 1 achieves state-of-the-art performance. We found that such a combination, as well as the individual models, do not perform as well for grading CLC texts, with the exception of the discourse-new model which produced good results during IELTS feature development, though not the highest ones possible. The same year, Elsner and Charniak (2011b) modified the entity-grid and augmented it with entity-specific features related to salience, coreference and types of entities, as well as proposed a variation of the entity-grid intended to integrate topical information (Elsner and Charniak, 2011a). More specifically, they use Latent Dirichlet Allocation (Blei et al., 2003) to learn topic-to-word distributions and then use as a feature the similarity between an entity and the subjects of the previous sentence.

Recently, Lin et al. (2011) adopted a different approach to tackle this task on news text. They propose a model that assesses coherence through discourse relations, where the underlying idea is that coherent texts exhibit a preferential ordering of discourse relations. Their implementation closely resembles the one in the entity-grid; however, they focus on modelling the transition of discourse relations in adjacent sentences using a discourse-role grid instead. Evaluation results indicate their model to be complementary to the entity-grid. Applying the above to AA on learner texts would also be an interesting direction for future work.

4.8 Conclusions

We evaluated coherence models and features in two different learner corpora under the AA grading task. On the publically-available FCE texts, we presented the first systematic analysis of a wide variety of previous and new models for assessing discourse coherence and cohesion, and evaluated their individual performance as well as their combinations for the AA task. Our goal was to examine the predictive power of a variety of coherence models by measuring the effect on performance when combined with an FCE AA system that achieves state-of-the-art results on predicting overall quality scores, but does not explicitly use discourse coherence and cohesion features, making it thus vulnerable to subversion. We successfully adapted ISA, an efficient and incremental variant distributional semantic model, to the task, and further identified ISA, LoWBOW, the POS IBM model and word

length as the best individual features for assessing coherence in FCE texts. A significant improvement over the AA system presented in Chapter 3 and the best published result on the FCE dataset were obtained by augmenting the system with an ISA-based local coherence feature. We also explored the robustness of the ISA model of local coherence on 'outlier' texts and achieved much better correlations with the examiner's grades for these texts in the FCE dataset. This should facilitate development of an automated system to detect essays consisting of high-quality but incoherent sequences of sentences.

As overall quality and coherence scores need not always correlate, we run further experiments on the IELTS dataset, which has been manually annotated with discourse coherence and cohesion scores. Previous work has mostly treated coherence as a binary classification problem, or as a pair-wise ranking task, in which a set of random permutations is generated per document and then performance is evaluated by measuring how many times a permutation is ranked lower than its original version. Our goal was to exploit coherence scores to directly identify appropriate feature types and coherence models, and investigate their contribution to overall AA coherence-grading performance. We were able to show the utility of some new feature types, while some bear similarities with the FCE features, such as lemma ngrams and complexity measures. However, it is difficult to know to what extent they are specifically measuring coherence (for example, locative nouns) given the high correlation between coherence and the overall score. We also adapted the LoWBOW model for assessing sequential content in texts, and showed evidence on both FCE and IELTS texts supporting our hypothesis that local histograms are useful. It is quite likely that further experimentation with LoWBOW features, given the large range of possible parameter settings, would yield better results too.

Finally, we investigated the extent to which feature spaces generalise across datasets. FCE features can achieve as high a performance on IELTS texts as 0.76, though the results are highly variable and correlation can get as low as 0.48. On the other hand, IELTS features exhibit higher stability, though correlations are not as high and lie between 0.56 and 0.66. Degradations in performance between IELTS and FCE are not surprising given the differences in the marking schemes and guidelines, as well as the underlying datasets. Nevertheless, results on both exams are close to the estimated upper bound and within the range of variation found amongst the four examiners who remarked the test scripts. This indicates that good performance can be achieved without the need to train on scripts that use identical tasks, therefore our approach requires less customisation compared to task-dependent methods. However, all our results are specific to ESOL CLC texts and may not generalise to other genres or ESOL attainment levels. Future work should also investigate a wider range of (learner) texts and further coherence models, such as that of Elsner and Charniak (2011a) and Lin et al. (2011).

CHAPTER 5

Analysing the 'black box'

In this chapter, we demonstrate how automated assessment systems can support SLA research when integrated with visualisation tools. We present a visual user interface supporting the investigation of a set of linguistic features discriminating between passing and failing FCE ESOL exam scripts. The system displays directed graphs to model interactions between features and supports exploratory search over FCE learner scripts. We illustrate how the interface can support the investigation of the co-occurrence of many individual features, and discuss how such investigations can shed light on understanding the linguistic abilities that characterise different levels of attainment and, more generally, developmental aspects of learner grammars.[1] Further, we evaluate the visualiser through usability studies, which is a key component in ensuring its quality, success, and adoption by the target user population; in our case, SLA researchers, teachers and assessors. Finally, preliminary experiments demonstrate that our approach also effectively contributes towards identifying patterns that can help us further improve performance of automated assessment systems. To the best of our knowledge, this is the first attempt to visually analyse and perform a linguistic interpretation of automatically-determined features that characterise learner English, as well as to illustrate how their visualisation can enhance the identification of new discriminative features.

Work presented in Sections 5.3 and 5.4 was accepted as a full paper in the joint workshop on Visualisation of Linguistic Patterns & Uncovering Language History from Multilingual Resources, European Chapter of the Association for Computational Linguistics (Yannakoudakis et al., 2012), as well as in the Learner Corpus Research conference (Alexopoulou et al., 2013).

5.1 Introduction

Advances in machine learning have led to self-contained out-of-the-box machine learning solutions that are more often than not viewed as 'black boxes'; that is, they are primarily investigated in terms of their input and output, while their internal characteristics may often be ignored. As a result, even after performing extensive 'feature engineering' as exemplified in Chapters 3 and 4 above, there may arise difficulties in interpreting results, in addition to overlooking patterns that could have potentially been used to make the model

[1] The linguistic interpretation of discriminative features has been done in collaboration with Dora Alexopoulou, with whom we also extensively discussed the requirements of the tool.

more effective. Generic approaches to AA have the advantage of modelling consistent 'marking criteria' regardless of the prompt delivered, while machine learning allows us to identify explicit cues in the data that determine the quality of a text. These cues represent the internal 'marking criteria' used to evaluate someone's proficiency level.

In order to assess the validity of AA systems, it is important we understand those criteria and what drives their discriminative power. Although opaque AA marking criteria might better secure the system from being 'gamed' or led astray, AA models are not a panacea, and their deployment largely depends on the ability to examine their characteristics, and, more specifically, whether their internal 'marking criteria' can be interpreted in a meaningful and useful way, whether they measure what is intended to be measured, whether they are accurate and fair, whether any kinds of bias have been introduced, and, in general, whether their development reflects sound pedagogy. It is therefore imperative the methodologies adopted are transparent. Attempts to game the system as a consequence of this may be unavoidable – we already know that standardised assessment 'suffers' from formulaic approaches to teaching and learning writing. Thus, it is of equal importance that at the same time we utilise methodologies that guard against threats to their validity.

5.2 Visualisation

In this chapter, we focus on the interpretation of AA 'marking criteria'. Visualisation techniques can help us shed light on AA 'black boxes', and inspect the features they yield as the most predictive of a learner's level of attainment. As data-driven approaches are quantitatively very powerful, visualisation can help us gain a deeper understanding on their workings. The latter is particularly important for learning models designed to imitate the value judgements examiners make when they mark a text. We build a visual user interface (hereafter UI) which allows investigation and interpretation of a set of linguistic features discriminating between passing and failing FCE ESOL exam scripts. The UI displays directed graphs to model interactions between features and supports exploratory search over FCE learner scripts. Our experiments demonstrate that proper analysis and visualisation of AA features can support SLA research, and, in particular, can shed light on understanding the linguistic abilities that characterise different levels of attainment and, more generally, developmental aspects of learner grammars. Additionally, we illustrate how hypothesis formation through visualisation of discriminative features can aid the identification of new discriminative features, and thus further contribute to informing the development of AA systems.

The UI is developed to analyse features described in Briscoe et al. (2010). Briscoe et al. have also treated FCE AA as a classification problem, and used a binary discriminative classifier to learn a linear threshold function that best discriminates passing from failing FCE scripts, and predict the class to which a script belongs. To facilitate learning of the classification function, the data should be represented appropriately with the most relevant set of features. As mentioned in the previous chapters, they found a discriminative feature set which includes, among other feature types, word and POS ngrams. We extract the discriminative instances of these two feature types and focus on their linguistic analysis. Table 5.1 presents a small subset ordered by discriminative weight. A major advantage in using (supervised) discriminative classifiers to support hypothesis formation over, for example, clustering techniques, is that they assign weights to features

Feature	Example
VM_RR (POS bigram: +)	*could clearly*
,_because (word bigram: −)	*, because of*
necessary (word unigram: +)	*it is necessary that*
the_people (word bigram: −)	**the people are clever*
VV∅_VV∅ (POS bigram: −)	**we go see film*
NN2_VVG (POS bigram: +)	*children smiling*

Table 5.1: Subset of features ordered by discriminative weight; + and − show their association with either passing or failing scripts.

representing their relative importance.

We believe the investigation of discriminative features can offer insights into assessment and into the linguistic properties characterising the relevant CEFR level (see Chapter 1, Section 1.4.2), which can, in turn, be exploited to identify new discriminative patterns that further improve performance of AA systems. However, the amount and variety of data potentially made available by the classifier is considerable, as it typically finds hundreds of thousands of discriminative feature instances. Even if investigation is restricted to the most discriminative ones, calculations of relationships between features can rapidly grow and become overwhelming. Discriminative features typically capture relatively low-level, specific and local properties of texts, so features need to be linked to the scripts they appear in to allow investigation of the contexts in which they occur. The scripts, in turn, need to be searched for further linguistic properties in order to formulate and evaluate higher-level, more general and comprehensible hypotheses which can inform reference level descriptions and understanding of learner grammars.

The appeal of information visualisation is to gain a deeper understanding of important phenomena that are represented in a database (Card et al., 1999) by making it possible to navigate large amounts of data for formulating and testing hypotheses faster, intuitively, and with relative ease. An important challenge is to identify and assess the usefulness of the enormous number of projections that can potentially be visualised. Exploration of (large) databases can quickly lead to numerous possible research directions; lack of good tools often slows down the process of identifying the most productive paths to pursue.

In our context, we require a tool that visualises features flexibly, supports interactive investigation of scripts instantiating them, and allows statistics about scripts, such as the co-occurrence of features or presence of other linguistic properties, to be derived quickly. One of the advantages of using visualisation techniques over command-line database search tools is that SLA researchers and related users, such as assessors and teachers, can access scripts, associated features and annotation intuitively without the need to learn query language syntax.

We modify previously-developed visualisation techniques (Battista et al., 1998) and build a visual UI supporting hypothesis formation about learner grammars through visualisation of discriminative features. Features are grouped in terms of their relative co-occurrence in the corpus and directed graphs are used in order to illustrate their relationships. Selecting different feature combinations automatically generates queries over FCE data and returns the relevant scripts as well as associations with meta-data and different types of errors committed by the learners. In the next sections we describe in detail the visualiser, illustrate how it can support the investigation of individual features, and

discuss how such investigations can shed light on the relationships between features and developmental aspects of learner grammars. Furthermore, we illustrate how hypothesis formation through discriminative features can aid the identification of new discriminative features. In the last section of this chapter, we evaluate the visualiser through usability testing and user feedback; ensuring its quality is essential to successful use by target users.

To the best of our knowledge, this is the first attempt to visually analyse as well as perform a linguistic interpretation of discriminative features that characterise learner English, whose analysis can also inform the development of AA systems. We would also like to point out that we also apply the visualiser to the publically-available FCE ESOL texts (see Chapter 2, Section 2.1.1.1) and make it available as a web service to other researchers.[2]

5.3 The English Profile visualiser

5.3.1 Basic structure and front-end

The English Profile (EP) visualiser is developed in Java and uses the Prefuse library (Heer et al., 2005) for the visual components. Figure 5.1 shows its front-end. Features are represented by labelled nodes and displayed in the central panel; positive features (i.e., those associated with passing the exam) are shaded in a light green colour while negative ones are light red.[3] The size of the node is used to visually encode feature frequencies; the smaller the node, the less frequent the feature. By hovering the mouse over the nodes, a tooltip text is displayed which describes the CLAWS tags for POS ngrams and gives short examples. A field at the bottom right supports searching for features/nodes that start with specified characters and highlights them in blue. An important aspect is the display of feature patterns, discussed in more detail in the next section (5.3.2).

5.3.2 Feature relations

Crucial to understanding discriminative features is finding the relationships that hold between them. We calculate co-occurrences of features at the sentence-level in order to extract 'meaningful' relations and possible patterns of use. Combinations of features that may be 'useful' are kept while the rest are discarded. 'Usefulness' is measured as follows:

Consider the set of all the sentences in the corpus $S = \{s_1, s_2, ..., s_N\}$ and the set of all the features $F = \{f_1, f_2, ..., f_M\}$. A feature $f_i \in F$ is associated with a feature $f_j \in F$, where $i \neq j$ and $1 \leq i, j \leq M$, if their relative co-occurrence score is within a predefined range:

$$\text{score}(f_j, f_i) = \frac{\sum_{k=1}^{N} \text{exists}(f_j, f_i, s_k)}{\sum_{k=1}^{N} \text{exists}(f_i, s_k)} \tag{5.1}$$

where $s_k \in S$, $1 \leq k \leq N$, exists() is a binary function that returns 1 if the input features occur in s_k, and $0 \leq \text{score}(f_j, f_i) \leq 1$. We group features in terms of their relative co-occurrence within sentences in the corpus and display these co-occurrence relationships as directed graphs. Two nodes (features) are connected by an edge if their score, based on Equation (1), is within a user-defined range (see example below). Given

[2]Available upon request: http://ilexir.co.uk/applications/ep-visualiser/

[3]Colours can be customised by the user.

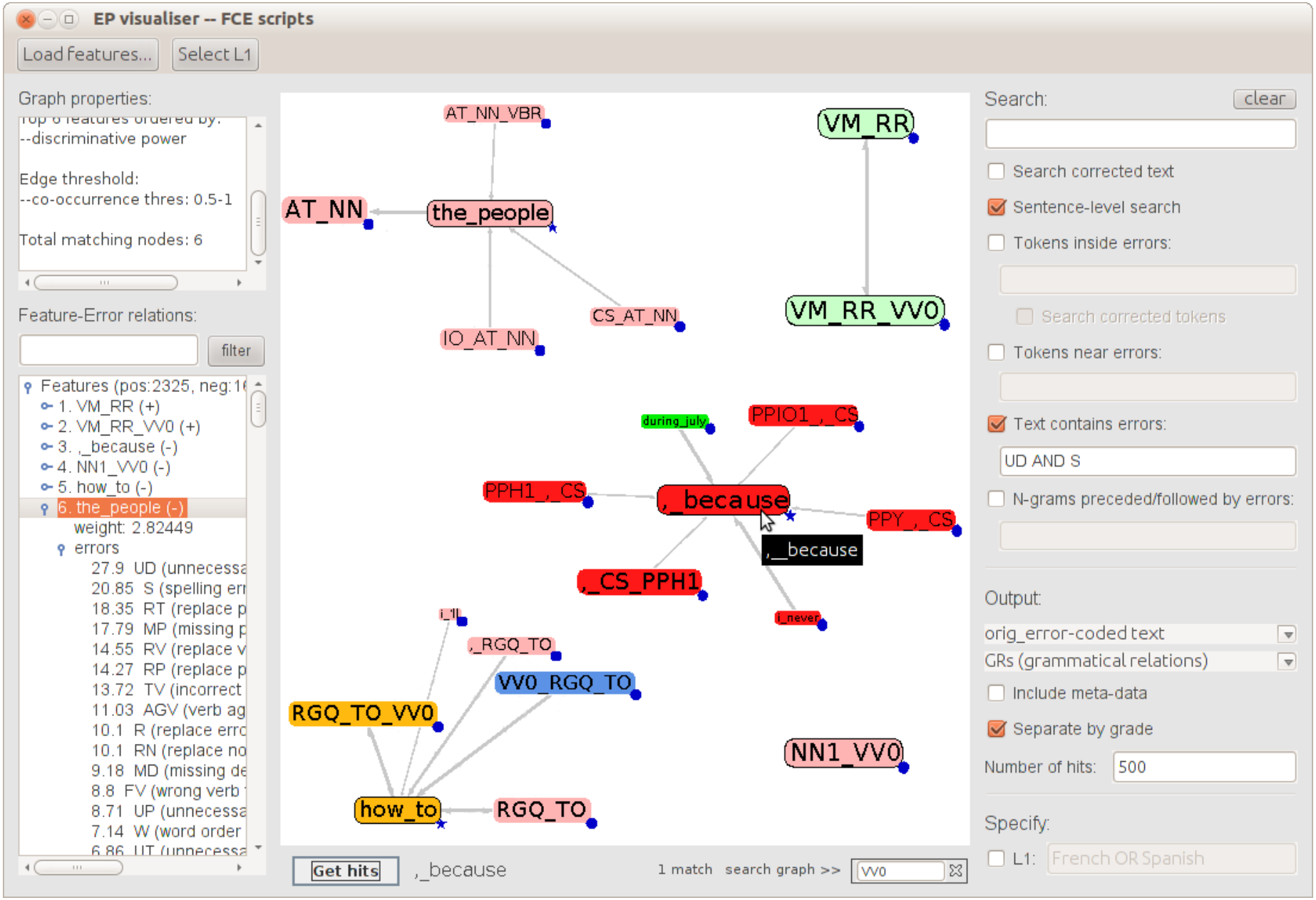


Figure 5.1: Front-end of the EP visualiser.

f_i and f_j, the outgoing edges of f_i are modelled using score(f_j, f_i) and the incoming edges using score(f_i, f_j). Feature relations are shown via highlighting of features when the user hovers the cursor over them, while the strength of the relations is visually encoded in the edge width.

