Task-Independent Features for Automated Essay Grading

Conference Paper · June 2015

DOI: 10.3115/v1/W15-0626

Torsten Zesch
University of Duisburg-Essen

Michael Wojatzki
University of Duisburg-Essen

Task-Independent Features for Automated Essay Grading

Torsten Zesch **Michael Wojatzki**
Language Technology Lab
University of Duisburg-Essen

Dirk Scholten-Akoun
Center of Teacher Education
University of Duisburg-Essen

Abstract

Automated scoring of student essays is increasingly used to reduce manual grading effort. State-of-the-art approaches use supervised machine learning which makes it complicated to transfer a system trained on one task to another. We investigate which currently used features are task-independent and evaluate their transferability on English and German datasets. We find that, by using our task-independent feature set, models transfer better between tasks. We also find that the transfer works even better between tasks of the same type.

1 Introduction

Having students write an essay is a widely used method for assessment, e.g. universities use essay writing skills as a proxy for the prospects of applicants. As manually grading essays is costly, automated essay grading systems are increasingly used becasue they – once developed – do not introduce additional costs for grading new essays.

Automated essay grading systems usually follow a supervised approach and yield a quality of holistic grading comparable to human performance (Valenti et al., 2003; Dikli, 2006). These systems make use of certain properties of essays (called features) in order to estimate the essay quality. In the grading process, these features are extracted and ratings are assigned according to the manifestations of the features (Attali and Burstein, 2006). In order to automatically learn the association between feature values and ratings a high amount of manually rated essays is required for training. Hence, it seems desirable to develop systems that work without this initial input, which means – expressed in terms of machine learning – that features should not be defined by the present task but by general essay grading. A task is defined here as prompting a group of humans to solve a particular writing task. Tasks differ in attributes such as the grade-level of underlying subjects or characteristics of the prompt.

Many kinds of features have been proposed for essay grading (Valenti et al., 2003; Dikli, 2006). They differ in the degree of dependency to the task at hand. There are features that are strongly dependent on a task, e.g. when they detect important words or topics (Chen and He, 2013). Other features are less dependent, e.g. when they capture general characteristics of essays like the number of words in the essay (Östling, 2013; Lei et al., 2014), usage of connectors (Burstein and Chodorow, 1999; Lei et al., 2014), etc.

We assume that a system which considers only task-independent features should perform well no matter what task it is trained on. However, it is unclear how much explanatory power the model might lose in this step. In this paper, we test this hypothesis by performing experiments with a state-of-the-art essay grading system on English and German datasets. We categorize features into task-dependent and task-independent ones and evaluate the difference in grading accuracy between the corresponding models. We find that the task-independent models show a better performance for both languages tested, but the resulting losses are relatively high in general. Moreover, we examine the tasks more closely and group them according to whether they offer a textual source as a reference point. We show that the transfer works better if the model is derived from the same task type.

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 224–232,
Denver, Colorado, June 4, 2015.

2 Features

In this section, we describe state-of-the-art features and how they relate to the quality of an essay. For each feature, we discuss whether it belongs to the strongly task-dependent or weakly task-dependent group.

2.1 Length Features

This very simple but quite valuable feature deals with the **essay length** (Mahana et al., 2012; Chen and He, 2013; Östling, 2013; Lei et al., 2014). The core idea is that essays are usually written under a time limit. So the amount of produced text can be a useful predictor of the productivity of the writer and thus the quality of the essay (Shermis and Burstein, 2002). Therefore, we measure the text length by counting all tokens and sentences in an essay. The degree of task-dependence of this feature is directly connected to the time limit.

The average **sentence length** in words and **word length** in characters can be an indicator for the degree of complexity a writer can master (Attali and Burstein, 2006; Mahana et al., 2012; Chen and He, 2013; Östling, 2013). As this is not particularly tied to a specific task, these features are weakly task-dependent.

2.2 Occurrence Features

According to Mahana et al. (2012) the occurrences of linguistic phenomena such as **commas, quotations**, or **exclamation marks** can serve as valuable features in a grade prediction. These features focus more on the structuring of an essay and are thus weakly task-dependent.

For tasks that are source-based (i.e. a source text is provided on which the task is based), we augment this approach by also counting **formal references** like citations and line references. Source-based features are obviously strongly task-dependent.

