

Automatic Identification of Discourse Moves in Scientific Article Introductions

Nick Pendar and Elena Cotos

Applied Linguistics and Technology Program

Iowa State University

Ames, IA 50011 USA

{pendar, ecotos}@iastate.edu

Abstract

This paper reports on the first stage of building an educational tool for international graduate students to improve their academic writing skills. Taking a text-categorization approach, we experimented with several models to automatically classify sentences in research article introductions into one of three rhetorical moves. The paper begins by situating the project within the larger framework of intelligent computer-assisted language learning. It then presents the details of the study with very encouraging results. The paper then concludes by commenting on how the system may be improved and how the project is intended to be pursued and evaluated.

1 Introduction and Background

Interest in automated evaluation systems in the field of language assessment has been growing rapidly in the last few years. Performance-based and high-stakes standardized tests (e.g., ACT, GMAT, TOEFL, etc.) have employed such systems due to their potential to yield evidence about the learners' language proficiency and/or subject matter mastery based on analyses of their constructed responses. Automated writing evaluation applications are also beginning to draw the attention of pedagogues who are much interested in assessment for learning, i.e., assessment used as a tool in gaining direction for remediation. Arguably, these technological innovations open up a wide range of possibilities for high-quality formative evaluation that can closely match teaching goals and tailor instruction to individual

learners by providing them with feedback and direction on their attainment of knowledge.

Traditionally, automated evaluation has been used for essay grading, but its potential could be successfully extrapolated to other genres in both first language (L1) and second language (L2) academic contexts. Existing scoring systems can assess various constructs such as topical content, grammar, style, mechanics, syntactic complexity, and even deviance or plagiarism (Burstein, 2003; Elliott, 2003; Landauer et al., 2003; Mitchell et al., 2002; Page, 2003; Rudner and Liang, 2002). Because learner writing is generally highly erroneous, an emerging research trend has focused on automated error detection in L2 output finding novel approaches to develop intelligent ways to assess ill-formed learner responses (Burstein and Chodorow, 1999; Chodorow et al., 2007; Han et al., 2006; Leacock and Chodorow, 2003). Various NLP and statistical techniques also allow for the evaluation of text organization, which is however limited to recognizing the five-paragraph essay format, thesis, and topic sentences. At present, to our knowledge, there is only one automated evaluation system, AntMover (Anthony and Lashkia, 2003), that applies intelligent technological possibilities to the genre of research reports—a major challenge for new non-native speaker (NNS) members of academia. AntMover is able to automatically identify the structure of abstracts in various fields and disciplines.

Academic writing pedagogues have been struggling to find effective ways to teach academic writing. Frodesen (1995) argues that the writing instruction for non-native speaker students should “help

initiate writers into their field-specific research communities” (p. 333). In support of this opinion, (Kushner, 1997) reasons that graduate NNS courses have to combine language and discourse with the skill of writing within professional norms. Various pedagogical approaches have been attempted to achieve this goal. For instance, (Vann and Myers, 2001) followed the inductive analysis approach, in which students examined the format, content, grammatical, and rhetorical conventions of each section of research reports. Supplements to this approach were tasks that required students to write journal entries about the rhetorical conventions of prominent journals in their disciplines and tasks that placed the experience of writing up research “in the framework of an interactive, cooperative effort with cross-cultural interaction” (Vann and Myers, 2001, p. 82). Later, after having followed a primarily skill-based approach, in which students wrote field-specific literature reviews, summaries, paraphrases, data commentaries, and other discipline-specific texts, Levis and Levis-Muller (2003) reported on transforming the course into a project-based writing one. The project consisted of carrying out original research, the topic of which, for the purpose of coping with discipline diversity, was the same for all students and was determined by the instructor. From the start, the students were provided with a limited set of articles, for instance, on cross-cultural adjustment, with which they worked to identify potential research questions for a further investigation and to write the literature review. This approach placed a heavy emphasis on collaboration as students worked in small groups on developing data-collection instruments and on data analysis. Oral presentations on group-research projects wrapped up the course.