For example, one of the highest-weighted positive discriminative features is VM_RR (see Table 5.1), which captures sequences of a modal auxiliary followed by an adverb as in *will always (avoid)* or *could clearly (see)*. Investigating its relative co-occurrence with other features using a score range of 0.8–1 and regardless of directionality, we find that VM_RR is related to the following: (i) POS ngrams: RR_VBØ_AT1, VM_RR_VBØ, VM_RR_VHØ, PPH1_VM_RR, VM_RR_VVØ, PPIS1_VM_RR, PPIS2_VM_RR, RR_VBØ; (ii) word ngrams: will_also, can_only, can_also, can_just. These relations show us the syntactic environments of the feature (i) or its characteristic lexicalisations (ii).

5.3.3 Dynamic creation of graphs via selection criteria

Questions relating to a graph display may include information about the most connected nodes, separate components of the graph, types of interconnected features, and so on. However, the functionality, usability and tractability of graphs is severely limited when the number of nodes and edges grows by more than a few dozen (Fry, 2007). In order to provide adequate information, but at the same time avoid overly complex graphs, we support dynamic creation and visualisation of graphs using a variety of selection criteria. The EP visualiser supports the flexible investigation of the top 4,000 discriminative features and their relations.

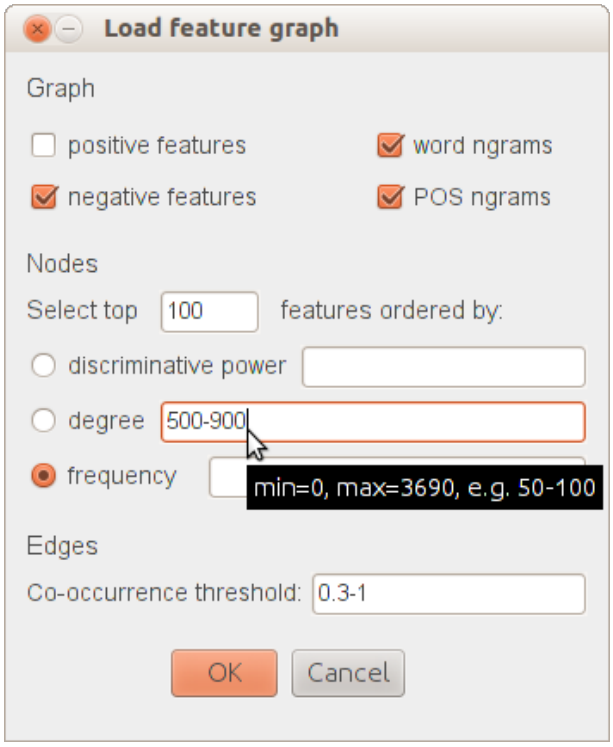


Figure 5.2: Selecting features as well as the co-occurrences to be visualised.

The *Menu* item on the top left of the UI in Figure 5.1 activates a panel that enables users to select the top N features to be displayed (Figure 5.2). The user can choose

whether to display positive and/or negative features, select ranking criteria, as well as define filters based on their characteristics. The last two can be defined in terms of the features' discriminative power,[4] degree (i.e., the total number of discriminative features with which a feature co-occurs at the sentence level), and frequency. For instance, a user can choose to investigate features that have a degree between 500 and 900, then rank them by their frequency and finally display the top 100. Highly-connected features might lead to useful insights on learner grammar while infrequent features, although discriminative, might not lead to useful hypotheses. Additionally, users can investigate co-occurrence relations and set different score ranges (using the *Co-occurrence threshold* field) according to Equation (1), which controls the edges to be displayed. By hovering the mouse over the text fields, a tooltip text is displayed that shows the range of values the users can choose from.

Figure 5.3a presents the graph of the 5 most frequent negative features, using a score range of 0.8–1. The system displays only one edge, while the rest of the features are isolated. However, these features might be related to other features from the list of 4,000 (which are not displayed since they are not found in the top N list of features). Blue aggregation markers in the shape of a circle, located at the bottom right of each node, are used to visually display that information. When a node with an aggregation marker is selected, the system automatically expands the graph and displays the related features. The marker shape of an expanded node changes to a star, while a different border stroke pattern is used to visually distinguish the revealed nodes from the top N. Figure 5.3b presents the expanded graph when the aggregation marker for the feature VVD_II is selected. If the same aggregation marker is selected twice, the graph collapses and returns to its original form.

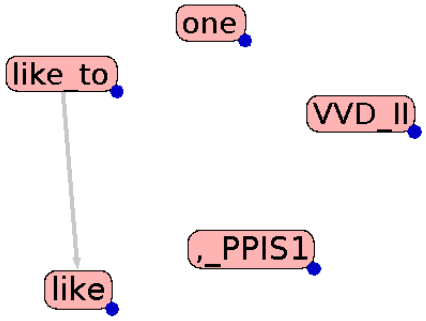
5.3.4 Feature–Error relations

The FCE texts have been manually error-coded (Nicholls, 2003) so it is possible to find associations between discriminative features and specific error types. The *Feature–Error relations* component on the left of Figure 5.1 displays a list of the features, ranked by their discriminative weight, together with statistics on their relations with errors. Feature–error relations are computed at the sentence level by calculating the proportion of sentences containing a feature that also contain a specific error (similar to Equation (1)). In the example in Figure 5.1, we see that 27% of the sentences that contain the bigram feature the_people also have an unnecessary determiner (UD) error, while 14% have a replace verb (RV) error.

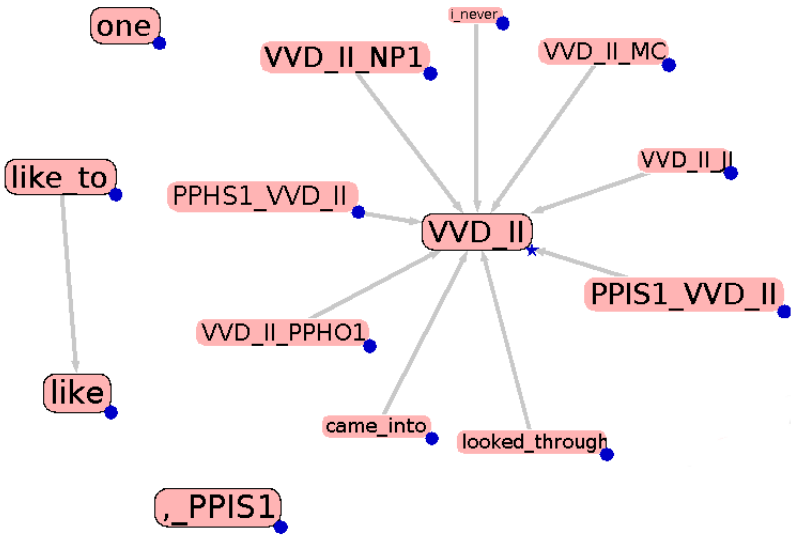
5.3.5 Searching the data

In order to allow the user to explore how features are related to the data, the EP visualiser supports browsing operations. Selecting multiple features – highlighted in yellow – and clicking on the button *get hits* returns relevant texts. The right panel of the front-end in Figure 5.1 displays a number of search and output options. The interface extends and integrates a command-line Lucene (Gospodnetic and Hatcher, 2004) CLC search tool developed by Gram and Buttery (2009), which allows for a wide range of specialised

[4]The higher the number, the lower the discriminative power of that feature; for example, a discriminative power of 200 means that the feature is ranked as number 200 in the list of discriminative features.



(a) Graph of the top 5 most frequent negative features using a score range of 0.8–1.



(b) Expanded graph when the aggregation marker for the feature VVD_II is selected.

Figure 5.3: Dynamic graph creation.

search queries on parsed and error-coded texts, including searches on the original and corrected scripts, as well as on meta-data. Users can choose to output the original/error-coded/POS-tagged text and/or the grammatical relations found by the RASP parser (Briscoe et al., 2006), while different colours are used in order to visually distinguish data annotations and enhance readability. Texts can be retrieved at the sentence or script level and separated according to grade, varying from A to E. Additionally, Boolean queries can be executed in order to examine occurrences of (selected features and) specific errors only.[5] Further other options include searching for scripts containing a token tagged inside or near a specific error, while the corrected texts can also be queried. Also, users can investigate scripts based on meta-data information, and specifically, native language (Gram and Buttery, 2009).

Figure 5.4 shows the display of the system when the features how_to and RGQ_TO_VV∅ (*how to* followed by a verb in base form) are selected. The text area in the centre displays sentences instantiating them. A search box at the top supports navigation, highlighting search terms in red, while a small text area underneath displays the current search query, the size of the database and the number of matching scripts or sentences. On the bottom left, the overall word count of all matching texts is also displayed. The *Errors by decreasing frequency* pane on the left shows a list of the errors found in the matching scripts, ordered by decreasing frequency. Three different tabs (lemma, POS and lemma_POS) provide information about and allow extraction of counts of lemmata and POSs inside an error tag.

5.3.6 Learner L1

Research on SLA has investigated the possible effect of a native language (L1) on the learning process. Using the *Menu* item on the top left corner of Figure 5.1, users can select the language of interest while the system displays a new window with an identical front-end and functionality. Feature–error statistics are now displayed per L1, while selecting multiple features returns scripts written by learners speaking the chosen L1.

At this point, we would also like to point out that we adopted a user-driven development of the visualiser based on the needs of an SLA researcher who acted as a design partner during the development of the tool and was eager to use and test it. There were dozens of meetings over a period of seven months, and the feedback on early interfaces was incorporated in the version described here. After the prototype reached a satisfactory level of stability, the final version overall felt enjoyable and inviting, as well as allowed her to form hypotheses and draw on different types of evidence in order to substantiate it (Alexopoulou et al., 2013).

5.4 Interpreting discriminative features: a case study

We now illustrate in greater depth how the EP visualiser can support interpretation of discriminative features through a case study. We investigate the POS trigram RG_JJ_NN1 (−), which is the 18th most discriminative (negative) feature. It corresponds to a sequence of a degree adverb followed by an adjective and a singular noun as in *very good boy*.

[5]For example, users can activate the *Text contains errors:* option and type 'R OR W'. This will return sentences containing replace or word order errors.

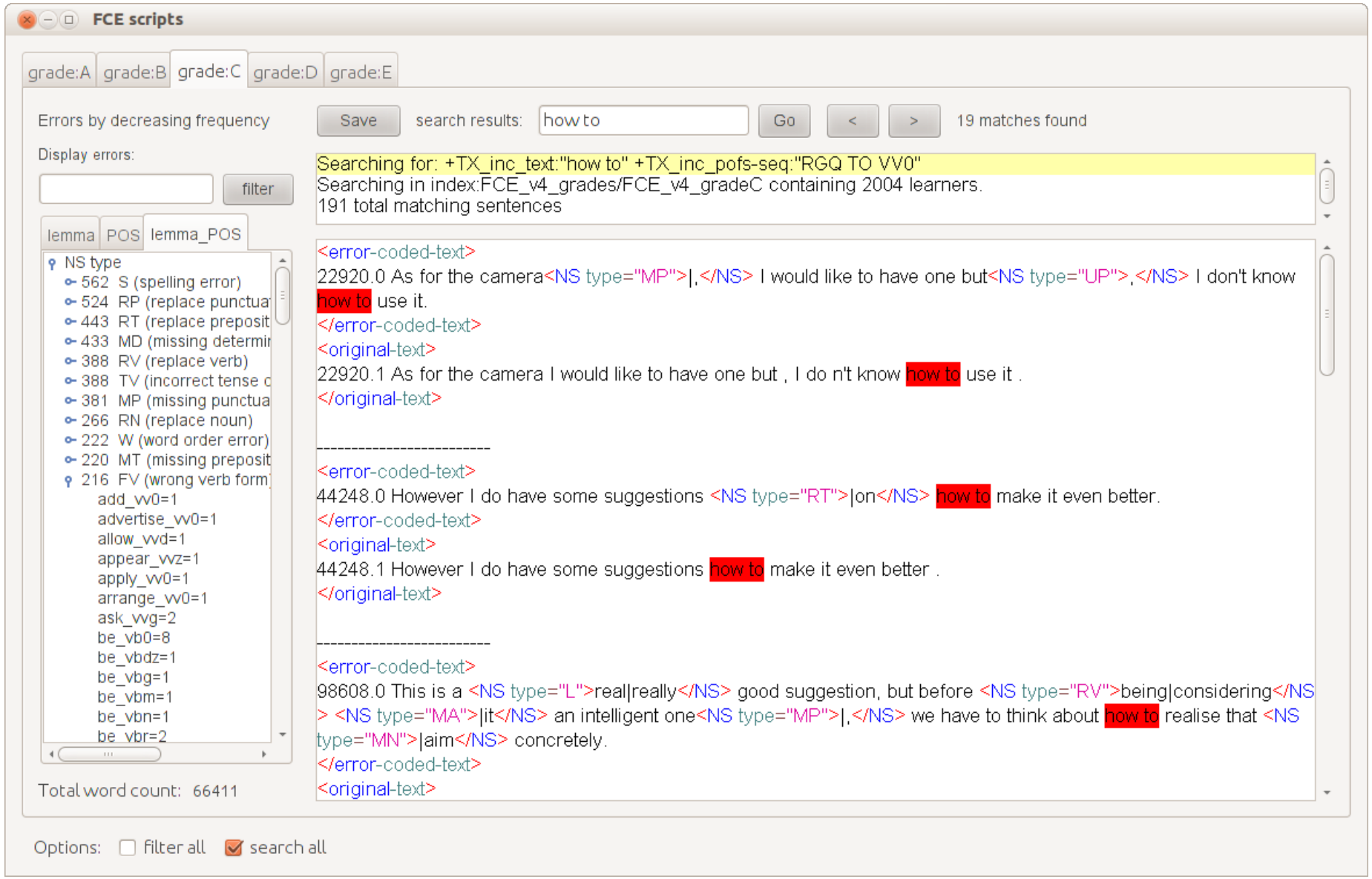


Figure 5.4: Sentences, split by grade, containing occurrences of how_to and RGQ_TO_VV∅. The list on the left gives error frequencies for the matching scripts, including the frequencies of lemmata and POS tags inside an error.

The question is why such a feature is negative since the string is not ungrammatical. Visualisation of this feature using the 'dynamic graph creation' component of the visualiser allows us to see the features it is related to. This offers an intuitive and manageable way of investigating the large number of underlying discriminative features by examining their associations with the feature of interest.

We find that RG_JJ_NN1 is related to its discriminative lexicalisation, very_good (−), which is the 513th most discriminative feature. Also, it is related to JJ_NN1_II (−) (e.g., *difficult sport at*), ranked 2,700th, which suggests a particular context for RG_JJ_NN1 when the noun is followed by a preposition. Searching for this conjunction of features in scripts, we get production examples like *1a,b,c*. Perhaps more interestingly, RG_JJ_NN1 is related to VBZ_RG (−) (ranked 243rd): *is* followed by a degree adverb. This relation suggests a link with predicative structures since putting the two ngrams together yields strings VBZ_RG_JJ_NN1 corresponding to examples like *1c,d*; if we also add _II we get examples like *1c*.