Using third party sources to support an argument can be a valuable hint for evidence (Bergler, 2006). Therefore, we use the approach of Krestel et al. (2008) to detect **direct, indirect**, and **reported speech** in essays. The approach relies on set of reporting verbs and rules to identify and distinguish these forms.

If a task is based on a certain text source, the occurrence of **core concepts** in the essay should be an indicator for high quality (Foltz et al., 1999). We determine core concepts from the source using words or phrases with a high tf.idf weight. Again these features are just meaningful if the related task offers a textual source.

2.3 Syntax Features

Variation in the syntactic structures used in an essay may indicate proficiency in writing (Burstein et al., 1998). Following Chen and He (2013), we operationalize this by measuring the ratio of distinct parse trees to all the trees and the average depths of the trees to compute **syntactic variation** features.

Further, the parsing trees are used to measure the proportion of **subordinate, causal** and **temporal clauses**. Causal and temporal clauses are detected by causal or temporal conjunctions that could be found in subordinate-clauses. For example, a subordinate clause beginning with *when* is considered as temporal. The detection of causal- and temporal clauses is used to enrich the syntactic variability by a discourse element (Burstein et al., 1998; Chen and He, 2013; Lei et al., 2014). As syntactic features are relatively independent of the task, we categorize them as weakly task-dependent.

2.4 Style Features

Another important aspect of essay quality is an appropriate style. Following Östling (2013), we use the relative ratio of POS-tags to detect style preferences of writers. We complemented this by a feature that measures the formality F of an essay (Heylighen and Dewaele, 2002) defined as:

$$F = \frac{\sum\limits_{i \in A} \frac{c(i)}{n} - \sum\limits_{j \in B} \frac{c(j)}{n} + 100}{2}$$

where A = {N, ADJ, PP, DET}, B = {PR, V, ADV, UH}, and n is the number of tokens in the text. The **formality**-feature should be strongly task-dependent, as the correct style depends on the task and the target audience.

The words used in the essay tell us something about the vocabulary the writer actively uses. In accordance with Chen and He (2013), we measure the

type-token-ratio to estimate whether an essay has a relatively rich or rather poor vocabulary.

As noted by Breland et al. (1994), word knowledge of a writer is highly tied to the corpus frequency of the words used. The lower the frequency the higher the writer's language proficiency. We model this idea by calculating the average **word frequency** in the Web1T-corpus (Brants and Franz, 2006). We expect this average frequency to be relatively stable and thus categorize the feature as weakly task-dependent.

2.5 Cohesion Features

The structure of an essay reflects the writer's ability to organize her ideas and compose a cohesive response to the task. Following Lei et al. (2014) the use of **connectives** (like *therefore* or *accordingly*) can be a hint for a cohesive essay. We count occurrences of connectives (from a fixed list) and normalize by the total number of tokens. As cohesion is relatively independent from the topic of an essay, we categorize this feature as weakly task-dependent.

2.6 Coherence Features

In order to make an essay understandable, writers need to ensure that the whole text is coherent and the reader can follow the argumentation (Chen and He, 2013; Lei et al., 2014). Features based on Rhetorical Structure Theory (William and Thompson, 1988) could be used (Burstein et al., 2001), but there are no reliable parsers available for German and performance is also not yet robust enough for English. Instead, we operationalize coherence measuring the **topical overlap** between adjacent sentences. We use similarity measures based on n-gram overlap and redundancy (e.g. of nouns). This operationalization of coherence is weakly task-dependent, as the degree of topical overlap is independent of the actual topic.

2.7 Error Features

Grammatical or spelling errors are one of the most obvious indicators of bad essays, but have been found to have only little impact on scoring quality (Chen and He, 2013; Östling, 2013). We add a simple rule-based **grammar error** feature in our system based on LanguageTool.[1] We do not expect grammar errors to be bound to specific topics and categorize the feature as weakly task-dependent.