The academic writing course discussed in the paragraph above is corpus- and genre-based, combining a top-down approach to genre analysis and a bottom-up approach to the analysis of corpora (Cortes, 2006). Cortes (2006) explains that the course was designed to better address the issues of genre-specificity and disciplinarity since some students who took the previous form of the course claimed that, although they were taught useful things, they did not learn to write the way researchers in their disciplines generally do. In the present format of the course, each student is pro-

vided with a corpus of research articles published in top journals of his/her discipline. Students conduct class analyses of their corpus according to guidelines from empirical findings in applied linguistics about the discourse tendencies in research article writing. Their task is to discover organizational and linguistic patterns characteristic of their particular discipline, report on their observations, and apply the knowledge they gain from the corpus analyses when writing a research article for the final project in the course.

2 Motivation

Although each of the pedagogical approaches mentioned in the previous section has its advantages, they all fail to provide NNS students with sufficient practice and remedial guidance through extensive individualized feedback during the process of writing. An NLP-based academic discourse evaluation software application could account for this drawback if implemented as an additional instructional tool. However, an application with such capabilities has not yet been developed. Moreover, as mentioned above, the effects of automated formative feedback are not fully investigated. The long-term goal of this research project is the design and implementation of a new automated discourse evaluation tool as well as the analysis of its effectiveness for formative assessment purposes. Named IADE (Intelligent Academic Discourse Evaluator), this application will draw from second language acquisition models such as interactionist views and Systemic Functional Linguistics as well as from the Skill Acquisition Theory of learning. Additionally, it will be informed by empirical research on the provision of feedback and by Evidence Centered Design principles (Mislevy et al., 2006).

IADE will evaluate students’ drafts of their academic writing in accordance with the course materials in terms of an adapted model of Swales’ (Swales, 1990; Swales, 2004) move schema as partially presented in Table 1. IADE will achieve this by conducting a sentence-level classification of the input text for rhetorical shifts. Given a draft of a research article, IADE will identify the discourse moves in the paper, compare it with other papers in the same discipline and provide feedback to the user.

Move 1	Establishing a Territory
Step 1:	Claiming Centrality
Step 2:	Making topic generalization(s) and/or
Step 3:	Reviewing previous research
Move 2	Establishing a niche
Step 1A:	Indicating a gap or
Step 1B:	Highlighting a problem or
Step 1C:	Question-raising or
Step 1D:	Hypothesizing or
Step 1E:	Adding to what is known or
Step 1F:	Presenting justification
Move 3	Occupying the niche
Step 1A:	Announcing present research descriptively or
Step 1:	Announcing present research purposefully
Step 2A:	Presenting research questions or
Step 2B:	Presenting hypotheses
Step 3:	Definitional clarifications and/or
Step 4:	Summarizing methods and/or
Step 5:	Announcing principal outcomes and/or
Step 6:	Stating the value of the present research and/or
Step 7:	Outlining the structure of the paper

Table 1: Discourse move model for research article introductions based on (Swales, 1990; Swales, 2004)

The development of IADE is guided by the principles of Evidence Centered Design (ECD), “an approach to constructing and implementing educational assessments in terms of evidentiary arguments” (Mislevy et al., 2006, p. 15). This design allows the program to identify the discourse elements of students’ work products that constitute evidence and to characterize the strength of this evidence about the writing proficiencies targeted for the purpose of formative assessment.

3 Discourse Move Identification

3.1 Data and Annotation Scheme

The discussions above imply that the first step in the development of IADE is automatic identification of discourse moves in research articles. We have approached this task as a classification prob-

	Discipline	Files
1.	Accounting	20
2.	Aero-space engineering	20
3.	Agronomy	21
4.	Applied linguistics	20
5.	Architecture	20
6.	Biology	20
7.	Business	20
8.	Chemical engineering	20
9.	Computer engineering	20
10.	Curriculum and instruction	20
11.	Economics	20
12.	Electrical engineering and power system	20
13.	Environmental engineering	20
14.	Food science & food service	20
15.	Health & human performance	20
16.	Industrial engineering	20
17.	Journalism	20
18.	Mechanical engineering	20
19.	Sociology	20
20.	Urban and regional planning	20

Table 2: Disciplines represented in the corpus for article introductions

lem. In other words, given a sentence and a finite set of moves and steps, what move/step does the sentence signify? This task is very similar to identifying the discourse structure of short argumentative essays discussed in (Burstein et al., 2003), the difference being in the genre of the essays and type of the discourse functions in question.