1a *It might seem to be* ***very difficult sport at*** *the beginning.*

1b *We know a lot about* ***very difficult situation in*** *your country.*

1c *I think it's* ***very good idea to*** *spending vacation together.*

1d *Unix* ***is very powerful system*** *but there is one thing against it.*

The associations between features already give an idea of the source of the problem. In the sequences including the verb *be* the indefinite article is omitted. So the next thing to investigate is if indeed RG_JJ_NN1 is associated with article omission, not only in predicative contexts, but more generally. The *Feature–Error relations* component of the UI reveals an association with MD (missing determiner) errors: 23% of sentences that contain RG_JJ_NN1 also have a MD error. The same holds for very_good, JJ_NN1_II and VBZ_RG with percentages 12%, 14% and 15% respectively. We then compared the number of MD errors per script across different types of scripts. Across all scripts the ratio MD:doc is 2.18, that is, approximately 2 MD errors per script; in RG_JJ_NN1 scripts this ratio goes up to 2.75, so that each script has roughly 3 MD errors. VBZ_RG follows with 2.68, JJ_NN1_II with 2.48, and very_good with 2.32. In scripts containing all features the ratio goes up to 4.02 (3.68 without very_good), and in scripts containing VBZ_RG_JJ the ratio goes up to 2.73. Also, in most of these scripts the error involves the indefinite article. The emerging picture then is that there is a link between these richer nominal structures that include more than one modifier and the omission of the article. Two questions arise: (i) why these richer nominals should associate with article omission and (ii) why only singular nouns are implicated in this feature.

Article omission errors are typical of learners coming from L1s lacking an article system (Hawkins and Buttery, 2010; Ionin and Montrul, 2010; Robertson, 2000). Trenkic (2008) proposes that such learners analyse articles as adjectival modifiers rather than as a separate category of determiners or articles. When no adjective is involved, learners may be aware that bare nominals are ungrammatical in English and provide the article. However, with complex adjectival phrases, learners may omit the article because of the presence of a degree adverb. In order to evaluate this hypothesis further we need to investigate if article omission is indeed more pronounced in our data with more complex

Language	f_1	f_2	f_3	f_4
all	0.26	0.40	0.02	0.03
Turkish	0.29	0.48	0.04	0.03
Japanese	0.17	0.39	0.02	0.02
Korean	0.30	0.58	0.06	0.03
Russian	0.35	0.52	0.03	0.03
Chinese	0.25	0.56	0.02	0.03
French	0.21	0.41	0.00	0.03
German	0.19	0.41	0.00	0.02
Spanish	0.27	0.32	0.00	0.03
Greek	0.30	0.35	0.02	0.02

Table 5.2: feature:doc ratios for different L1s.

adjectival phrases e.g., *very difficult situation* than with simpler ones e.g., *nice boy* and whether this is primarily the case for learners from L1s lacking articles.

Again, using the *Errors by decreasing frequency* pane we found that the MD:doc ratio in scripts containing the bigram JJ_NN1 is 2.20. Additionally, in scripts containing JJ_NN1 and not RG_JJ_NN1 it goes down to 2.04. These results are much lower compared to the MD:doc ratio in scripts containing RG_JJ_NN1 and/or the features with which it is related (see above), further supporting our hypothesis. We also found the ratio of RG_JJ_NN1 (f_1) occurrences per document across different L1s, as well as the ratio of VBZ_RG_JJ (f_2), VBZ_RG_JJ_NN1 (f_3) and RG_JJ_NN1_II (f_4). As shown in Table 5.2 there is no correlation between these features and the L1, with the exception of f_1 and f_2 which are more pronounced in Korean and Russian speakers, and of f_3 which seems completely absent from French, German and Spanish which all have articles. The exception is Greek which has articles but uses bare nominals in predicative structures.

However, a more systematic pattern is revealed when relations with MD errors are considered (using the *Feature–Error relations* and *Errors by decreasing frequency* components for different L1s). As shown in Table 5.3, there is a sharp contrast between L1s with articles (French, German, Spanish and Greek) and those without (Turkish, Japanese, Korean, Russian, Chinese), which further supports our hypothesis. A further question is why only the singular article is implicated in this feature. The association with predicative contexts may provide a clue. Such contexts select nominals which require the indefinite article only in the singular case; compare *Unix is (a) very powerful system* with *Macs are very elegant machines*.

In summary, navigating the UI, we formed some initial interpretations for why a particular feature is negatively discriminative. In particular, nominals with complex adjectival phrases appear particularly susceptible to article omission errors by learners of English with L1s lacking articles. The example illustrates not just the usefulness of visualisation techniques for navigating and interpreting large amounts of data, but, more generally the relevance of features weighted by discriminative classifiers. Despite being superficial in their structure, POS ngrams can pick up syntactic environments linked to particular phenomena. In this case, the features do not just identify a high rate of article omission errors, but, importantly, a particular syntactic environment triggering higher rates of such errors.

	sentences%		MD:doc	
Language	f_1	f_2	f_1	f_2
all	23.0	15.6	2.75	2.73
Turkish	45.2	29.0	5.81	5.82
Japanese	44.4	22.3	4.48	3.98
Korean	46.7	35.0	5.48	5.31
Russian	46.7	23.4	5.42	4.59
Chinese	23.4	13.5	3.58	3.25
French	6.9	6.7	1.32	1.49
German	2.1	3.0	0.91	0.92
Spanish	10.0	9.6	1.18	1.35
Greek	15.5	12.9	1.60	1.70

Table 5.3: f_1/f_2 relations with MD errors for different L1s, where sentences% shows the proportion of sentences containing f_1/f_2 that also contain a MD.

5.5 Improving automated assessment

Herein, we present preliminary results on how visualisation of (FCE) discriminative features facilitates the identification of patterns that can help us further improve performance of automated assessment systems. More specifically, in the previous section we demonstrated how visualisation of the 18th most discriminative feature allowed us to form hypotheses on the dependence between syntactic phenomena and the presence of MD errors. Two POS ngrams (among others), RG_JJ_NN1 and VBZ_RG_JJ, involving the use of complex adjectival phrases, exhibited a high contrast in MD errors between learners from L1s with articles and without. Putting the two ngrams together yields the POS fourgram VBZ_RG_JJ_NN1, which corresponds to incorrect constructions in the data, for example, *is very clever idea*. Our hypothesis then is that adding this ngram to our best performing FCE model, presented in Chapter 4, Section 4.4.1, will further improve performance as it should be highly discriminative, in addition to RG_JJ_NN1 and VBZ_RG_JJ. The FCE feature-types include POS unigrams, bigrams, and trigrams, but not fourgrams; thus, the classifier cannot automatically identify discriminative features of this type. It is worth mentioning at this point that, in general, POS fourgrams as a feature type were not found to be discriminative during development of our AA models.

We refer to the best performing FCE model as the baseline, and investigate its performance when this particular POS fourgram is added on top of this system's feature set. Table 5.4 presents the results on the FCE texts from the examination year 2001. Although POS fourgrams, in general, did not improve AA-system performace, adding this specific sequence of POS tags as a feature (VBZ_RG_JJ_NN1) improves r by 0.008, getting closer to the upper bound on this evaluation measure, while the improvement is significant at $\alpha = 0.059$ using a one-tailed test. No effect was observed in ρ, which is expected since the baseline model is very close to the upper bound for this measure. Further, we calculated the average correlation between the AA system's predicted scores, the FCE grades and those provided by the four senior ESOL examiners who remarked the 2001 texts (Table 5.5). Again, the extended model improves over the baseline, which confirms our hypothesis that the POS fourgram is discriminative. The results illustrate

Features	r	ρ
Baseline	0.749	0.790
+VBZ_RG_JJ_NN1	**0.757**	**0.791**
Upper bound	0.796	0.792

Table 5.4: Performance on the exam scripts drawn from the examination year 2001.

Features	r	ρ
Baseline	0.727	0.736
+VBZ_RG_JJ_NN1	**0.732**	**0.738**

Table 5.5: Average correlation between the AA model, the FCE dataset grades, and 4 additional examiners on the exam scripts from year 2001.

that hypothesis formation through visualisation of features, and, in general, visual exploration of features, can also aid the development of AA systems (through identification of new discriminative textual cues), and further improve their performance. In other words, discriminative features can be interpreted in a meaningful way, and this, in turn, can be used to enhance automated assessment of text quality.

5.6 User evaluation

5.6.1 Experimental design and participants

As mentioned previously, we adopted a user-driven paradigm for the development of the visualiser based on the needs of an SLA researcher who acted as a design partner during the development of the tool and was eager to use and test it. To further assess as well as measure the effectiveness and efficiency of the EP visualiser, we conducted a small-scale controlled usability study as the first stage for eliciting overlooked requirements and getting quantitative and qualitative feedback to inform future developments. Evaluation of visual presentations, visualisation systems and, generally, of computer-based interfaces is a key component to ensure their quality, success, and adoption by the target user population — in our case, SLA researchers, teachers and assessors. For example, poor system usability may lead to low user effectiveness, increased errors in completing tasks, and consequently low adoption rates. We compared the EP visualiser against an existing CLC search tool, OpenInsight (described in more detail in Section 5.6.2), which is available to authors and writers working for CUP and to members of staff at Cambridge ESOL. OpenInsight is developed by CUP, who gave us permission to use the tool.[6]

Our main goal is to evaluate the EP visualiser as a candidate substitute for OpenInsight for searches relating primarily, but not exclusively, to linguistic discriminative features. OpenInsight is a strong baseline mainly due to its simplicity and ease of use. Additionally, it has been available to CUP and Cambridge ESOL for at least nine years and has been used to inform their work. The latter is important, as particularly Cambridge ESOL employees represent a target user population for the visualiser. As Hearst (2009) notes, "another way to bias the outcome of a usability study is to compare a new design against an unrealistic baseline system". In this experiment, we measure four main variables,

[6]We should, however, note that there are other alternative tools currently available as well.

task completion time, mouse events needed to complete a task, number of errors and user satisfaction. Mouse events are measured in terms of mouse clicks performed by participants, while user satisfaction is measured using questionnaires to assess subjective satisfaction for each tool. Although response time and mouse events do not necessarily reflect search success, they will give us an indication of the processes involved in using each system. Errors are assessed by counting, out of all tasks, the number of tasks users did not successfully complete.

Many factors can affect the usability studies of a system, whether it is a Graphical User Interface (GUI), which allows users to perform actions through graphical components, such as OpenInsight (see next section), a Visual User Interface (VUI), which investigates the mapping between visual presentations and the users' mental model, such as the EP visualiser, or a Text-based User Interface (TUI), which is typically based on text commands. These factors include user characteristics, such as experience, domain knowledge and cognitive skills, as well as the choice of tasks given to the users (Chen and Zhang, 2007; Hearst, 2009; Nielsen, 1993). Hearst (2009) notes that, when comparing a new interface against one with which users are familiar, it is commonly observed that users, most of the time, prefer the original one at first. For example, although GUIs have in general been identified to be superior to TUIs, research has shown that for expert users, a GUI interface may not always be preferred (e.g., Chen and Zhang, 2007; D'Ydewalle et al., 1995; Whiteside et al., 1985).

Four female subjects working at Cambridge ESOL's Research and Validation group volunteered to participate in our study.[7] Two of the participants' research involves second language acquisition and assessment and have prior knowledge in using OpenInsight, and the other two assist on group-related projects and are novice users. More specifically, two users had no experience with OpenInsight, one had prior experience in using basic system components (beginner), and the final one was an expert user with more than two years of experience with the tool.

Our study uses a within-subjects design, where each participant uses both OpenInsight and the EP visualiser. Such a design is common when comparing interfaces (Hearst, 2009). Advantages include the requirement for fewer subjects, compared to a between-subjects design, while variance between experimental conditions may also be systematic. On the other hand, this may introduce order effects, that is, the order in which the systems are presented to the users can have an influence on them and bias the results (see Chapter 2, Section 2.6.2 for more details). We used a blocked design to counterbalance for order effects, and each interface was randomly assigned to, and was the starting view for half of the participants.

Prior to evaluation, each participant was individually introduced to the systems' features and characteristics during a one-hour session for each tool. During these sessions, they were instructed to complete eight practice tasks, as directed by the evaluator, and received feedback about the accuracy of each action. This is a common methodology adopted in order for users to get familiar with the system and relax into the tasks (Hearst, 2009). The main evaluations, as well as the introductory sessions, were conducted in a quiet room in Cambridge ESOL's offices, mainly because there were difficulties in using OpenInsight outside Cambridge ESOL, in addition to it being easier for the subjects not to have to travel, at times that were most convenient for them, and lasted up to an hour

[7] In rigorous large-scale evaluation techniques participant characteristics, such as gender, should not be underspecified, as it can limit generalisability of the results.

per system. Further, the manuals for each tool were available to the participants while performing the tasks to ensure minimal memorisation required. During the main evaluation, participants were asked to complete two sets of actions, each one consisting of four tasks, while the study was also counterbalanced for task types. The tasks were similar across tools, but not identical, to minimise learning effects; they are described in detail in Section 5.6.3. To ensure that the interface evaluations were equivalent, we only tested functions that are available in both systems. Pilot testing preceded the main evaluation, where a non-domain expert working in NLP also participated, and allowed us to determine the clarity of the tasks, as well as to ensure the possibility of completion with each system.

Due to the difficulties in accessing OpenInsight from machines other than the ones in Cambridge ESOL, as well as in installing new software on those, we were unfortunately unable to run both systems on the same computer. OpenInsight is network-dependent, that is, search results depend on and are obtained through a network connection, in contrast to the EP visualiser, which accesses the underlying database locally. On the one hand, this allows us to evaluate OpenInsight in the exact same state as it is available to the users in their everyday environment. On the other hand, however, we do realise that this setup introduces confounding variables; our evaluation metrics though include number of errors, mouse events and questionnaires, in addition to task completion times.

5.6.2 OpenInsight: CLC search tool

OpenInsight is a GUI available to employees in CUP and Cambridge ESOL, providing access to the CLC through database queries, and is briefly described in Nicholls (2003). As mentioned earlier, Gram and Buttery (2009) have developed a command-line CLC search tool that provides the opportunity for a wide range of specialised searches, which we also extend and integrate as a key component in the EP visualiser. Nevertheless, we chose to use OpenInsight as, typically, graphical interfaces exhibit higher simplicity and usability rates compared to command-line tools. In what follows, we give a brief overview of OpenInsight's basic functionality.

Figure 5.5 shows OpenInsight's front-end. The button on the top left corner, containing a magnifying glass over a red tick, is used to begin searching the data. Selecting it enables the window presented in Figure 5.6, which allows for a wide range of searches. Users can choose the datasets they want to investigate, for example, the FCE and CPE exam scripts (named FCE_LNR and CPE_LNR respectively in the tool) and perform a normal or collocation search under *Search Options*, as well as specify the search query in the *Search string(s)* field at the bottom right. The *PoS* button next to it displays a list of the CLC errors tags and allows selection of and searching for specific occurrences. The *Filter* field at the top right can be used to apply Boolean filters with respect to various meta-data, such as language, age, grades and exam year, while the *Advanced* button below further restricts the results to be retrieved.

When the search is completed, a new window appears which allows users to examine instantiations of the search query. Figure 5.7 illustrates an example output for the term *actual*. The matching results are displayed one per line by default, and the matching word is centred and highlighted in red, while information relative to the data is displayed at the top of the window. Selecting the option *Stats* under the tools menu in Figure 5.5 enables a window that provides advanced searches within the matching results, such as

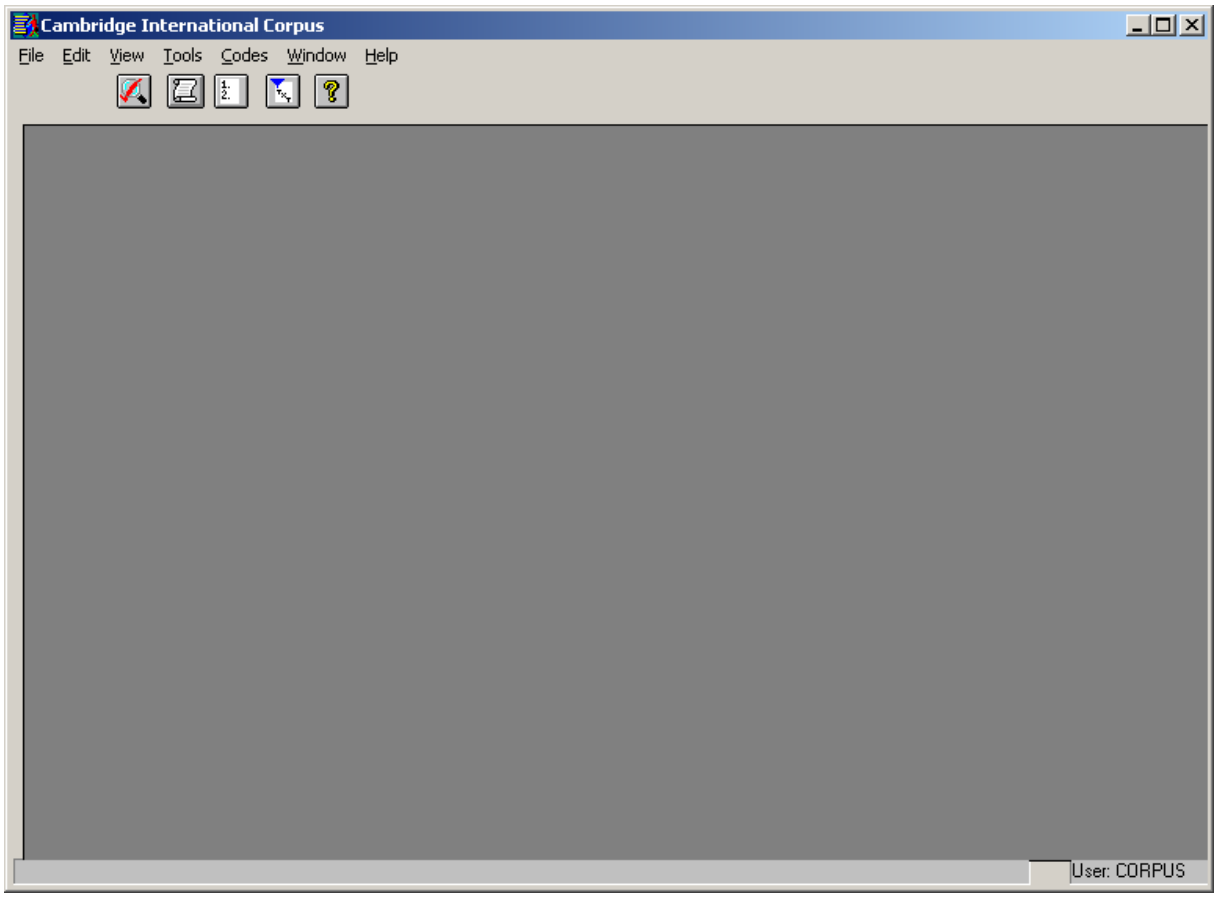


Figure 5.5: OpenInsight front-end: initial screen.