2.8 Readability Features

We use a set of established **readability** features (Flesch, Coleman-Liau, ARI, Kincaid, FOG, Lix, and SMOG), that rely on normalized counts of words, letters, syllables or other phenomena (like abbreviations) which affect the readability (McLaughlin, 1969; McCallum and Peterson, 1982; Smith and Taffler, 1992). Depending on which writing style is considered as appropriate, high scoring essays might be associated with different levels of readability. However, a certain level of formal writing is required for most essays and very simple or very complex writing are both indicators for bad essays. Thus, we categorize the features as weakly task-dependent.

2.9 Task-Similarity Features

For source-based essays, we can determine the **task similarity** of an essay by computing the similarity between essay and the task specific source (Östling, 2013). There should be a certain degree of similarity between the source and the essay, but if the similarity is too high the essay might be plagiarized. We use Kullback–Leibler divergence between source and essay.

A variant of this feature computes the **corpus similarity** to a neutral background corpus (Brown corpus (Marcus et al., 1993) in our case) in order to determine whether the essay was written specific enough.

While the corpus similarity should be weakly task-dependent, the task similarity is of course strongly dependent on the task.

2.10 Set-Dependent Features

So far, all features have only used the characteristics of a single essay, but it is also useful to take the whole set of essays into account. Instead of detecting characteristics of an individual essay the differences between essays in the set is examined. Set-based features can be based on topics (Burstein et al., 1998) or n-grams (Chen and He, 2013). We use **word n-gram** features for the 1,000 most frequent uni-, bi- and tri-grams in the essay set. Following

[1] https://www.languagetool.org

Chen and He (2013), we further use the same number of **POS n-grams** as features.

As a consequence of writing conventions, wording in an essay usually differs between regions in a text. For example, words that indicate a summary or a conclusion are indicators for a good essay only if they occur at the end, not at the beginning. Thus, we partition the essay in n equally sized parts based on word counts (we found five parts to work well) and compute **partition word n-grams** using the same settings as described above.

As all features described in this section deal with frequent wording or essay topics, they are strongly task-dependent.

3 Experimental Setup

We now describe the experimental setup used to examine our research question regarding task-independent models.

3.1 Datasets

As we want to compare models across tasks, we need datasets that contain different tasks.

English A suitable English dataset is the ASAP essay grading challenge.[2] The dataset contains eight independent tasks of essay-writing with each about 1,800 graded essays (except the last one with only 723). The essays were written by students in grade levels between 7 and 10 of a US high-school. The tasks cover a wide range of different settings and can be grouped on whether they were source-based or not:

The **source-based tasks** have in common that the participants first received a text as input and then had to write an essay that refers to this source. The following task belong to this group:

- Task 3: Given a source of someone who is traveling by bicycle, students should describe how the environment influences the narrator.
- Task 4: On the basis of the text 'winter hibiscus' participants should explain why the text ends in a particular way.
- Task 5: Students were requested to describe the mood of a given memoir.
- Task 6: Based on an excerpt on the construction of the Empire State Building, participants had to describe the obstacles the builders faced.

The **opinion tasks** ask for an opinion about a certain topic, but without referring to a specific source text.

- Task 1: Students should convince readers of a local newspaper of their opinion on the effects computers have on people.
- Task 2: Participants were asked to write about their opinion on whether certain media should be banned from libraries. They were prompted to include own experiences.
- Task 7: Participants should freely write on 'patience'. They could either write entirely free or about a situation in which they or another person proved patience.
- Task 8: Participants were told to tell a true story in which laughter was a part.

As the different tasks use different scoring schemes, we use holistic scores and normalize to a scale from 0 to 9 in order to make the trained model exchangeable.

German The German dataset contains two independent tasks each with 197 and 196 annotated essays. The essays were written by first-year university students of degree programs for future teachers. Both writing tasks had in common that the participants first received a text as an input. After reading the given text they were supposed to write an essay by summarizing the argumentative structure of the text. However, students were also asked to include their own pro and contra arguments.

- T1: Students were requested to summarize and to discuss a newspaper article of a national German newspaper which deals with an educational topic.
- T2: Participants were asked to summarize and to discuss a newspaper article of a national German newspaper which focusses on the quality of contributions in the participatory media.

Again, we use the holistic scores. No normalization was necessary as both tasks use the same 6-point scoring scale.