The corpus used in this study was compiled from an existing corpus of published research articles in 44 disciplines, used in an academic writing graduate course for international students. The corpus contains 1,623 articles and 1,322,089 words. The average length of articles is 814.09 words. We made a stratified sampling of 401 introduction sections representative of 20 academic disciplines (see Table 2) from this corpus of research articles. The size of this sub-corpus is 267,029 words; each file is on average 665.91 words long, resulting in 11,149 sentences as data instances.

The sub-corpus was manually annotated based on Swales’ framework by one of the authors for moves

and steps (see Figure 1 for an example). The markup scheme includes the elements presented in Table 1. Annotation was performed at sentence level, each sentence being assigned at least one move and almost always a step within that move as specified in the markup scheme.¹ The scheme allowed for multiple layers of annotation for cases when the same sentence signified more than one move or more than one step. This made it possible to capture an array of the semantic shades rendered by a given sentence.

```
<intro_m3 step="description">
<intro_m3 step="method">
<intro_m3 step="purpose">
  This paper presents an
  application of simulation,
  multivariate statistics,
  and simulation metamodels
  to analyze throughput of
  multiproduct batch chemical
  plants.
</intro_m3>
</intro_m3>
</intro_m3>
```

Figure 1: A sample annotated sentence

3.2 Feature Selection

In order to classify sentences correctly, we first need to identify features that can reliably indicate a move/step. We have taken a text-categorization approach to this problem.² In this framework each sentence is treated as a data item to be classified, and is represented as an n -dimensional vector in the \mathcal{R}^n Euclidean space. More formally, a sentence s_i is represented as the vector $\bar{s}_i = \langle f_1, f_2, \dots, f_n \rangle$ where each component f_j of the vector \bar{s}_i represents a measure of feature j in the sentence s_i . The task of the learning algorithm is to find a function $F : S \rightarrow C$ that would map the sentences in the corpus S to classes in $M = \{m_1, m_2, m_3\}$ (where m_1 , m_2 , and m_3 stand for Move 1, Move 2, and Move 3, respectively). In this paper, for simplicity, we are assuming that F is a many-to-one function; however, it should be kept in mind that since sentences may

¹Only in two instances a step was not assigned.

²For an excellent review, see (Sebastiani, 2002).

signify multiple moves, in reality the relation may be many-to-many.

An important problem here is choosing features that would allow us to classify our data instances into the classes in question properly. In this study we focused on automatically identifying the major moves in the introduction section of research articles (i.e., m_1, m_2, m_3). Due to the sparseness of data, we have not attempted to identify the steps within the moves at this time.

We extracted word unigrams, bigrams and trigrams (i.e., single words, two word sequences, and three word sequences) from the annotated corpus. Subsection 3.5 reports the results of some of our experiments with these feature sets.

The following steps were taken in preprocessing:

1. All tokens were stemmed using the NLTK³ port of the Porter Stemmer algorithm (Porter, 1980). This allows us to represent lexically related items as the same feature, thus reducing interdependence among features and also helping with the sparse data problem.
2. All numbers in the texts were replaced by the string `_number_`.
3. In case of bigrams and trigrams, the tokens inside each n -gram were alphabetized to capture the semantic similarity among n -grams containing the same words but in a different order. This tactic also reduces interdependence among features and helps with the sparse data problem.
4. All n -grams with a frequency of less than five were excluded. This measure was also taken to avoid overfitting the classifier to the training data.

The total number of each set of n -grams extracted is shown in Table 3.