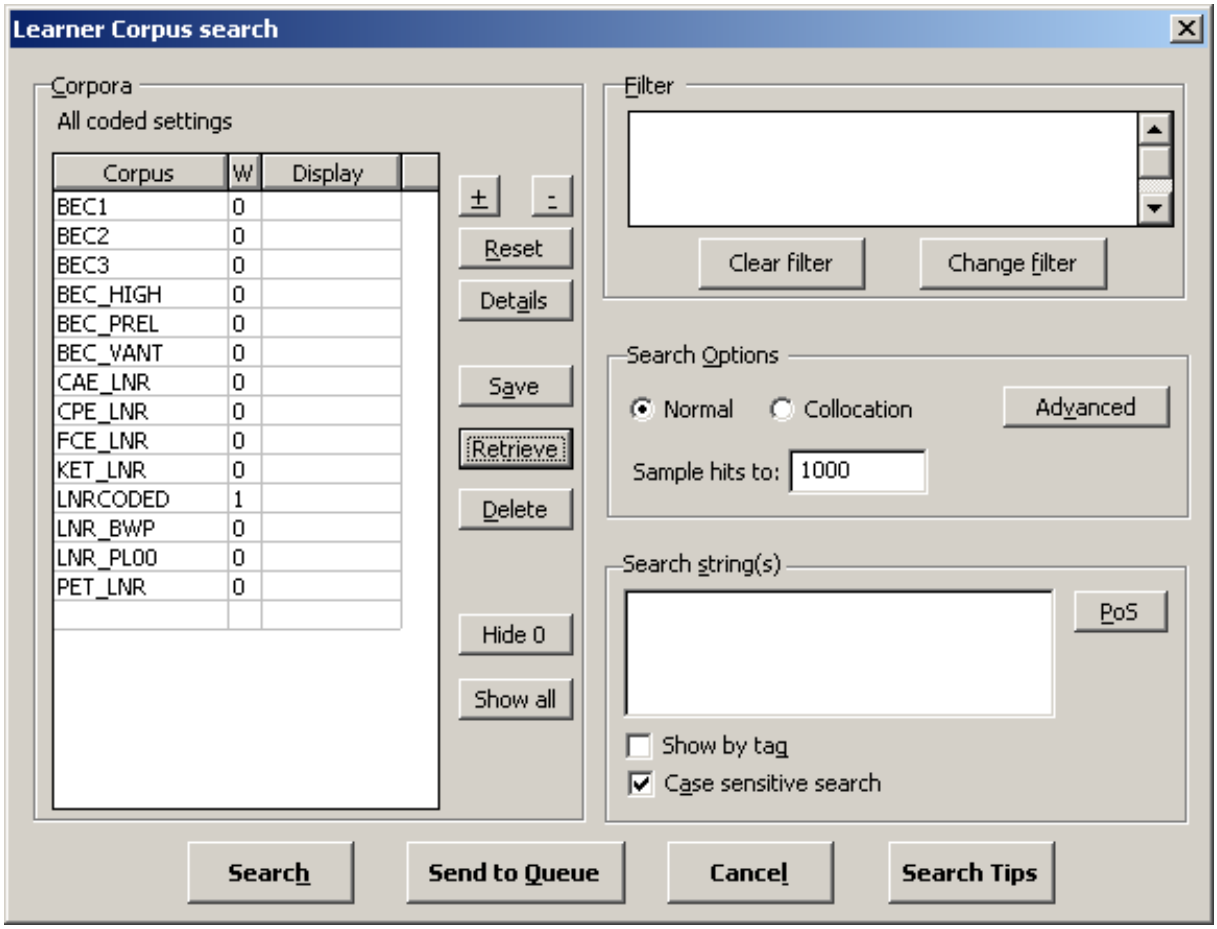


Figure 5.6: OpenInsight: starting a new search.

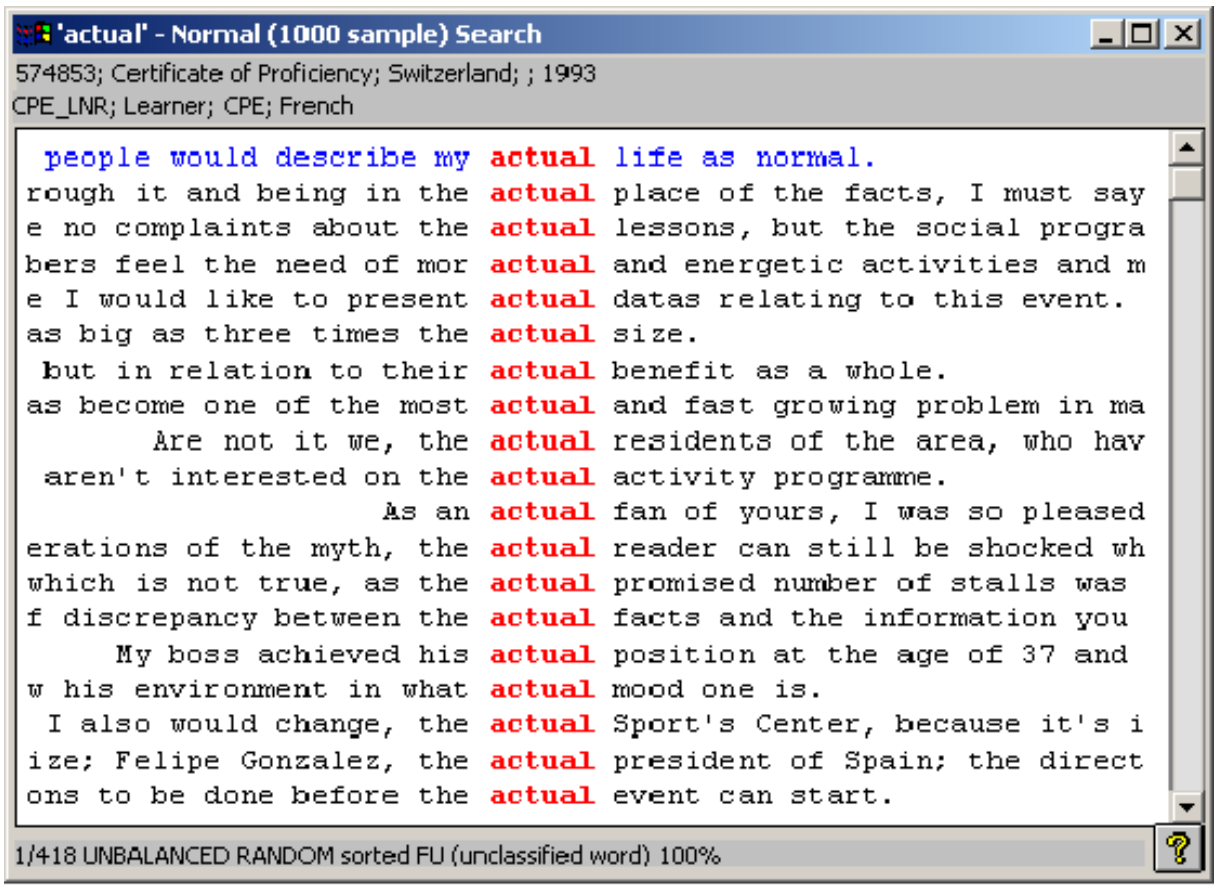


Figure 5.7: OpenInsight search results: the corpus browser window.

identifying the words or errors occurring in specific positions around the search term, and sorting based on frequency.

5.6.3 Tasks

During the main evaluation sessions, all participants completed two sets of tasks, each one consisting of four sub-tasks. The first set of tasks involved searches relating to discriminative features, while the second focused on general searches. Our primary goal was to evaluate the two systems with respect to flexibility in searching for discriminative features and related statistics; however, as the EP visualiser was developed with a focus on these features, whereas OpenInsight not, we also assessed them in a more general setting to identify whether one is superior to the other for other types of searches and examine the extent to which they may be complementary. A system supporting a wide range of functionalities would be a powerful tool for data-informed hypothesis creation in general.

To ensure that the comparison between systems was equivalent, we only tested functionality that is available in both. Since the automatically determined features are discriminative on the CLC FCE texts, the evaluation setup involved searching in this text collection.[8] As mentioned earlier, the tasks were similar across interfaces, but not identical, in order to minimise learning effects. Those relating to discriminative features were

[8]Adapting the EP visualiser to different sets of features and/or datasets is quite straightforward, though the version described here does not currently support this.

identified through a discussion with our main SLA user who guided the development of the visualiser, while the ones involving general searches were inspired by the search examples presented in Gram and Buttery (2009), focusing on extracting statistics with an emphasis on hypothesis formation. However, none of the tasks required that the participants generate hypotheses, and they were all simple and specific. Hypothesis creation is an inherently open-ended task, and this makes it difficult to use appropriate objective measures for its evaluation. As Zuk et al. (2006) note, high-level cognitive issues are hard to measure with quantitative user studies. The first set of tasks is presented below:

Set A: tasks related to discriminative features

1. Out of all word ngram features, find the most frequent among those that have a discriminative rank between [rank 1] and [rank 2] (where 1 represents the most discriminative).

2. Given the feature found in task 1, find one word ngram feature with which it co-occurs in the same sentence.

3. Given the feature found in task 1, find the most frequent error with which it co-occurs in the same sentence (you can just write down the error tag).

4. Given the feature found in task 1, find the total number of documents produced by Romance learners (in particular, Spanish, Portuguese, French and Italian) that contain it.

In this set, the first task was slightly different between versions to counterbalance for learning effects. More specifically, rank 1 and rank 2 were set to 50 and 150, or 500 and 600. We tried to introduce as much control as possible to the answers of the tasks to facilitate their similarity, though an erroneous answer may affect the results for subsequent tasks. For example, the rank 1 and rank 2 values were chosen so that in both cases, the candidate word ngrams were approximately the same in number, around forty. Further, we made sure that, in both cases, the most frequent word ngrams were positively discriminative features and of the same length – word unigrams. The correct answers to task 1 were highly frequent function words, *as* and *or*.

However, the EP visualiser was specifically built to answer those types of questions; thus, using these tasks in a comparison with OpenInsight, which was not developed with these goals in mind (though, as mentioned earlier, we explicitly chose functions that are available in both systems), may introduce a bias over our system. On the other hand, the development of a system that would allow users to perform feature-related searches in a fast, easy and flexible way was one of our main motivations; thus, such a comparison will allow us to assess this quantitatively and measure its potential superiority with respect to a tool that is currently available to the users. However, we also performed a more general assessment, and our second set of tasks emphasised searches that are not (directly) related to discriminative features, but are more generic. Such a study will allow us to identify the extent to which the tools posses complementary properties and to elicit overlooked requirements in the hopes of further advancing system development. These tasks are presented below (set B):

Set B: tasks related to general searches
1. Find the total number of learners (or, equivalently, documents) in the dataset.
2. Find the total number of [error type] errors in the dataset.
3. Find the average number of [error type] errors per learner in passing scripts only. To calculate this, you need to divide the total number of [error type] errors by the total number of documents (or, equivalently, learners) in passing scripts (you can just write down the two parts of the formula).
4. Out of all opportunities for error, how many result in [error type] error? To find this, divide the total number of [error type] errors by the total number of words in the dataset (you can just write down the two parts of the formula).

The error type selected for questions 2 to 4 was either a replace verb error (RV) or a replace noun error (RN). Answers to tasks similar to the first one, which do not necessarily vary between interfaces, are nevertheless system-dependent, and the search process is not likely to be biased when comparing the different tools, while users were also instructed to perform all the necessary steps to give an answer. During the evaluation sessions, users had access to system manuals, lists of discriminative features, and error tags and their description. Timing for each task started after the users had read it and ended as soon as they had completed it, while errors were not propagated for assessing the correctness of an answer.

5.6.4 Results

Descriptive statistics for the main variables of response time, click counts and errors are listed in Table 5.8. The mean response time was six times longer with OpenInsight compared to the visualiser on tasks related to discriminative features (task set A), while on generic tasks the differences are smaller with a mean completion time of 2.44 minutes with OpenInsight and of 1.62 with the visualiser. The number of mouse clicks in OpenInsight was relatively large compared to the visualiser and varied between a mean of 14 clicks on the second task set and 35 on the first. On the other hand, the mean counts for the visualiser exhibit more stability and stay around 5 clicks. Response time and mouse events do not necessarily reflect search success and are less critical compared to task accuracy. The mean number of errors is close to zero on set A with the visualiser, and around 0.3 on both sets with OpenInsight. However, user errors are higher when the visualiser is used to run task set B compared to set A, and 25% of the tasks are answered incorrectly.

In summary, the visualiser is much better on task set A, and OpenInsight is better on task set B than on task set A. The differences between the two systems are larger for set A, and these results support the hypothesis that the visualiser may be better suited for discriminative-feature searches, though the small differences on the second set (task set B) also suggest that it is as good as OpenInsight on generic searches. Given our small sample sizes though, we were unable to run significance tests. We also note at this point that task completion times may be confounded to some extent by the lack of control in computer characteristics.

		Time (min)		Click counts		Errors	
Task set	**System**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Set A	OI	6.38	3.48	35.62	20.20	0.31	0.48
Set A	EPV	1.06	0.54	5.12	3.30	0.06	0.25
Set B	OI	2.44	1.93	14.06	12.12	0.38	0.50
Set B	EPV	1.62	1.26	5.06	3.43	0.25	0.45

Table 5.8: Mean and standard deviation for task completion time, mouse event (click) counts and task accuracy when participants used OpenInsight (OI) or the EP visualiser (EPV) to complete task set A, which focused on searching discriminative features and related statistics, and task set B, focusing on generic searches.

	Overall time satisfaction		Answer confidence	
System	**Mean**	**SD**	**Mean**	**SD**
OI	4.00	1.41	4.00	1.15
EPV	6.50	0.58	5.00	0.82

Table 5.9: Mean and standard deviation for satisfaction scores of the overall amount of time it took each participant to complete both sets of tasks with OpenInsight (OI) or the EP visualiser (EPV), as well as confidence scores of their answers to the tasks. Scores range from 1 (strongly disagree) to 7 (strongly agree).

To enable further analyses, participants were asked to score on a scale from 1 (strongly disagree) to 7 (strongly agree) their overall satisfaction with the amount of time it took to complete both sets of tasks[9] and the confidence of their answers, as well as to select the tasks they found harder to complete while using the different systems. Their answers were elicited right after they had finished the experiments with a system. As presented in Table 5.9, users were more satisfied with the visualiser's overall completion time, which also increased their confidence with respect to their responses, though, again, we do not know the extent to which these differences are significant. Table 5.10 shows that most users found it hard to complete generic tasks with the visualiser and feature-related tasks with OpenInsight. Task A3, which asked users to find the error with which a feature co-occurs most frequently in the same sentence, was perceived to be the hardest task in set A when using OpenInsight. This may be due to the fact that users had to search for this themselves, in contrast to the visualiser, which automatically displays that information. Task 3 from set B, asking users to find the average number of a specific error type per candidate in passing FCE scripts, was the one identified by the users as the hardest to complete with the EP visualiser. This can be explained by the fact that our system does not explicitly display passing and failing texts, but rather only categorises them per grade, in contrast to OpenInsight.