[2] https://www.kaggle.com/c/asap-aes

Group	Feature
strongly task-dependent	essay length partition word n-gram POS n-gram word n-gram
	**core concepts* **formal references* **task similarity*
weakly task-dependent	connectives commas/quotations/exclamation corpus similarity direct, indirect and reported speech formality grammar error readability subordinate, causal & temporal clauses syntactic variation topical overlap type-token-ratio word frequency word/sentence length

Table 1: List of features grouped into strongly and weakly task-dependent. Source-based features (marked with a *) are not used in our experiments.

3.2 Strongly vs. Weakly Dependent Features

Our theoretic considerations on the commonly used features show that they differ in their degree of dependence on a specific essay writing task. As not all tasks refer to a source, we exclude – for the sake of comparability – features that rely heavily on the source text, i.e. features like core concepts. We argue that set-dependent features are strongly task-dependent and most others are weakly dependent. Table 1 gives an overview of the two feature groups used in our experiments. The **full** feature set uses both strongly and weakly task-dependent features, while the **reduced** set only uses the weakly task-dependent ones.

3.3 Essay Grading System

In order to ensure a fair comparison, we re-implemented a state-of-the-art essay grading system based on DKPro TC (Daxenberger et al., 2014)[3] which ensures easy reproducibility and replicability.

Our system takes a set of graded essays and performs preprocessing using tokenization, POS-tagging, stemming, and syntactic parsing.[4] The feature extraction takes a list of features (either the **full** or **reduced** set of features) and extracts the corresponding feature values from the instances. The machine learning algorithm[5] then learns a model of essay quality from the extracted features.

In a second and independent step, the learned model is applied in order to grade essays. In the usual in-task setting (our baseline), we train on a part of the available data for a specific essay writing task and then evaluate on the held-out rest (10-fold cross validation). In our task-adaptation setting, we train the model on all the data for one task, but evaluate on another task.

For the German essays, we need to adapt some components of the system. For example, the lists of **connectives**, **causal** and **temporal** clause detection were replaced by German equivalents. The detection of **direct, indirect, and reported speech** was done following Brunner (2013). Further, **corpus similarity** was computed based on the Tiger corpus (Brants et al., 2004) instead of the Brown corpus, and the **word frequency** was calculated using the German part of Web1T. In all other aspects, the English and German setups are equal.

3.4 Evaluation Metric

Following the recommendation of the ASAP challenge, we use as evaluation metric **quadratic weighted kappa** computed as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

with $O_{i,j}$ as the number of times one annotator graded j and the other i, with $E_{i,j}$ as the expected grades given a random distribution and with

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

as the weight of the grades. The metric produces a value for the agreement between the human gold standard and the machine grading.

[3] version: 0.7

[4] The preprocessing was realized with the DKPro Core 1.7.0 components used within DKPro TC: BreakIterator, TreeTagger, SnowballStemmer and StanfordParser.

[5] Support Vector Machine provided by DKPro TC

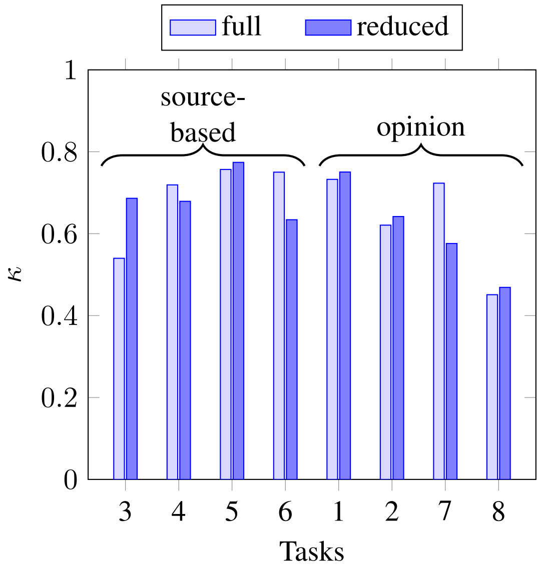

Figure 1: ASAP dataset: Comparison of the full and reduced model.

4 Results

We now report and discuss the results of our task adaptation experiments. The difference in performance will be an indicator of how well the models can be transferred from one essay set to another. We first establish the within-task results as a baseline and then compare them with the cross-task results.