To identify which n -grams are better indicators of moves, odds ratios were calculated for each as follows:

$$OR(t_i, m_j) = \frac{p(t_i|m_j) \cdot (1 - p(t_i|\bar{m}_j))}{(1 - p(t_i|m_j)) \cdot p(t_i|\bar{m}_j)} \quad (1)$$

³<http://www.nltk.org>

<i>n</i> -gram	Number
unigrams	3,951
bigrams	8,916
trigrams	3,605

Table 3: Total number of *n*-grams extracted

where $OR(t_i, m_j)$ is the odds ratio of the term (*n*-gram) t_i occurring in move m_j ; $p(t_i|m_j)$ is the probability of seeing the term t_i given the move m_j ; and $p(t_i|\bar{m}_j)$ is the probability of seeing the term t_i given any move other than m_j . The above conditional probabilities are calculated as maximum likelihood estimates.

$$p(t_i|m_j) = \frac{\text{count}(t_i \text{ in } m_j)}{\sum_{k=1}^N \text{count}(t_k \text{ in } m_j)} \quad (2)$$

where N is the total number of *n*-grams in the corpus of sentences S .

Finally, we selected terms with maximum odds ratios as features. Subsection 3.5 reports on our experiments with classifiers using *n*-grams with highest odds ratios.

3.3 Sentence Representation

As mentioned in the previous subsection, each sentence is represented as a vector, where each vector component f_i represents a measure of feature i in the sentence. Usually, in text categorization this measure is calculated as what is commonly known as the tf.idf (term frequency times the inverse document frequency), which is a measure of the importance of a term in a document. However, since our “documents” are all sentences and therefore very short, we decided to only record the presence or absence of terms in the sentences as Boolean values; that is, a vector component will contain either a 0 for the absence of the corresponding term or a 1 for its presence in the sentence.

3.4 Classifier

We chose to use Support Vector Machines (SVM) for our classifier (Basu et al., 2003; Burges, 1998; Cortes and Vapnik, 1995; Joachims, 1998; Vapnik, 1995). SVMs are commonly used to solve classification problems by finding hyperplanes that best classify data while providing the widest margin possible

between classes. SVMs have proven to be among the most powerful classifiers provided that the representation of the data captures the patterns we are trying to discover and that the parameters of the SVM classifier itself are properly set.

SVM learning is a supervised learning technique where the system is provided a set of labeled data for training. The performance of the system is then measured by providing the learned model a set of new (labeled) data, which were not present during the training phase. The system then applies the learned model on the new data and provides its own inferred labels. The labels provided by the system are then compared with the “true” labels already available. In this study, we used a common technique known as *v*-fold cross validation, in which data are divided into *v* equal-sized groups (either by random sampling or by stratified sampling). Then, the system is trained on all but one of the groups and tested on the remaining group. This process is repeated *v* times until all data items have been used in training and validation. This technique provides a fairly accurate view of how a model built on the whole data set will perform when given completely new data. All the results reported in the following subsection are based on five-fold cross validation experiments.

We predominantly used the machine learning environment RAPIDMINER (Mierswa et al., 2006) in the experimentation phase of the project. The SVMs were set to use the RBF kernel, which maps samples into a higher dimensional space allowing for capturing non-linear relationships among the data and labels. The RBF kernel has two parameters, C and γ . These parameters help against overfitting the classifier on the training data. The values of these parameters is not known before hand for each data set and may be found through an exhaustive search of different parameter settings (Hsu et al., 2008). In this study, we used $C = 2^3$ and $\gamma = 2^{-9}$, which were arrived at through a search of different parameter settings on the feature set with 3,000 unigrams. The search was performed by performing five-fold cross validation on the whole data set using models built with various combinations of C and γ values. Admittedly, these parameters are not necessarily the best parameters for the other feature sets on which exhaustive searches should be performed. This is

the next step in our project.

3.5 Evaluation

We performed five-fold cross validation on 14 different feature sets as summarized in Table 4. The results of these experiments are summarized in Figures 2–4. Accuracy shows the proportion of classifications that agreed with the manually assigned labels. The other two performance measures, precision and recall, are commonly used in information retrieval, text categorization, and other NLP applications. For each category, precision measure what proportion of the items assigned to that category actually belonged to it, and recall measures what proportion of the items actually belonging to a category were labeled correctly. The measures reported here (macro-precision $\hat{\pi}^M$ and macro-recall $\hat{\rho}^M$) are weighted means of class precision and recall over the three moves.