It is interesting to note that these results do not fully reflect the tasks which most of the users answered (in)correctly (Table 5.11); although task B3 did cause the largest number of errors when using the visualiser, the results indicate this as the hardest task for OpenInsight too. In addition, A1 is the one that causes the second largest error count with

[9]This will allow us to see whether users are satisfied with the overall completion times, regardless of how fast or slow they were with each system.

System	A1	A2	A3	A4	B1	B2	B3	B4
OI	2	2	3	1	0	0	2	1
EPV	0	1	0	0	1	1	4	1

Table 5.10: Number of participants who marked specific tasks as being hard to answer with either OpenInsight (OI) or the EP visualiser (EPV). A1 represents task 1 from set A, A2 represents task 2 from set A, and so on.

System	A1	A2	A3	A4	B1	B2	B3	B4
OI	3	2	0	0	0	1	4	1
EPV	0	0	0	1	0	1	2	1

Table 5.11: Number of participants who answered specific tasks incorrectly with either OpenInsight (OI) or the EP visualiser (EPV). A1 represents task 1 from set A, A2 represents task 2 from set A, and so on.

OpenInsight, while A3 was answered correctly by all users. Errors in A1 can be explained by the task's requirement to search through discriminative features, a functionality which OpenInsight does not directly support, thus making it error-prone for this type of search.

In addition to the above, each participant's satisfaction with respect to various system aspects was measured using the 'Usefulness, Satisfaction, and Ease of use' (USE) questionnaire (Lund, 2001). The questions are constructed based on a seven-point Likert scale, and users rate their agreement with each statement in a scale from 1 to 7, where 1 represents 'strongly disagree' and 7 'strongly agree'. The statements focus on four dimensions, usefulness, ease of use, ease of learning (which is strongly related to ease of use), and satisfaction, and are presented in detail in Appendix G. Subjects rated the EP visualiser higher by mean scores for each of the USE dimensions. Table 5.12 summarises these findings. Although the visualiser involves the use of more complex functions compared to OpenInsight, the mean values for ease of use and learning are consistently higher.

As soon as the users completed both experimental conditions, they were given another set of questions to allow us to further assess their previous responses. Although the visualiser appears to be as good as, and in some cases better than, OpenInsight, the success of a tool depends on its adoption by the users. Participants were asked to compare the two systems and indicate their preference in different scenarios. This evaluation consisted of the following four questions:

Q1: Which system do you prefer for tasks similar to set A?

Q2: Which system do you prefer for tasks similar to set B?

Q3: Overall, which system do you prefer?

Q4: Choose one of the following preference relationships among systems:

(a) EP visualiser is better than OpenInsight.

(b) EP visualiser is as good as OpenInsight.

(c) EP visualiser is worse than OpenInsight.

	Usefulness		Ease of use		Ease of learning		Satisfaction	
System	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
OI	3.62	0.34	3.95	0.78	4.94	0.72	3.71	1.11
EPV	5.16	0.89	5.23	0.50	5.50	0.41	5.50	0.36

Table 5.12: Mean and standard deviation for subjective satisfaction scores, measured using the USE questionnaire, when participants used OpenInsight (OI) or the EP visualiser (EPV) to complete all tasks. Satisfaction scores range from 1 (strongly disagree) to 7 (strongly agree).

System	Q1	Q2	Q3	Q4
OI	1	1	1	1 (as good as)
EPV	3	3	3	3 (better than)

Table 5.13: Number of participants who selected either OpenInsight (OI) or the EP visualiser (EPV) with respect to questions 1 to 4 above (Q1–Q4).

As presented in Table 5.13, the majority of the participants expressed a preference for the visualiser. Three out of four users would select this system for both types of tasks, as well as overall, and one indicated the visualiser being as good as OpenInsight. This can be seen as a success for a tool that participants had only briefly experienced. It is interesting to note that the users who preferred OpenInsight with respect to each of these questions are all different.

Finally, we wanted to investigate the extent to which participants find feature visualisation and their graph display useful through the following questions:

Q5: Visualisation of discriminative features is useful.

Q6: Graphs are useful for visualising features.

Q7: Graphs are useful for visualising feature relations.

Q8: Graphs are easily interpretable for displaying features and their relations.

The first question, Q5, focuses on the usefulness of visualising discriminative features. Our hypothesis is that visualisation of such features can offer insights into assessment and the linguistic properties characterising learner levels, which was supported by the case study presented in Section 5.4. Although the question does not exemplify their application, the results presented in Table 5.14 suggest that users highly recognise the usefulness in visualising them. Further, participants tended to agree that graphs are useful for visualising features and the relations between them, and that their visual interpretation is straightforward. The questions were given to the users when they had completed the experiment with the EP visualiser.

5.6.5 User feedback

During the evaluation sessions, participants tended to think aloud and voice their thoughts, confusions and preferences, which was also encouraged by the evaluator. Additionally, the USE questionnaire consisted of sections asking them to list the most negative and positive

Q5		Q6		Q7		Q8	
Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
6.25	0.50	5.75	0.96	6.75	0.50	6.25	0.96

Table 5.14: Mean and standard deviation for subjective scores given to questions 5 to 7 (Q5–Q7) regarding discriminative-features visualisation. Scores range from 1 (strongly disagree) to 7 (strongly agree).

aspects of each tool, while the questions focusing on user preferences (questions Q1-Q4 in the previous section) further requested users to provide a short explanation on their decision. These also gave us valuable feedback for future development and are presented below for each system.

OpenInsight

The most positive comments and aspects of OpenInsight include the following, as worded by the users: "a lot of data and statistics are easily available", "not too difficult to pick up", "it is user friendly", "it is straightforward and simple to use, once you know it", "you can accomplish a range of tasks", "fairly easy to fully master", "you can do the same thing in various ways", "you can apply various filters to the data", "I like the KWIC view" (that is, 'Key Word In Context' view, in which the search terms are highlighted and centred, while displaying their context on either side).

The most negative comments and aspects of OpenInsight include: "fairly rigid system", "it is not very intuitive", "aesthetically not very attractive", "slow on most types of searches", "requires many steps to complete the tasks (e.g., applying more than three native-language filters)", "filters are hard to apply". Additionally, some users got frustrated with the system's speed, especially during completion of task set A relating to discriminative features.

As we can see, some positive and negative aspects are contradictory, and these differences mainly arose because of the users' different levels of familiarity with the system. Moreover, aesthetics and tasks seem to have an effect on users' perception, which can partly be seen from the negative aspects listed (discussed further in the next section). In general, users tended to agree on the simplicity of the system, in addition to its enhanced functionality.

EP visualiser

The most positive comments and aspects of the visualiser according to the users include: "fast and responsive", "quick to search", "[graphs are] pretty", "it is visually pleasing", "graphics, diagrams, colour (I love colour)", "felt like I knew what I was doing", "powerful in terms of the types of searches that can be run", "allows break-down of results across many variables (e.g., L1s, grades)", "I like [the] division of [data into] grades", "intuitive once you've mastered it", "graphs is a new technique, but I can see the advantages".

The most negative aspects of this tool according to user feedback include: "if you display many features, [the] graph becomes very busy", "large numbers are not separated by comma" (when the visualiser displays statistics), "cannot filter by some criteria, e.g., pass/fail scripts", "some functionality is not intuitive (e.g., shift+mouse-click to display feature relations)", "doesn't show how long you have to wait for the results to be retrieved"

(in contrast to OpenInsight which uses a progress bar), "I'm not used to clicking on the display for searches" (the visualiser supports selection and search of features primarily using the mouse), "I found the window for selecting features confusing" (see Figure 5.2), "didn't quite get how many components work", "quite complex for the occasional user wishing to carry out simple tasks", "does not have a 'clear all' button [to de-select selected features] so you are sure future searches have no bias", "graphs are moving", "I can't decide about the need for the features[/nodes] to dynamically move around, but otherwise the graph visualisation is very good". The last comments in particular highlight the system functions that most users found confusing. The graphs render dynamically, and this was the reason why three out four participants thought that the system was busy and was actually processing some information in the background, in which case they may have had to wait until it gets ready to be used. Furthermore, using the feature-selection panel to choose the features users want to investigate, in combination with them having to also explicitly select the nodes to be able to browse the texts instantiating them, was a difficult process for the users, who assumed that during search the system will automatically retrieve texts instantiating all the currently displayed features. Similar to OpenInsight, but to a lesser extent, there were some concerns related to the system's speed.

In general, users' reactions followed a similar pattern in that the visualiser is more complex and not as straightforward, although questionnaire ratings were higher compared to OpenInsight (discussed in the next section). On the other hand, they commented favourably on aesthetics and speed. Further improvements to the system were also suggested, such as adding an option that would automatically show only those feature–error relations in the left panel in Figure 5.1, that correspond to the features currently displayed in the central panel. Another user suggested that it would be useful to directly display a feature's frequency, in addition to visualising it through the node's size (as mentioned in Section 5.3, the visualiser currently adjusts the size of the nodes based on the frequency of the features they represent), as well as to add a 'select-and-go' functionality in the search-results window (see Figure 5.4) that would automatically highlight the tokens tagged inside errors in the texts. Regarding the graphs, one participant proposed to allow users to select clusters from the graphs displayed for further, separate investigation.

The questionnaires regarding system comparison and user preferences elicited the following responses: "[the] EP tool was quicker, newer and easy to learn but I forgot one step. Could learn to use the system adequately well in a short time though", "OpenInsight is better for lexical analysis, when you need to explore 2 or more combinations of lexical items", "OpenInsight is more straightforward", "the tools are complementary since they're built for different things", "confidence [regarding answers to the tasks] will probably increase [for both systems] if used for longer". Interestingly, one of the users valued newness of the visualiser, while the same participant also characterised OpenInsight as a museum: "no matter how hard you try and how many times you go around it, you wouldn't be able to see everything, in contrast to the visualiser".

Finally, one of the subjects found the tools complementary to each other given their different purposes. Despite the availability of search tools to the target user population, none currently directly supports feature-related searches, though some are flexible enough to be employed this way too, and our study investigated these differences quantitatively. On generic types of searches it was found to be equally as good as an existing system, though, as discussed in the next section, studies focusing on specific system functionalities

and/or tasks may seem too narrow for comparing one tool against another. Designing further studies with a larger range of tasks and/or combining the best properties from each tool would be interesting avenues for future research.

5.6.6 Discussion

Overall, participants completed tasks faster using the EP visualiser compared to OpenInsight, made fewer errors and needed fewer mouse-clicks on average, and these were also reflected in subjective satisfaction scores collected through a series of questionnaires. The results indicate that the visualiser has a high usability rating, particularly for searches related to discriminative features, in terms of effectiveness, efficiency and user satisfaction, while it is as good as OpenInsight on more generic tasks as measured by the mean number of erroneous responses, though we were not able to assess these with significance tests. Further, users tended to agree on our hypothesis that feature visualisation is useful and graphs are a good means to display them and to facilitate their interpretation. Results from our study and participants' feedback during the evaluation sessions and on positive and negative aspects of the systems gave us several interesting directions for future work.

On the other hand, our lack of control on computer characteristics may have introduced confounding variables to the experiment, especially for the timing results. System differences may be affected by machine loads, architecture, performance, and so on, though the tools compared in this study are inherently different, in that one is network-based, whereas the other is not. Additionally, our experiment was small-scaled and focused on getting initial quantitative results with respect to the visualiser's effectiveness in primarily searching discriminative features, reflecting its comparison to a different system to facilitate future development. A large pool of participants whose characteristics are also controlled is necessary to confirm these findings and investigate their generalisability. Furthermore, controlled experiments focusing on specific system functionalities and tasks may seem too narrow for comparing one tool against another. Task difficulty and participant differences, such as expert knowledge in a topic and cognitive skills, can have effects on the results (Hearst, 2009), while previous research has shown that different types of tasks may lead to contradictory results (see, for example, Staggers and Kobus, 2000, for a discussion on results regarding the superiority of GUIs). Moreover, Ben-Bassat et al. (2006) have found an interdependence between perceived aesthetics and usability in questionnaire-based assessments, and showed that users' preferences are not necessarily based only upon performance, while aesthetics are considered too. As Hearst (2009) notes, "In some cases it is desirable to separate effects of usability from those of branding or of visual design". Of course, this largely depends on the study's goals. All the above are issues that need to be further investigated and factored into future experimental designs to test the generalisability of the results.

5.7 Conclusions

We have demonstrated how visual presentation of machine-learned features can support SLA research and point to a range of interesting patterns in learner data, which can further facilitate informed development and improvement of AA systems. More specifically, we integrated highly-weighted discriminative linguistic features into a graph-based visualiser to support an in-depth analysis. We presented a coordinated approach, using search

tools along with a graph visualisation, combined with easy access to the underlying raw data, and described a case study of how the system allows SLA researchers to investigate the data and form hypotheses about intermediate-level learners. This analysis, in turn, allowed us to improve the performance of the FCE AA system significantly ($\alpha = 0.059$). Although the usefulness of the EP visualiser should be confirmed through more rigorous evaluation techniques, such as longitudinal case studies (Munzner, 2009; Shneiderman and Plaisant, 2006) with a broad field of experts, our initial usability studies are encouraging and show that the visualiser has high ratings. Parameter control is essential in system-evaluation studies to ensure minimisation of confounding variables and generalisability of the results. Careful selection and evaluation of the tasks themselves, pilot testing, and the use of various evaluation metrics and questionnaires that would allow assessment from several perspectives, need also be factored into experimental designs.

Future work should include development, testing and evaluation of the UI with a wider range of users, and be directed towards investigation and evaluation of different visualisation techniques of machine-learned or -extracted AA features that support hypothesis formation about learner grammars, and more generally, facilitate knowledge discovery. Evaluation metrics is important to be further investigated and extended so as to assess generated hypotheses using domain experts, as well as the extent to which such systems can inform AA-system development. Though several approaches have been proposed for linguistic visualisation (see Chapter 2, Section 2.6.1 for an overview), our work differs by using visualisation as a search tool for hypothesis generation, as well as to identify new discriminative features that further improve performance of automated assessment systems.

chapter 6

Conclusion

The main goals of this work have been to develop automated assessment models of English-learner writing, and investigate their internal characteristics in order to support second language acquisition research, as well as to facilitate the development of AA systems and improve their performance.

We have addressed automated assessment as a supervised discriminative machine learning problem, investigated various aspects of text quality, and released the first publically available shared dataset for training and testing such systems and comparing their performance. We have showed experimentally that generic approaches can achieve performance close to the upper bound, as defined by the level of agreement between human examiners, whilst also having the advantage of requiring smaller sample sizes and representing consistent 'marking criteria' regardless of the prompts. Further, they are less likely to need re-training or tuning for new prompts or assessment tasks.

In Chapters 3 and 4, we approached the automated writing assessment task from two different perspectives, linguistic competence, focusing on lexical and grammatical properties, as well as errors committed and discourse coherence and cohesion. We presented state-of-the-art results, identified new techniques that outperform previously developed ones, and performed a systematic assessment of several models and features. We examined model generalisation to different learner corpora and observed lexical and POS ngrams to arise prominently in our models, though in different forms, GR complexity measures extracted using the RASP system and the error rate estimated using a large background corpus. An interesting avenue for future work would be to explore feature types that provide a more invariant framework, though perhaps at the cost of near optimal performance. Models suggested in previous coherence research had a minimal effect in our evaluation, and this suggests that learner data and/or framing the task as a scoring problem is a distinct subcase of coherence assessment. Further, we examined and to some extent addressed validity issues of automated assessment systems, and, more specifically, their robustness to subversion by writers who understand something of their workings. Surprisingly, there is very little published data on the robustness of existing systems, although this is critical for their deployment.

Although in terms of their output automated assessment models simply return a score, implicit in its computation is the identification of positive and negative discriminative features that contribute to its calculation. In Chapter 5, we used visualisation techniques to shed light on how automated assessment models function and inspect the features they yield as the most predictive of a learner's level of attainment. We built a visual

user interface which aids the development and partially automates hypothesis formation about learner grammars, as demonstrated by a case study, as well as the identification of new discriminative features. The visualiser supports exploratory search over a corpus of learner texts using directed graphs of automatically determined linguistic features discriminating between passing and failing exam scripts, while a preliminary small-scale user study demonstrated its high usability ratings. To the best of our knowledge, this is the first attempt to visually analyse and perform a linguistic interpretation of automatically determined features that characterise learner English, as well as to demonstrate how this, in turn, can inform system development and improve the model's performance. Future work could usefully be directed towards identifying ways for (automatically) providing feedback to students based on positive and negative discriminative features as part of self-assessment and self-tutoring systems.