4.1 Baseline: Within-Task Models

Figure 1 gives an overview of the results obtained when training a dedicated model for each task, either with the strongly task-dependent full model or the weakly task-dependent reduced model. Task8 shows very low performance due to the much smaller amount of available training data. We expected that the full model would always perform better than the reduced model, but we get a mixed picture instead. It seems that even within a task, the full feature set overfits on specific words used in the training data while they do not need to be necessarily mentioned in order to write a good essay.

Figure 2 shows the results for the German essays. The kappa values are much lower than for the English essays. This can be explained by the fact that the German tasks focus more on content issues than on language proficiency aspects, as the German essays are targeted towards university students compared to school students for the English essays. As content issues are hardly covered by our features, the results could probably be improved by adding content features like the occurrence of core concepts (see 2.2). However, for both German tasks we see the expected drop in performance when going from the full to the reduced model although it is rather small.

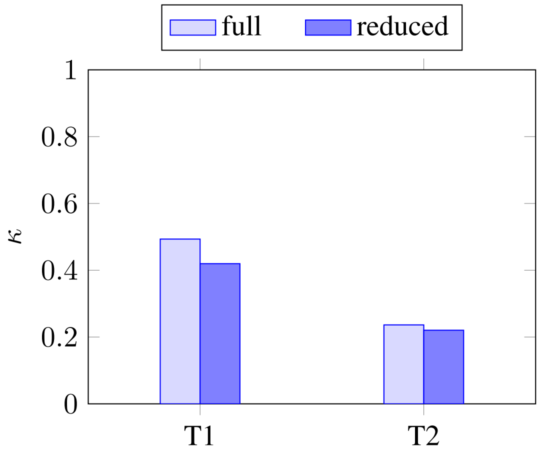

Figure 2: German dataset: Comparison of the full and reduced model

After having calculated the baselines, we can now transfer the models and determine the loss associated with the transfer.

4.2 Experiment: Cross-Task Models

We now examine the task-adaptivity of models by training on one task and testing on another, and then compare the result to the baseline established above.

Table 2 shows the resulting loss in performance for the full model. The table rows represent the tasks on which the model has been trained and the columns the tasks on which the trained model was tested. The average loss over all model transfers is .42, which shows that the full models do not work very well when transferred to another task.[6] For most cases, the observed behavior is symmetric, i.e. we see a similar drop when training on task 5 and testing on 4 or training on 4 and testing on 5. Though there are some remarkable exceptions. The model

[6] Note that the average loss in terms of quadratic weighted kappa is not equal the mean, as Fishers-Z transformation (Fisher, 1915) has to be performed before averaging variance ratios like quadratic weighted kappa.

		Opinion				Source-based			
		3	4	5	6	1	2	7	8
Opinion	3	-	-0.41	-0.24	-0.44	-0.42	-0.56	-0.34	-0.43
	4	-0.35	-	-0.48	-0.48	-0.35	-0.55	-0.38	-0.41
	5	-0.41	-0.47	-	-0.55	-0.13	-0.46	-0.25	-0.35
	6	-0.46	-0.59	-0.61	-	-0.43	-0.36	-0.45	-0.13
Source-based	1	-0.45	-0.60	-0.55	-0.65	-	+0.01	-0.37	-0.12
	2	-0.46	-0.60	-0.60	-0.63	-0.40	-	-0.61	-0.10
	7	-0.39	-0.49	-0.42	-0.53	-0.28	-0.19	-	-0.19
	8	-0.41	-0.53	-0.50	-0.60	-0.52	-0.24	-0.33	-

Table 2: Loss of the **full** models compared with using the tasks own model (loss >-0.3 highlighted)

		Opinion				Source-based			
		3	4	5	6	1	2	7	8
Opinion	3	-	-0.11	-0.29	-0.25	-0.66	-0.61	-0.31	-0.46
	4	-0.04	-	-0.24	-0.24	-0.67	-0.60	-0.29	-0.46
	5	-0.23	-0.18	-	+0.03	-0.54	-0.60	-0.16	-0.44
	6	-0.41	-0.34	-0.24	-	-0.39	-0.57	-0.06	-0.40
Source-based	1	-0.54	-0.43	-0.45	-0.37	-	-0.12	-0.07	-0.20
	2	-0.48	-0.40	-0.48	-0.43	-0.35	-	-0.36	-0.05
	7	-0.54	-0.39	-0.39	-0.38	-0.09	-0.28	-	-0.25
	8	-0.56	-0.49	-0.57	-0.50	-0.49	-0.25	-0.31	-