$$\hat{\pi}^\mu = \frac{TP}{TP + FP} \quad (3)$$

$$\hat{\rho}^\mu = \frac{TP}{TP + FN} \quad (4)$$

$$\hat{\pi}^M = \frac{\sum_{i=1}^{|C|} w_i \hat{\pi}_i}{|C|} \quad (5)$$

$$\hat{\rho}^M = \frac{\sum_{i=1}^{|C|} w_i \hat{\rho}_i}{|C|} \quad (6)$$

The figures show that the unigram models result in the best recall and the trigram models, the best precision. Generally, we attribute lower recall to the sparseness of the data. Access to more training data will help improve recall. We should also note the behavior of the models with respect to bigram features. As seen on Figures 3 and 4, increasing the size of the bigram feature set causes a decline in model precision and a rise in model recall. Considering that there are far more frequent bigrams than unigrams or trigrams (cf. Table 4), this behavior is not surprising. Including more bigrams will increase recall because there are more possible phrases to indicate a move, but that will also result in a decline in precision because those bigrams may also frequently appear in other moves. It also seems that a model employing unigram, bigram and trigrams all will perform better than each individual model. We are planning to experiment with these feature sets, as well.

	Terms	N
1	Unigrams	1,000
2		2,000
3		3,000
4	Bigrams	1,000
5		2,000
6		3,000
7		4,000
8		5,000
9		6,000
10		7,000
11		8,000
12	Trigrams	1,000
13		2,000
14		3,000

Table 4: Feature sets used in experiments

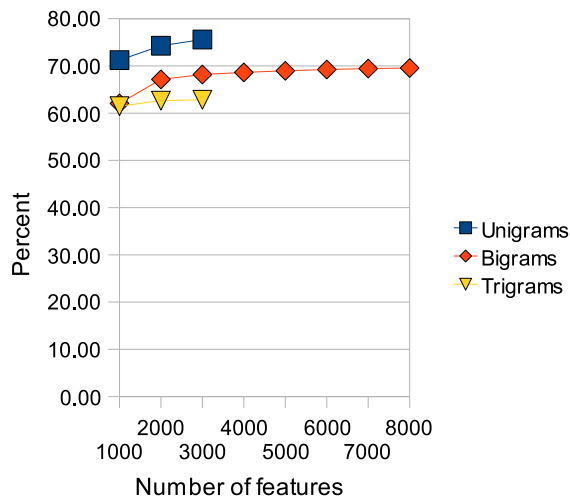


Figure 2: Model accuracy for different feature sets

Error analysis revealed that Move 2 is the hardest move to identify. It most frequently gets misclassified as Move 1. In the future, it might be helpful to make use of the relative position of the sentence in text in order to disambiguate the move involved. In addition, further investigation is needed to see what percentage of Move 2 sentences identified as Move 1 by the system also have been labeled Move 1 by the annotator. Recall that some of the sentences had multiple labels and in this study we are only considering single labels per sentence.