To date, automated assessment systems have both been incorporated in writing assessment (for example, e-Rater, Attali and Burstein, 2006) and used as instructional tools in classrooms (for example, Criterion, Burstein et al., 2003). Although a large body of research has experimentally demonstrated that automated assessment models can produce scores indistinguishable from human raters, the range of techniques applied to the task is developing, with innovative methodologies introduced to accommodate higher-order qualities. The extent to which such models are used operationally makes the investigation of the accuracy, robustness and transparency of automated assessment systems a research priority. Several other issues, such as the use of more 'sophisticated' evaluation measures relating to their internal characteristics, comparisons against external criteria, and construct validity, which relates to the conceptual fit between what is intended to be measured and what is actually being measured, need to be further factored into their design. As the technology keeps advancing, the performance of current automated assessment systems should be considered a baseline rather than an upper bound.

APPENDIX A

CLC ERROR TAXONOMIES

AG	Agreement error
AGA	Anaphora agreement error
AGD	Determiner agreement error
AGN	Noun agreement error
AGV	Verb agreement error
AGQ	Quantifier agreement error
AS	Agreement structure error
C	Countability error
CD	Wrong determiner because of noun countability
CE	Complex error
CL	Collocation or tautology error
CN	Countability of noun error
CQ	Wrong quantifier because of noun countability
DA	Derivation of anaphor error
DC	Derivation of link word error
DD	Derivation of determiner error
DI	Incorrect determiner inflection
DJ	Derivation of adjective error
DN	Derivation of noun error
DQ	Derivation of quantifier error
DT	Derivation of preposition error
DV	Derivation of verb error
DY	Derivation of adverb error
FA	Wrong anaphor form
FC	Wrong link word form
FD	Incorrect determiner form
FJ	Wrong adjective form
FN	Wrong noun form
FQ	Wrong quantifier form
FT	Wrong preposition form
FV	Wrong verb form

FY	Wrong adverb form
IA	Incorrect anaphor inflection
ID	Idiom wrong
IJ	Incorrect adjective inflection
IN	Incorrect noun inflection
IQ	Incorrect quantifier inflection
IV	Incorrect verb inflection
IY	Incorrect adverb inflection
L	Inappropriate register
M	Missing error
MA	Missing anaphor
MC	Missing link word
MD	Missing determiner
MJ	Missing adjective
MN	Missing noun
MP	Missing punctuation
MQ	Missing quantifier
MT	Missing preposition
MV	Missing verb
MY	Missing adverb
NE	No error
R	Replace error
RA	Replace anaphor
RC	Replace link word
RD	Replace determiner
RJ	Replace adjective
RN	Replace noun
RP	Replace punctuation
RQ	Replace quantifier
RT	Replace preposition
RV	Replace verb
RY	Replace adverb
S	Spelling error
SA	Spelling American
SX	Spelling confusion
TV	Incorrect tense of verb
U	Unnecessary error
UA	Unnecessary anaphor
UC	Unnecessary link word
UD	Unnecessary determiner
UJ	Unnecessary adjective
UN	Unnecessary noun
UP	Unnecessary punctuation
UQ	Unnecessary quantifier
UT	Unnecessary preposition
UV	Unnecessary verb
UY	Unnecessary adverb

W	Word order error
X	Incorrect negative formation

APPENDIX B

EXAMPLE FCE SCRIPTS

Example script — error annotation has been removed:

```
<learner><head sortkey="TR252*0100*2000*01">
  <candidate>
     <personnel>
        <language>Spanish</language>
        <age>16-20</age>
     </personnel>
     <score>28.0</score>
  </candidate>
  <text>
     <answer1>
       <question_number>1</question_number>
       <exam_score>3.2</exam_score>
         <coded_answer>
          <p>Dear Helen:</p>
          <p>I've recived your letter and I am pleased to have won
             because I needed some days to relax myself and to leave
             the city, which is very extressing.</p>
          <p>I only can travel on July because I am working in an
             office and I must ask my boss for a holiday and it's the
              mounth he can give me, so I hope it isn't a problem for
              you.</p>
          <p>To spend the two weeks I would prefer to live in tents,
             that's in my opinion a way to be nearer the enviroment
             and the animals, although I don't mind living in log
             cabins.</p>
          <p>In spite of liking all the sports you wrote in your
             letter I am only good at climbing and sailing, because I
              am used to practise them with my father since I was a
             child.</p>
          <p>I have got a doubt, it is not very important, but I
             would like to know if I need any money or all is payed,
             the kind of clothes I should wear, if the weather is
             good or bad... and all the extra information you can
             send me.</p>
```

```
          <p>Thank you very much for the prize.</p>
          <p>Yours faithfully.</p>
          <p>Eliza</p>
         </coded_answer>
     </answer1>
     <answer2>
       <question_number>3</question_number>
       <exam_score>3.3</exam_score>
         <coded_answer>
          <p>Why people like so much go shopping?</p>
          <p>In my opinion the majority of the people that go
             shopping are women who must buy food and other kind of
             things for her family, so for them, shopping is a very
             boring activity that is repeated everyday and that drive
              them crazy when they have to choose the best product
             and the cheapest price to save some money.</p>
          <p>Othrewise, there are people who think the opposite; such
              as: children, teenagers, men; but not always, because
             they go to buy hardly ever, only when they need clothes
             or something for their job or school. In that way
             shopping can be as funny as you want, although if you do
              it very often, it will be as an obligation in the
             future.</p>
          <p>To sum up, all you make in small cuantities is good and
             funny, but don't encrease them if you don't want to feel
              uncomfortable.</p>
         </coded_answer>
     </answer2>
  </text>
</head></learner>
```

Example script — error annotation has been retained:

```
<learner><head sortkey="TR798*0100*2000*01">
  <candidate>
     <personnel>
        <language>Chinese</language>
        <age>16-20</age>
     </personnel>
     <score>24.0</score>
  </candidate>
  <text>
     <answer1>
       <question_number>1</question_number>
       <exam_score>3.2</exam_score>
         <coded_answer>
          <p>Dear Mr Ryan<NS type="RP"><i>.</i><c>,</c></NS></p>
          <p>Thanks for <NS type="DD"><i>you</i><c>your</c></NS>
             letter. I am so <NS type="RJ"><i>exciting</i><c>excited
             </c></NS> that I have won the first prize. I will give
             you all <NS type="MD"><c>the</c></NS> information you
```

```
      need and ask some questions.</p>
    <p>I <NS type="TV"><i>could</i><c>can</c></NS> only travel
      <NS type="RT"><i>on</i><c>in</c></NS> July. As you know,
       I am a student and the nearest holiday is <NS type="MD
      "><c>the</c></NS> summer holiday. But I have booked a
      flight <NS type="UT"><i>to</i></NS> home at the
      beginning of <NS type="S"><i>Auguest</i><c>August</c></
      NS>. And also I would like to go <NS type="RT"><i>on</i
      ><c>in</c></NS> summer.</p>
    <p><NS type="MT"><i>The</i><c>Concerning the</c></NS>
      accommodation<NS type="MP"><c>,</c></NS> I would like to
       <NS type="RV"><i>live</i><c>stay</c></NS> in tents<NS
      type="RP"><i>. Because</i><c>, because</c></NS> I <NS
      type="TV"><i>never <NS type="RV"><i>live</i><c>stay</c
      ></NS></i><c>have never stayed</c></NS> in <NS type="MD
      "><c>a</c></NS> <NS type="AGN"><i>tents</i><c>tent</c></
      NS> before. I think it <NS type="TV"><i>is</i><c>will be
      </c></NS> great and I want to try it.</p>
    <p>I like doing sports. I would like to play basketball and
       golf when I am at the Camp. I play basketball a lot and
       I am a member of our college <NS type="RN"><i>term</i><
      c>team</c></NS>. But I am not very good at golf.</p>
    <p>And also I want to ask some questions. What clothes
      should I <NS type="TV"><i>taken</i><c>take</c></NS>? How
       much money should I <NS type="TV"><i>taken</i><c>take</
      c></NS>? And how <NS type="TV"><i>could</i><c>can</c></
      NS> we meet at the airport? I am looking forward <NS
      type="MT"><c>to</c></NS> your reply.</p>
     <p>Yours sincerely.</p>
    </coded_answer>
</answer1>
<answer2>
  <question_number>2</question_number>
  <exam_score>2.3</exam_score>
    <coded_answer>
     <p>As our class is going to <NS type="RV"><i>mark</i><c>
      make</c></NS> a short video about daily life at college,
       I <NS type="TV"><i>write</i><c>am writing</c></NS> this
       report to suggest some lessons and activities which
      should be filmed.</p>
     <p>1. English lesson. Because <NS type="UQ"><i>all</i></NS>
       students in <NS type="MD"><c>the</c></NS> English class
       are from all over the world<NS type="RP"><i>. We</i><c
      >, we</c></NS> can talk <NS type="MT"><c>about</c></NS>
      everybody's <NS type="FN"><i>feeling</i><c>feelings</c
      ></NS> <NS type="MT"><c>about</c></NS> living and <NS
      type="FV"><i>study</i><c>studying</c></NS> <NS type="RT
      "><i>at</i><c>in</c></NS> a foreign country.</p>
     <p>2. Computing lesson and computer room. <NS type="S"><i>
      Nowdays</i><c>Nowadays</c></NS> <NS type="MD"><c>the</c
```

```
></NS> <NS type="RP"><i>internet</i><c>Internet</c></NS>
 <NS type="RV"><i>makes</i><c>brings</c></NS> us closer
and closer. We can get all <NS type="UA"><i>what</i></NS
> we want on <NS type="MD"><c>the</c></NS> <NS type="RP
"><i>internet</i><c>Internet</c></NS>. It's one of the
most important things in our life now.</p>
<p>3. <NS type="S"><i>Liabrary</i><c>Library</c></NS>. We
not only borrow books from <NS type="MD"><c>a</c></NS> <
NS type="S"><i>liabrary</i><c>library</c></NS> but also
study at <NS type="MD"><c>a</c></NS> <NS type="S"><i>
liabrary</i><c>library</c></NS>. <NS type="MD"><i>
Library</i><c>The library</c></NS> is very important in
our daily life.</p>
<p>4. Canteen. <NS type="RP"><i>Everyday</i><c>Every day</c
></NS> we go to <NS type="MD"><c>the</c></NS> canteen <
NS type="FV"><i>have</i><c>to have</c></NS> lunch, no
matter <NS type="MC"><c>whether</c></NS> you <NS type="
TV"><i>bought</i><c>buy</c></NS> food from there or you
take your own food.</p>
<p>5. Football. What do you do after class? <NS type="DN"><
i>Joging</i><c>Jogging</c></NS> or <NS type="UV"><i>
doing</i></NS> some sports? You can't forget football.</
p>
<p><NS type="AGA"><i>These</i><c>This</c></NS> <NS type="
AGV"><i>are</i><c>is</c></NS> what I think should be
filmed. If any of you have other <NS type="FN"><i>
suggestion</i><c>suggestions</c></NS>, we can discuss <
NS type="MA"><c>this</c></NS> again. But I think these
five <NS type="AGN"><i>lesson</i><c>lessons</c></NS> or
activities are <NS type="MD"><c>the</c></NS> most common
 in our daily life at college.</p>
</coded_answer>
</answer2>
</text>
</head></learner>
```

APPENDIX C

EXAMPLE IELTS SCRIPT

Example script[1] — error annotation has been removed:

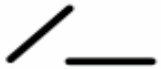
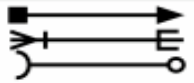
```
<learner>
  ...
  <text>
     <answer1>
         ...
         <coded_answer>
          <p>The table shows the percentage of ...</p>
          <p>The highest proportion (75%) was in ... students in
             Education. The same category had high percentage of
             student who were employed in previous job. However, the
             lowest proportion accured in ...</p>
          <p>Other area, ... had the best figures ...</p>
          ...
         </coded_answer>
     </answer1>
     <answer2>
         ...
         <coded_answer>
           <p>It is important that people can be producing the
              excellent outcome ...</p>
           <p>Furthermore, it is no controversial between colleagues
              ...</p>
           ...
         </coded_answer>
     </answer2>
  </text>
  ...
</learner>
```

[1] As IELTS examination scripts are not publically available, we have only reproduced a small amount of information.

appendix D

Properties and best uses of visual encodings

Properties and Best Uses of Visual Encodings

Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	**yes**	**infinite**	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	**optional** (alphabetical or numbered)	**infinite**	Good	Good	Good	Good
	length	**yes**	**many**	Good	Good		
	size, area	**yes**	**many**	Good	Good		
	angle	**yes**	medium/few	Good	Good		
	pattern density	**yes**	few	Good	Good		
	weight, boldness	**yes**	few		Good		
	saturation, brightness	**yes**	few		Good		
	color	no	few (< 20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	**infinite**			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	**yes**	few		Good		

Noah Iliinsky • ComplexDiagrams.com/properties • 2012-06

appendix E

Examples of outlier scripts

Modification 1(a): randomly order word unigrams within a sentence:

```
<p>Sir/Madam Dear ,</p>
<p>I with newspaper writing I which am your local ago read in a days
   advertisement reference two to .</p>
<p>weekend which was the my would this to took I dissatisfaction , ,
   As activities like your by to organised I article part express
   college in .</p>
<p>that the your advertisement course ill afternoon , had the First
   be last to because but , of cancelled course the mentions teacher
   the was cancelled , only all .</p>
<p>you local history Secondly registered body mention the that course
    for , no .</p>
<p>this course like I that attended would you 32 to to So inform .</p
   >
<p>is and very Thirdly you courses the as were , as far painting ,
   popular say what photography true they .</p>
<p>courses the in interested kind are that for that being of reason
   people these popular think I so these activities very is .</p>
<p>was coteporary and the Apart this was from the very experienced
   teachers equipment particullarly .</p>
<p>your cost the weekend mentions Furthemore article over 100 that ,
    .</p>
<p>not That true is .</p>
<p>The lunch cost was only 60 including .</p>
<p>so mention than term , activities run but sooner next be it
   college to , weekend not another you is the will least Last
   planning .</p>
<p>I disappointing will to the you tell another lit to , would you my
    kind which not thank write as I was and article would mention
   letter like weekend , would to as you to I like you the for you
   that bud Finaly attention like to truth .</p>
<p>Yours faithfully ,</p>
```

Modification 1(b): randomly order word bigrams within a sentence:

```
<p>Dear Sir/Madam ,</p>
<p>days ago your advertisement writing with which I a local newspaper
    two I am read in reference to .</p>
<p>was organised by the college , , which your article activities
   weekend express my As I took part dissatisfaction to in this I
   would like to .</p>
<p>all , had to was ill the course , because your advertisement the
   course mentions that the last First of be cancelled , but the
   teacher cancelled only afternoon .</p>
<p>body registered that no local history you mention Secondly , for
   the course .</p>
<p>to this So I 32 attended would like to inform you that course .</p
   >
<p>as far very popular as the photography courses painting and say ,
   they were what you Thirdly , is true .</p>
<p>courses so interested in I think reason for these kind are very
   being these of activities popular is that people that the .</p>
<p>teachers was and the this the very experienced equipment was
   particullarly coteporary Apart from .</p>
<p>mentions that the weekend Furthemore , cost over 100 your article
    .</p>
<p>not true That is .</p>
<p>was only 60 The cost including lunch .</p>
<p>activities weekend be sooner next term , so planning to run
   another college is than you , the Last but not least it will
   mention .</p>
<p>would like weekend was I would my letter that the to write I would
    to thank disappointing as which will tell you like to like you
   attention to you for , and you kind another article you mention
   Finaly , lit to bud I not as the truth .</p>
<p>Yours faithfully ,</p>
```

Modification 1(c): randomly order word trigrams within a sentence:

```
<p>Dear Sir/Madam ,</p>
<p>your advertisement which two days ago with reference to a local
   newspaper I am writing I read in .</p>
<p>activities weekend , , I would which was organised my
   dissatisfaction to part in this like to express by the college As
   I took your article .</p>
<p>be cancelled , course had to mentions that the cancelled only the
   First of all but the course was ill , because the teacher , your
   advertisement last afternoon .</p>
<p>body registered for mention that no Secondly , you the local
   history course .</p>
<p>So I would like to inform you that 32 attended to this course .</p
   >
<p>, they were courses is true painting and photography far as the
   what you say Thirdly , as very popular .</p>
<p>being these courses so popular is the reason for these kind of I
   think that that people are very interested in activities .</p>
```

```
<p>very experienced and the teachers was the equipment was Apart from
    this particullarly coteporary .</p>
<p>the weekend cost Furthemore , your article mentions that over 100
    .</p>
<p>That is not true .</p>
<p>The cost was only 60 including lunch .</p>
<p>it will be least , the college is planning activities weekend next
    to run another term , so Last but not sooner than you mention .</
   p>
<p>bud I would write another article my letter , as you mention kind
   attention to would like to tell you that and I would the weekend
   was not as disappointing which will lit like you to like to thank
   you for you Finaly , I to the truth .</p>
<p>Yours faithfully ,</p>
```

Modification 1(d): randomly order sentences within a script:

```
<p>You must be very happy because you have been offed two jobs .</p>
<p>Also working in a museum include big tips from the tourists .</p>
<p>Last but not last , I hope you will find my opinion helpful I wish
    your problem will be solved , but I do n't think that it 's about
    a problem .</p>
<p>Yours faithfully ,</p>
<p>First of all , you mus n't panic .</p>
<p>Thirdly , as far as the painting and photography courses is true
   what you say , they were very popular .</p>
<p>I think that the reason for being these courses so popular is that
    people are very interested in these kind of activities .</p>
<p>So I would like to inform you that 32 attended to this course .</p
   >
<p>But you should wonder which working place is more convenient to
   you .</p>
<p>I hope you are fine .</p>
<p>Apart from these , despite the fact that it is a summer job you
   could be employed for the whole next year , if you are lucky .</p>
<p>Drop me a line .</p>
<p>You 'd better not to travel far away from your home .</p>
<p>Whatever choice you will make I 'm sure it will be the best for
   you .</p>
<p>That is not true .</p>
<p>Furthemore , your article mentions that the weekend cost over 100
    .</p>
<p>Finaly , I would like to tell you that the weekend was not as
   disappointing as you mention bud I would like to thank you for you
    kind attention to my letter , and I would like you to write
   another article which will lit to the truth</p>
```

Modification 2: Swap words that have the same POS within a sentence:

```
<p>local Sir/Madam , I am writing with advertisement in your
   newspaper which I read to a Dear reference two days ago .</p>
```

```
<p>As I took college by this activities weekend , which was organised
    to the dissatisfaction , I would express to like your article in
   my part .</p>
<p>last of only , your course mentions that the teacher had to be
   cancelled , because the course was ill , but the advertisement
   cancelled all the First afternoon .</p>
<p>Secondly , you mention that the history registered for no local
   course body .</p>
<p>So I would inform to like you that 32 attended to this course .</p
   >
<p>Thirdly , as true as the photography and painting courses is
   popular what you say , they were very far .</p>
<p>I think that the being for kind these activities so popular is
   that people are very interested in these reason of courses .</p>
<p>from Apart this the teachers was very experienced and the
   equipment was particullarly coteporary .</p>
<p>over 100 , your article mentions that the weekend cost Furthemore
    .</p>
<p>That is not true .</p>
<p>The cost was only 60 including lunch .</p>
<p>next but not least , the term is planning to mention another
   activities weekend Last college , so it will be sooner than you
   run .</p>
```

APPENDIX F

Discourse connectives

Addition
additionally, again, also, and, besides, equally, finally, further, furthermore, in addition, indeed, more, moreover, next, too.

Comparison
again, also, compared to, compared with, in comparison to, in comparison with, in the same manner, in the same way, likewise, similar, similarly.

Contrast
alternatively, although, (it may) be the case that, besides, but, conversely, despite, different from, granted, however, in contrast, it is true that, (it) may (be the case that), nevertheless, notwithstanding, on the contrary, on the other hand, regardless, whereas, while, yet.

Conclusion
basically, finally, in all, in brief, in conclusion, in a nutshell, in short, in summary, on the whole, therefore, to conclude, to sum up, to summarise.

APPENDIX G

USE QUESTIONNAIRE

Usefulness

a. It helps me be more effective.
b. It helps me be more productive.
c. It is useful.
d. It gives me more control over the activities in my life.
e. It makes the things I want to accomplish easier to get done.
f. It saves me time when I use it.
g. It meets my needs.
h. It does everything I would expect it to do.

Ease of Use

a. It is easy to use.
b. It is simple to use.
c. It is user friendly.
d. It requires the fewest steps possible to accomplish what I want to do with it.
e. It is flexible.
f. Using it is effortless.
g. I can use it without written instructions.
h. I don't notice any inconsistencies as I use it.
i. Both occasional and regular users would like it.
j. I can recover from mistakes quickly and easily.
k. I can use it successfully every time.

Ease of Learning

a. I learned to use it quickly.
b. I easily remember how to use it.

c. It is easy to learn to use it.

d. I quickly became skilful with it.

Satisfaction

a. I am satisfied with it.

b. I would recommend it to a friend.

c. It is fun to use.

d. It works the way I want it to work.

e. It is wonderful.

f. I feel I need to have it.

g. It is pleasant to use.

(Lund, 2001)

Bibliography

Alexopoulou, T., Yannakoudakis, H., and Salamoura, A. (2013). Classifying intermediate Learner English: a data-driven approach to learner corpora. In Granger, S., Gilquin, G., and Meunier, F., editors, *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, Corpora and Language in Use – Proceedings 1, pages 11–23. Louvain-la-Neuve: Presses universitaires de Louvain.

AMIDA (2007). Augmented multi-party interaction with distance access. Technical report, `www.amidaproject.org`.

Andersen, Ø. E. (2011). *Grammatical error prediction*. PhD thesis, University of Cambridge.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-Rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3):1–30.

Attali, Y., Powers, D. E., Freedman, M., Harrison, M., and Obetz, S. A. (2008). Automated scoring of short-answer open-ended GRE subject test items. Technical Report GREB-04-02, RR-08-20, ETS, `http://www.ets.org/Media/Research/pdf/RR-08-20.pdf`.

Axelsson, M. W. (2000). USE – the uppsala student english corpus: an instrument for needs analysis. *ICAME Journal*, 24:155–157.

Banko, M. and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33. ACL.

Barker, F., Post, B., Schmidt, E., and McCarthy, M. (2011). Identifying criterial aspects of pronunciation in L2 English across CEFR levels: Implications for language learning. In Angouri, J., Daller, M., and Treffers-Daller, J., editors, *Proceedings of the 44th Annual Meeting of the British Association for Applied Linguistics*, The impact of Applied Linguistics, pages 17 – 21. Scitsiugnil Press, UK.

Baroni, M., Lenci, A., and Onnis, L. (2007). ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the workshop on Cognitive Aspects of Computational Language Acquisition*, pages 49–56. ACL.

Barzilay, R. and Lapata, M. (2008). Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1):1–34.

Barzilay, R. and Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 113–120. ACL.

Battista, G. D., Eades, P., Tamassia, R., and Tollis, I. G. (1994). Algorithms for drawing graphs: an annotated bibliography. *Computational Geometry: Theory and Applications*, 4(5):235–282.

Battista, G. D., Eades, P., Tamassia, R., and Tollis, I. G. (1998). *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR Upper Saddle River, NJ, USA.

Ben-Bassat, T., Meyer, J., and Tractinsky, N. (2006). Economic and subjective measures of the perceived value of aesthetics and usability. *ACM Transactions on Computer-Human Interaction*, 13(2):210–234.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc, Secaucus, NJ, USA.

Blandford, A., Cox, A., and Cairns, P. (2008). Controlled experiments. In Cairns, P. A. and Cox, A. L., editors, *Research Methods for Human-Computer Interaction*, pages 1–16. Cambridge University Press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993 – 1022.

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426.

Bös, S. and Opper, M. (1998). Dynamics of batch training in a perceptron. *Journal of Physics A: Mathematical and General*, 31(21):4835–4850.

Boyd, A., Zepf, M., and Meurers, D. (2012). Informing Determiner and Preposition Error Correction with Word Clusters. In *Proceedings of the 7th workshop on Innovative Use of NLP for Building Educational Applications*, pages 208 – 215. Association for Computational Linguistics.

Briscoe, T. (2006). An introduction to tag sequence grammars and the RASP system parser. Technical Report UCAM-CL-TR-662, University of Cambridge, Computer Laboratory, `http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-662.pdf`.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In *Proceedings of the ACL-Coling'06 Interactive Presentation Session*, pages 77–80. ACL.

Briscoe, T., Medlock, B., and Andersen, Ø. E. (2010). Automated assessment of ESOL free text examinations. Technical Report UCAM-CL-TR-790, University of Cambridge, Computer Laboratory, `http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-790.pdf`.

Brooke, J. and Hirst, G. (2012). Measuring Interlanguage: Native Language Identification with L1-influence Metrics. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 1 – 6, Istanbul, Turkey. European Language Resources Association (ELRA).

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Burstein, J. (2003). The e-Rater scoring engine: Automated essay scoring with natural language processing. In Shermis, M. D. and Burstein, J., editors, *Automated essay scoring: A cross-disciplinary perspective*, pages 113–121. Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., and Leacock, C. (2003). Criterion: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence*, pages 3–10. American Association for Artificial Intelligence.

Burstein, J., Kukich, K., Wolff, S., Lu, C., and Chodorow, M. (1998a). Computer analysis of essays. In *Proceedings of the annual meeting of the National Council of Measurement in Education*, pages 1–13.

Burstein, J., Kukich, K., Wolff, S., Lu, C., and Chodorow, M. (1998b). Enriching automated essay scoring using discourse marking. In *Workshop on Discourse Relations and Discourse Marking*, pages 15 – 21. ACL.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris, M. D. (1998c). Automated scoring using a hybrid feature identification technique. *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pages 206–210.

Burstein, J., Tetreault, J., and Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684. ACL.

Callear, D., Jerrams-Smith, J., and Soh, V. (2001). CAA of short non-MCQ answers. In *Proceedings of the 5th Computer-Assisted Assessment (CAA) Conference*, pages 1 – 14. Loughborough: Loughborough University.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann, USA.

Chae, J. and Nenkova, A. (2009). Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–147. ACL.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, 21(4):505–524.

Charniak, E. and Elsner, M. (2009). EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 148–156. ACL.

Chen, J.-W. and Zhang, J. (2007). Comparing Text-based and Graphic User Interfaces for Novice and Expert Users. In *AMIA Annual Symposium Proceedings Archive*, pages 125–129. American Medical Informatics Association.

Chen, L. and Yoon, S. Y. (2011). Detecting structural events for assessing non-native speech. In *Proceedings of the 6th workshop on Innovative Use of NLP for Building Educational Applications*, pages 38–45. Association for Computational Linguistics.

Chen, Y. Y., Liu, C. L., Chang, T. H., and Lee, C. H. (2010). An Unsupervised Automated Essay Scoring System. *IEEE Intelligent Systems*, 25(5):61–67.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, USA.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Collins, C. M. (2010). *Interactive Visualizations of natural language*. PhD thesis, University of Toronto.

Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL*, 21(2):259–279.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge.

Culy, C., Lyding, V., and Dittmann, H. (2011). Structured Parallel Coordinates: a visualization for analyzing structured language data. In *Proceedings of the 3rd International Conference on Corpus Linguistics, CILC-11*, pages 485–493.

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

Dale, R., Anisimoff, I., and Narroway, G. (2012). HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the 7th workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–62. Association for Computational Linguistics.

De Felice, R. and Pulman, S. G. (2008a). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling) – Volume 1*, pages 169–176. Association for Computational Linguistics.

De Felice, R. and Pulman, S. G. (2008b). Automatic detection of preposition errors in learner writing. *CALICO Workshop on Automatic Analysis of Learner Language*, 26(3):512–528.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Díaz-Negrillo, A. and Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista española de lingüística aplicada*, 19:83–102.

Dickinson, M., Kübler, S., and Meyer, A. (2012). Predicting Learner Levels for Online Exercises of Hebrew. In *Proceedings of the 7th workshop on Innovative Use of NLP for Building Educational Applications*, pages 95–104. Association for Computational Linguistics.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1):1 – 36.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., and Plaisant, C. (2007). Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 213–222. ACM.

Dumas, J. F. and Redish, J. C. (1999). *A Practical Guide to Usability Testing*. Intellect Ltd, UK.

D'Ydewalle, G., Leemans, J., and Rensbergen, J. V. (1995). Graphical versus character-based word processors: an analysis of user performance. *Behaviour & Information Technology*, 14(4):208–214.

Eades, P. and Sugiyama, K. (1990). How to draw a directed graph. *Journal of Information Processing*, 13(4):424–437.

Ellis, R. (1997). *The Study of Second Language Acquisition*. Oxford Introduction to Language Study. Oxford University Press, Oxford.

Elsner, M. (2011). *Generalizing Local Coherence Modeling*. PhD thesis, Brown University.

Elsner, M. and Charniak, E. (2008). Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 41–44. ACL.

Elsner, M. and Charniak, E. (2010). The same-head heuristic for coreference. In *Proceedings of the Association for Computational Linguistics*, pages 33 – 37. ACL.

Elsner, M. and Charniak, E. (2011a). Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1179–1189. ACL.

Elsner, M. and Charniak, E. (2011b). Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129. ACL.

Ericsson, P. F. and Haswell, R. H. (2006). *Machine scoring of student essays: truth and consequences*. Utah State University Press, Logan, Utah.

Escalante, H. J., Solorio, T., and Montes-y-Gómez, M. (2011). Local Histograms of Character N-grams for Authorship Attribution. In *Proceedings of the 49th Annual Meeting on Association for Computational Linguistics*, pages 288–298. ACL.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *The Behavioral and brain sciences*, 32(5):429 – 448.

Faller, A. J. (1981). An Average Correlation Coefficient. *Journal of Applied Meteorology*, 20:203–205.

Feldman, R. and Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, Cambridge.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In Evert, S., Kilgarriff, A., and Sharoff, S., editors, *Proceedings of the 4th Web as Corpus Workshop*, pages 47–54.

Ffrench, A., Bridges, G., and Beresford-Knox, J. (2012). Quality Assurance: A Cambridge ESOL system for managing Writing examiners. In *University of Cambridge ESOL Examinations Internal Research Notes 49*, pages 11–17. University of Cambridge ESOL Examinations.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Fry, B. (2007). *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. O'Reilly Media, Canada.

Galaczi, E., Post, B., Li, A., and Graham, C. (2011). Measuring L2 English Phonological Proficiency: Implications for Language Assessment. In Angouri, J., Daller, M., and Treffers-Daller, J., editors, *Proceedings of the 44th Annual Meeting of the British Association for Applied Linguistics*, The impact of Applied Linguistics, pages 67 – 72. Scitsiugnil Press, UK.

Gansner, E. R., Koutsofios, E., North, S. C., and Vo, K. P. (1993). A technique for drawing directed graphs. *IEEE Transactions on Software Engineering*, 19(3):214–230.

Gao, J., Misue, K., and Tanaka, J. (2009). A Multiple-Aspects Visualization Tool for Exploring Social Networks. In *Human Interface and the Management of Information. Information and Interaction*, pages 277–286. Springer Berlin Heidelberg.

Garcia, E. (2010). A Tutorial on Correlation Coefficients. Technical report, `http://web.simmons.edu/~benoit/lis642/a-tutorial-on-correlation-coefficients.pdf`.

Gillard, P. and Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In Granger, S., editor, *Learner English on computer*, pages 159–171. Addison Wesley Longman.

Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420.

Gospodnetic, O. and Hatcher, E. (2004). *Lucene in Action*. Manning Publications, Greenwich.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior research methods, instruments, & computers: a journal of the Psychonomic Society, Inc.*, 36(2):193–202.

Gram, L. and Buttery, P. (2009). A tutorial introduction to iLexIR Search. Internal report.

Granger, S. (1994). The Learner Corpus: a revolution in applied linguistics. *English Today*, 10(03):25–33.

Granger, S. (2002). A bird's-eye view of learner corpus research. In Granger, S., Hung, J., and Petch-Tyson, S., editors, *Computer learner corpora, second language acquisition and foreign language teaching*, pages 3 – 33. John Benjamins Publishing.

Granger, S. (2003a). Error-tagged learner corpora and CALL: a promising synergy. *CALICO journal*, 20(3):465–480.

Granger, S. (2003b). The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Teachers of English to Speakers of Other Languages (TESOL) Quarterly*, 37(3):538–546.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Aijmer, K., editor, *Corpora and Language Teaching*, pages 13–32. John Benjamins, Amsterdam, The Netherlands.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman Pub Group, UK.

Hawkins, J. and Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In Taylor, L. and Weir, C., editors, *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment – Proceedings of the ALTE Cambridge Conference, April 2008*, pages 158 – 175. Cambridge University Press, Cambridge.

Hawkins, J. and Buttery, P. (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, 1(1):1–23.

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

Hearst, M. A. (2009). *Search user interfaces*. Cambridge University Press, Cambridge.

Heer, J. and Boyd, D. (2005). Vizster: visualizing online social networks. In *IEEE Symposium on Information Visualization (INFOVIS)*, pages 32–39. IEEE.