One question that might arise is how much infor-

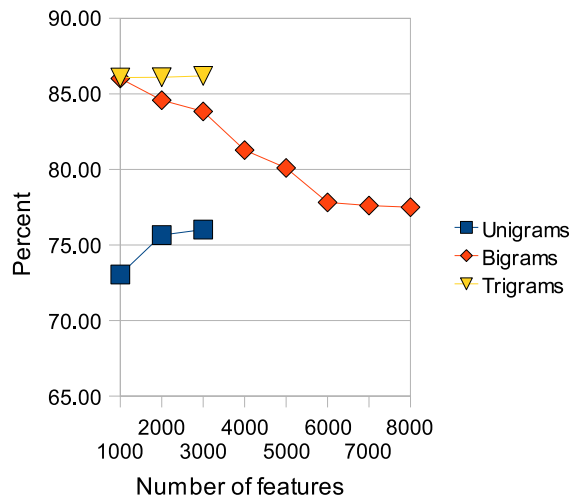


Figure 3: Model precision for different feature sets

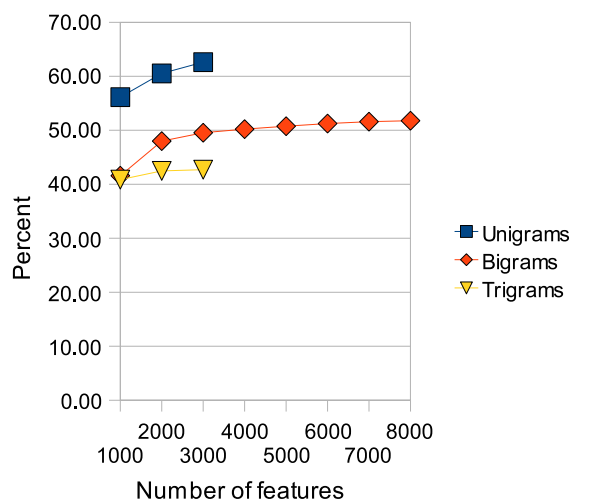


Figure 4: Model recall for different feature sets

mation about the discipline of the article contributes to classification accuracy. In other words, how discipline-dependent are our features? We also ran a set of experiments with the same features plus information about the scientific discipline in which each sentence was written. The change in system performance was not significant by any means, which suggests that our extracted features are not discipline-dependent.

3.6 Interannotator agreement

In order to get a clearer picture of the difficulty of the problem, we asked a second annotator to annotate a portion of the sub-corpus used in this study. The second annotations were done on a sample of files across all the disciplines adding up to 487 sentences. Table 5 contains a summary of the agreements between the two annotators.

	Move 1	Move 2	Move 3
No. agreed	457	452	480
$P(A)$	0.938	0.928	0.986
κ	0.931	0.919	0.984

Table 5: Interannotator agreement on 487 sentences.

Interannotator agreement κ , which is the probability of agreement minus chance agreement, is calculated as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (7)$$

where $P(A)$ represents observed probability of agreement, and $P(E)$ is the expected probability of agreement, i.e., chance agreement. Given three moves and uniform distribution among them, $P(E) = (\frac{1}{3})^2$. Therefore, the two annotators had an average κ of 0.945 over the three moves.

3.7 Limitations

This research is in its initial stages and naturally it has many limitations. One issue involves some of the choices we made in our experiments such as choosing to alphabetize the n -grams and choosing particular values for C and γ . We will be experimenting with non-alphabetized n -grams and also experimenting with different kernel parameters to find optimal models.

4 Discussion

This paper set out to identify rhetorical moves in research article introductions automatically for the purpose of developing IADE, an educational tool for helping international university students in the United States to improve their academic writing skills. The results of our models based on a relatively small data set are very encouraging, and research on improving the results is ongoing.

Apart from system accuracy, there are also some pedagogical issues that we need to keep in mind in the development of IADE. Warschauer and Ware (2006) call for the development of a classroom research agenda that would help evaluate and guide the application of automated essay scoring in the writing pedagogy. Based on a categorization developed by Long (1984), they propose three directions for research: product, process, and process/product, where “product refers to educational outcome (i.e., what results from using the software), process refers to learning and teaching process (i.e., how the software is used), and process/product refers to the interaction between use and outcome” (p. 10). On the level of evaluating technology for language learning in general, Chapelle (2007) specifies three targets for evaluation: “what is taught in a complete course”, “what is taught through technology in a complete course”, and “what is taught through technology” (p. 30). In the first case, an entire technology-based course is evaluated, in the second case, CALL materials used for learning a subset of course objectives, and in the third case, the use of technology as support and enhancement of a face-to-face course.

This project needs to pursue the third direction in both of these trends by investigating the potential of the IADE program specifically designed to be implemented as an additional component of a graduate course to improve non-native speaker students’ academic writing skills. Since this program will represent a case of innovative technology, its evaluation, as well as the evaluation of any other new CALL applications, according to Chapelle (2007), is “perhaps the most significant challenge teachers and curriculum developers face when attempting to introduce innovation into language education” (p. 30). Therefore, the analysis of the effectiveness of IADE will be conducted based on Chapelle’s (2001) framework, which has proven to provide excellent guidance for research of evaluative nature⁴.