Heer, J., Card, S. K., and Landay, J. A. (2005). Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430, New York, USA. ACM.

Herman, I., Melançon, G., and Marshall, M. S. (2000). Graph visualization and navigation in information visualization: a survey. *IEEE transactions on visualization and computer graphics*, 6(1):24–43.

Higgins, D. and Burstein, J. (2007). Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 1–12.

Higgins, D., Burstein, J., and Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2):145 – 159.

Higgins, D., Burstein, J., Marcu, D., and Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 185–192. ACL.

Horsky, J., McColgan, K., Pang, J. E., Melnikas, A. J., Linder, J. A., Schnipper, J. L., and Middleton, B. (2010). Complementary methods of system usability evaluation: surveys and observations during software design and development cycles. *Biomedical informatics*, 43(5):782–790.

Iliinsky, N. and Steele, J. (2011). *Designing Data Visualizations*. O'Reilly Media, Canada.

Ionin, T. and Montrul, S. (2010). The role of L1 transfer in the interpretation of articles with definite plurals in L2 English. *Language Learning*, 60(4):877–925.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142. Springer-Verlag London, UK.

Joachims, T. (1999). Making large scale SVM learning practical. In Schölkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169 – 184. MIT Press.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Johnson, J. S. and Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1):60–99.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. New jersey: Pearson Education Inc, 2nd edition.

Jurgens, D. and Stevens, K. (2010). The S-Space package: an open source package for word space models. In *Proceedings of the Association for Computational Linguistics 2010 System Demonstrations*, pages 30–35. ACL.

Käki, M. and Aula, A. (2008). Controlling the complexity in comparing search user interfaces via user studies. *Information Processing & Management*, 44(1):82–91.

Kakkonen, T., Myller, N., and Sutinen, E. (2004). Semi-automatic evaluation features in computer-assisted essay assessment. In *Proceedings of the 7th IASTED International Conference on Computers and Advanced Technology in Education*, pages 456 – 461.

Kakkonen, T. and Sutinen, E. (2008). Evaluation criteria for automatic essay assessment systems – there is much more to it than just the correlation. In *Proceedings of the 16th International Conference on Computers in Education*, pages 111–116.

Keppel, G., Saufley, W., and Tokunaga, H. (1992). *Introduction to Design and Analysis: A Student's Handbook*. Books in psychology. Worth Publishers, USA.

Kochmar, E. (2011). *Identification of a writer's native language by error analysis*. MPhil thesis, University of Cambridge.

Kochmar, E., Andersen, Ø. E., and Briscoe, T. (2012). HOO 2012 Error Recognition and Correction Shared Task: Cambridge University Submission Report. In *Proceedings of the 7th workshop on the Innovative Use of NLP for Building Educational Applications, HOO Shared Task*, pages 242–250. Association for Computational Linguistics.

Kohavi, R., Henne, R. M., and Sommerfield, D. (2007). Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, pages 959 – 967, New York, USA. ACM Press.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181.

Koppel, M., Schler, J., and Zigdon, K. (2005). Automatically determining an anonymous author's native language. In *Proceedings of the 2005 IEEE international conference on Intelligence and Security Informatics*, pages 209–217, Berlin, Heidelberg. Springer-Verlag.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Language teaching methodology series. Pergamon, Oxford.

Kukich, K. (2000). Beyond Automated Essay Scoring. *IEEE Intelligent systems*, 15(5):22–27.

Lafferty, J. and Lebanon, G. (2005). Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163.

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2011). Seven Guiding Scenarios for Information Visualization Evaluation. Technical Report 2011-992-04, University of Calgary, `http://hal.inria.fr/docs/00/72/30/57/PDF/Lam_2011_SGS.pdf`.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Landauer, T. K., Laham, D., and Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Shermis, M. and Burstein, J. C., editors, *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112.

Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In Shafto, M. G. and Langley, P., editors, *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417. Mawhwah, NJ: Erlbaum.

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '98*, pages 90–95. ACM.

Leacock, C. and Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Lebanon, G., Mao, Y., and Dillon, J. (2007). The locally weighted bag-of-words framework for document representation. *Journal of Machine Learning Research*, 8(10):2405–2441.

Leech, G. (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing*, 8(4):275–281.

Lenneberg, E. H. (1967). *Biological foundations of language*. Wiley, Oxford.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):5–55.

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information*. Princeton Architectural Press, USA.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically Evaluating Text Coherence Using Discourse Relations. In *Proceedings of the 49th Annual Meeting on Association for Computational Linguistics*, pages 997–1006. ACL.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Long, P. M. and Servedio, R. A. (2007). Discriminative learning can succeed where generative learning fails. In *Learning Theory*, pages 319–334. Springer Berlin Heidelberg.

Lonsdale, D. and Strong-Krause, D. (2003). Automated rating of ESL essays. In *Proceedings of the HLT-NAACL 2003 workshop on Building Educational Applications Using Natural Language Processing*, pages 61 – 67. ACL.

Lund, A. M. (2001). Measuring usability with the USE questionnaire. *Usability Interface*, 8(2):3–6.

Lyding, V., Lapshinova-Koltunski, E., Degaetano-Ortlieb, S., and Dittmann, H. (2012). Visualising Linguistic Evolution in Academic Discourse. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 44–48. Association for Computational Linguistics.

Mack, R. and Nielsen, J. (1995). Usability inspection methods: Executive summary. In *Human-computer interaction*, pages 170–181. Morgan Kaufmann Publishers Inc.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mao, Y., Dillon, J. V., and Lebanon, G. (2007). Sequential document visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1208–1215.

McClosky, D., Charniak, E., and Johnson, M. (2008). *BLLIP North American News Text, Complete*. Linguistic Data Consortium, Philadelphia.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press, Edinburgh.

Meyer, M., Munzner, T., DePace, A., and Pfister, H. (2010a). MulteeSum: a tool for comparative spatial and temporal gene expression data. *IEEE transactions on Visualization and Computer Graphics*, 16(6):908–917.

Meyer, M., Wong, B., Styczynski, M., Munzner, T., and Pfister, H. (2010b). Pathline: A tool for comparative functional genomics. *Computer Graphics Forum*, 29(3):1043–1052.

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, 29(4):495 – 512.

Milton, J. C. P. and Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. In Flowerdew, L. and Tong, A. K., editors, *Entering text*, pages 127–143. Hong Kong: The Hong Kong University of Science and Technology.

Miltsakaki, E. and Kukich, K. (2000). Automated evaluation of coherence in student essays. In *Proceedings of LREC 2000*, pages 1–8.

Miltsakaki, E. and Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(01):25–55.

Mitchell, T., Russell, T., Broomhead, P., and Aldridge, N. (2002). Towards robust computerised marking of free-text responses. In *Proceedings of the 6th CAA Conference*, pages 233 – 249. Loughborough University.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Mohler, M., Bunescu, R., and Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752 – 762. ACL.

Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921 – 928.

Nesi, H. (2008). Corpora and English for Academic Purposes. In *Proceedings of the 6th Languages for Specific Purposes International Seminar*, LSP: Interfacing Language with other Realms, Universiti Teknologi Malaysia, Johor Bahru: Malaysia.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In *How to use corpora in language teaching*, pages 125 – 149. John Benjamins Publishing.

Ng, A. Y. (2012). Coursera Machine Learning courses.

Ng, A. Y. and Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in neural information processing systems 14*, pages 841–848. MIT Press, USA.

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

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 373–380, New York, USA. ACM.

Nielsen, J. (1993). *Usability engineering*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Nitin, M., Tetreault, J., and Chodorow, M. (2012). Exploring Grammatical Error Correction with Not-So-Crummy Machine Translation. In *Proceedings of the 7th workshop on the Innovative Use of NLP for Building Educational Applications*, pages 44–53. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. ACL.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

O'Malley, J. M. and Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. The Cambridge Applied Linguistics Series. Cambridge University Press, Cambridge.

Osborne, J. (2011). Oral learner corpora and the assessment of fluency in the Common European Framework. In Frankenberg-Garcia, A., Flowerdew, L., and Aston, G., editors, *New Trends in Corpora and Language Learning*, Research in Corpus and Discourse, chapter 11, pages 181 – 198. Continuum International Publishing Group.

Page, E. (2003). Project essay grade: PEG. In Shermis, M. D. and Burstein, J. C., editors, *Automated essay scoring: A cross-disciplinary perspective*, pages 43–54.

Page, E. B. (1967). Grading essays by computer: progress report. In *Proceedings of the Invitational Conference on Testing Problems*, pages 87–100.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, 14(2):210–225.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559 – 572.

Peng, X., Ke, D., and Xu, B. (2012). Automated Essay Scoring Based on Finite State Transducer: towards ASR Transcription of Oral English Speech. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 50–59. ACL.

Perer, A. and Shneiderman, B. (2006). Balancing Systematic and Flexible Exploration of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700.

Pérez-Marín, D., Pascual-Nieto, I., and Rodríguez, P. (2009). Computer-assisted assessment of free-text answers. *The Knowledge Engineering Review*, 24(4):353–374.

Persing, I., Davis, A., and Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. Association for Computational Linguistics.

Phillips, S. E., Society for the Advancement of Excellence in Education Staff, and TASA Institute Staff (2007). *Automated Essay Scoring: A Literature Review*. Volume 30 of SAEE Research Series. Society for the Advancement of Excellence in Education.

Pitler, E., Louis, A., and Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554. ACL.

Plaisant, C. (2004). The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced Visual Interfaces - AVI '04*, pages 109 – 116, New York, USA. ACM Press.

Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M. G., Smith, M. N., Clement, T., and Lord, G. (2006). Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 141–150. ACM.

Poesio, M., Kabadjov, M. A., Vieira, R., Goulart, R., and Uryupina, O. (2005). Does discourse-new detection help definite description resolution? In *Proceedings of the 6th International Workshop on Computational Semantics*, pages 1 –12.

Pomerleau, D. A. (1989). ALVINN: an autonomous land vehicle in a neural network. Technical Report AIP-77, Pittsburgh, PA: Carnegie Mellon University, `http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA218975`.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., and Kukich, K. (2002). Stumping e-Rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2):103–134.

Pravec, N. (2002). Survey of learner corpora. *ICAME journal*, 26:81–114.

Preston, D. and Goodman, D. (2012). Automated Essay Scoring and The Repair of Electronics. Technical report, `http://snap.stanford.edu/class/cs341-2012/reports/03-Preston_cs341_-_Dan_and_Danny_-_Final.pdf`.

Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications using Natural Language Processing*, pages 9–16. Association for Computational Linguistics.

Robertson, D. (2000). Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16(2):135–172.

Rohrdantz, C., Hautli, A., Mayer, T., and Butt, M. (2011). Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Meeting of the Association for Computational Linguistics*, pages 305–310. ACL.

Rosé, C., Roque, A., Bhembe, D., and Vanlehn, K. (2003). A hybrid text classification approach for analysis of student essays. In *Proceedings of the HLT-NAACL workshop on Building Educational Applications using Natural Language Processing*, pages 68–75. Association for Computational Linguistics.

Rowland, C. F., Pine, J. M., Lieven, E. V. M., and Theakston, A. L. (2005). The incidence of error in young children's Wh-questions. *Journal of speech, language, and hearing research : JSLHR*, 48(2):384–404.

Rozovskaya, A. and Roth, D. (2011). Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 924–933. Association for Computational Linguistics.

Rozovskaya, A., Sammons, M., and Roth, D. (2012). The UI System in the HOO 2012 Shared Task on Error Correction. In *Proceedings of the 7th workshop on the Innovative Use of NLP for Building Educational Applications, HOO Shared Task*, pages 272 – 280. Association for Computational Linguistics.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Rubin, A. (2009). *Statistics for evidence-based practice and evaluation*. Cengage Learning, USA.

Rudner, L., Garcia, V., and Welch, C. (2006). An evaluation of IntelliMetric essay scoring system. *The Journal of Technology, Learning, and Assessment*, 4(4):1 – 22.

Rudner, L. and Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2):3–21.

Sahlgren, M. (2005). An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, pages 1–9. Citeseer.

Saville, N. and Hawkey, R. (2010). The English Profile Programme – the first three years. *English Profile Journal*, 1(1):1 – 14.

Schiftner, B. (2008). Learner Corpora of English and German: What is their status quo and where are they headed. *Vienna English Working Papers*, 17(2):47–78.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press, Cambridge.

Shermis, M. and Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: analysis. In *Annual National Council on Measurement in Education Meeting*, pages 1–54.

Shermis, M. D. and Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates, NJ, USA.

Shneiderman, B. and Plaisant, C. (2006). Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1 – 7. ACM.

Sleator, D. D. K. and Templerley, D. (1995). Parsing English with a link grammar. In *Proceedings of the 3rd International Workshop on Parsing Technologies*, pages 1 – 14. Association for Computational Linguistics.

Smola, A. J. (1996). *Regression estimation with support vector learning machines*. Master's thesis, Technische Universität München.

Soricut, R. and Marcu, D. (2006). Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 803–810. ACL.

Staggers, N. and Kobus, D. (2000). Comparing response time, errors, and satisfaction between text-based and graphical user interfaces during nursing order tasks. *Journal of the American Medical Informatics Association*, 7(2):164–176.

Steele, J. and Iliinsky, N. (2010). *Beautiful Visualization*. O'Reilly Media, Canada.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

Sukkarieh, J. Z., Pulman, S. G., and Raikes, N. (2003). Auto-marking: using computational linguistics to score short, free text responses. In *Proceedings of the 29th Annual Conference of the International Association for Educational Assessment (IAEA)*, pages 1–15.

Swanson, B. and Charniak, E. (2012). Native Language Detection with Tree Substitution Grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 193–197. ACL.

Swanson, B. and Yamangil, E. (2012). Correction Detection and Error Type Selection as an ESL Educational Aid. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 357–361. Association for Computational Linguistics.

Tandalla, L. (2012). Scoring Short Answer Essays. Technical report, `https://kaggle2.blob.core.windows.net/competitions/kaggle/2959/media/TechnicalMethodsPaper.pdf`.

Tono, Y. (2003). Learner corpora: design, development and applications. In Archer, D., Rayson, P., Wilson, A., and McEnery, T., editors, *Proceedings of the 2003 Corpus Linguistics Conference*, pages 800–809.

Tory, M. and Möller, T. (2005). Evaluating visualizations: do expert reviews work? *Computer Graphics and Applications, IEEE*, 25(5):8–11.

Trenkic, D. (2008). The representation of English articles in second language grammars: Determiners or adjectives? *Bilingualism: Language and Cognition*, 11(1):1–18.

Valenti, S., Neri, F., and Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2:319–330.

Van Ham, F., Wattenberg, M., and Viégas, F. B. (2009). Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1169–1176.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, USA.

Viégas, F. B., Wattenberg, M., and Feinberg, J. (2009). Participatory visualization with Wordle. *IEEE transactions on Visualization and Computer Graphics*, 15(6):1137–1144.

West, R., Park, Y., and Levy, R. (2011). Bilingual random walk models for automated grammar correction of ESL author-produced text. In *Proceedings of the 6th workshop on Innovative Use of NLP for Building Educational Applications*, pages 170–179. Association for Computational Linguistics.

Whiteside, J., Jones, S., Levy, P. S., and Wixon, D. (1985). User performance with command, menu, and iconic interfaces. *ACM SIGCHI Bulletin*, 16(4):185–191.

Wiemer-Hastings, P. and Graesser, A. C. (2000). Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 8(2):149–169.

Williams, C. (2008). The Cambridge Learner Corpus for researchers on the English Profile Project – Version 2. Technical report, University of Cambridge ESOL Examinations. Internal report.

Williams, E. J. (1959). The Comparison of Regression Variables. *Journal of the Royal Statistical Society*, 21(2):396–399.

Williamson, D., Xi, X., and Breyer, F. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Williamson, D. M. (2009). A Framework for Implementing Automated Scoring. In *Proceedings of the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education*, pages 1 – 39, San Diego, CA.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, 25(2-3):309–336.

Wong, S. M. J. and Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610. ACL.

Yannakoudakis, H. and Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In *Proceedings of the 7th workshop on the Innovative Use of NLP for Building Educational Applications, NAACL*, pages 33–43. ACL.

Yannakoudakis, H., Briscoe, T., and Alexopoulou, T. (2012). Automating Second Language Acquisition Research: Integrating Information Visualisation and Machine Learning. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 35–43. ACL.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180 – 189. ACL.

Yoon, S. Y. and Higgins, D. (2011). Non-English Response Detection Method for Automated Proficiency Scoring System. In *Proceedings of the 6th workshop on Innovative Use of NLP for Building Educational Applications*, pages 161–169, Portland, Oregon. Association for Computational Linguistics.

Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., and Carpendale, S. (2006). Heuristics for information visualization evaluation. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–6. ACM.