References

- Laurence Anthony and George V. Lashkia. 2003. Mover: A machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communication*, 46(3):185–193.
- A. Basu, C. Watters, and M. Shepherd. 2003. Support vector machines for text categorization. In *HICSS '03: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4*, page 103.3, Washington, DC, USA. IEEE Computer Society.
- Christopher J. C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.
- Jill C. Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative english. In *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*, pages 68–75, Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, College Park, Maryland.
- Jill C. Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.
- Jill C. Burstein. 2003. The e-rater text registered scoring engine: Automated essay scoring with natural language processing. In Shermis and Burstein (Shermis and Burstein, 2003), pages 113–121.
- Carol Chapelle. 2001. *Computer applications in second language acquisition*. Cambridge University Press, New York.
- Carol Chapelle. 2007. Challenges in evaluation of innovation: Observations from technology research. *Innovation in Language Learning and Teaching*, 1(1):30–45.
- Martin Chodorow, Joel R. Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30.
- Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.
- Viviana Cortes. 2006. Exploring genre and corpora in the English for academic writing class. *Manuscript submitted for publication*. Manuscript submitted for publication.
- Scott Elliott. 2003. Intellimetric™: From here to validity. In Shermis and Burstein (Shermis and Burstein, 2003), pages 71–86.
- Jan Frodesen. 1995. Negotiating the syllabus: A learning-centered, interactive approach to ESL graduate writing course design. In Diane Belcher and George Braine, editors, *Academic Writing in a Second Language: Essays on Research and Pedagogy*, pages 331–350. Ablex Publishing Corporation, NJ.

⁴see (Jamieson et al., 2005)

- Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 2(2):115–129.
- Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2008. A practical guide to support vector classification. Unpublished manuscript.
- Joanne Jamieson, Carol Chapelle, and Sherry Preiss. 2005. CALL evaluation by developers, a teacher, and students. *CALICO Journal*, 23(1):93–138.
- T Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *In Proceedings of ECML-98, 10th European Conference on Machine Learning*.
- Shimona Kushner. 1997. Tackling the needs of foreign academic writers: A case study. *IEEE Transactions on Professional Communication*, 40:20–25.
- Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Shermis and Burstein (Shermis and Burstein, 2003), pages 87–112.
- Claudia Leacock and Martin Chodorow. 2003. Automated grammatical error detection. In Shermis and Burstein (Shermis and Burstein, 2003), pages 195–207.
- John Levis and Greta Muller-Levis. 2003. A project-based approach to teaching research writing to nonnative writers. *IEEE Transactions on Professional Communication*, 46(3):210–220.
- Michael Long. 1984. Process and product in ESL programme evaluation. *TESOL Quarterly*, 18(3):409–425.
- I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. 2006. YALE (now: RAPIDMINER: Rapid prototyping for complex data mining tasks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*.
- R. Mislevy, I. Steinberg, R. Almond, and J. Lukas. 2006. Concepts, terminology, and basic models of evidence-centered design. In D. Williamson, R. Mislevy, and I. Bejar, editors, *Automated scoring of complex tasks in computer-based testing*, pages 15–47. Lawrence Erlbaum Associates, Mahwah, NJ.
- Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses. In *Proceedings of the 6th International Computer Assisted Assessment Conference*, pages 233–249, Loughborough University.
- Ellis Batten Page. 2003. Project Essay Grade. In Shermis and Burstein (Shermis and Burstein, 2003), pages 43–54.
- Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.
- Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using Bayes’ theorem. *The Journal of Technology, Learning and Assessment*, 1(2):3–21.
- Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.
- Mark D. Shermis and Jill C. Burstein, editors. 2003. *Automated Essay Scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates, Mahwah, NJ.
- John Swales. 1990. *English in Academic and Research Settings*. Cambridge University Press, Cambridge.
- John Swales. 2004. *Research Genres: Exploration and applications*. Cambridge University Press, Cambridge.
- Roberta Vann and Cynthia Myers. 2001. Academic ESL options in a large research university. In Ilona Leki, editor, *Academic Writing Programs, Case Studies in TESOL Practice Series*. TESOL, Alexandria, VA.
- Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, Berlin.
- Mark Warschauer and Paige Ware. 2006. Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, 10(2):1–24.