Table 3: Loss of the **reduced** models compared with using the tasks own model (loss >-0.3 highlighted)

trained on set 1 performs even better on set 2 than its own model, while training on set 2 and testing on set 1 results in a .4 drop. In addition, all source-based models (1, 2, and 7) work quite well as models for set 8 – the drop is only about .1 in all those cases. However, set 8 has relatively little training data so that this might be rather an effect of the other models being generally of higher quality than a task transfer effect.

The same procedure was carried out for the model with the reduced feature set that excludes task-dependent features. The results are shown in table 3. We see that the average loss is reduced (.36 compared to .42 for the full model) which is in line with our hypothesis that the reduced feature set should transfer better between tasks. However, the effect is not very strong when averaged over all tasks.

We also observe noticeable difference in the transferability between the groups (source-based vs. opinion tasks). Looking only within the source-based tasks the loss falls between +.03 and -.41, while for training on the opinion tasks and yields much higher losses (from -.37 to -.57). The same effect can be found for the opinion tasks (with the exceptions of set 7). In order to better see the difference, we show the average loss for each group in table 4. It is obvious that a transfer within source-based or opinion tasks works much better than across the groups. Within a group, the loss is only half as big as between groups.

	Opinion	Source-based
Opinion	-0.22	-0.46
Source-based	-0.47	-0.23

Table 4: Average loss of reduced model by task type

We perform the same set of experiments on the German data set. The results of the full model are shown in table 5a and the results of the reduced model are shown in figure 5b. Again the losses of the reduced model are much smaller than of the full model confirming our results on the English dataset.

5 Conclusion

In this work, we investigated the research question to what extend supervised models for automatic essay

	T1	T2
T1	-	-0.15
T2	-0.47	-

(a) Full

	T1	T2
T1	-	-0.07
T2	-0.28	-

(b) Reduced

Table 5: Loss on the German dataset

grading can be transferred from one task to another. We discussed a wide range of features commonly used for essay grading regarding their task dependence and found that they can be categorized into strongly and weakly task-dependent. Our hypothesis was that the latter model should transfer better between tasks. In order to test that, we implemented a state-of-the-art essay grading system for English and German and examined the task transferability by comparing the baseline performance (training on the actual task) with the models trained on the other tasks. We found, consistent with our hypothesis, that the reduced models performed better on average. The transfer worked even better if the underlying tasks are similar in terms of being source-based or opinionated. The fact that the losses on average are still quite high raises the question of whether a more fine-grained discrimination of features is necessary or whether models for essay grading can be transferred at all.

In future work we plan to further investigate the connection of task attributes to their task-transferability (e.g. the language proficiency level of participants or differences in the task description). In addition, we think that there are facets of quality that are independent of tasks, like the complexity of essays. Grading essays not only holistically, but according to facets is likely to transfer better between tasks and at the same time provides teachers with reliable sub-scores that may support their decisions without the demand of training data.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Sabine Bergler. 2006. Conveying attitude with reported speech. In *Computing attitude and affect in text: Theory and applications*, pages 11–22. Springer.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram corpus version 1.1. *Google Inc*.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2(4):597–620.

Hunter M Breland, Robert J Jones, Laura Jenkins, Marion Paynter, Judith Pollack, and Y Fai Fong. 1994. The college board vocabulary study. *ETS Research Report Series*, 1994(1):i–51.

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and linguistic computing*, 28(4):563–575.

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative english speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, pages 68–75. Association for Computational Linguistics.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. Automated scoring using a hybrid feature identification technique. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 206–210. Association for Computational Linguistics.

Jill Burstein, Claudia Leacock, and Richard Swartz. 2001. *Automated evaluation of essays and short answers*. Loughborough University Press.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *EMNLP*, pages 1741–1752.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. Dkpro tc: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland, June. Association for Computational Linguistics.

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Ronald A Fisher. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, pages 507–521.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. Automated essay scoring: Applications to educational technology. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, volume 1999, pages 939–944.

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340.

Ralf Krestel, Sabine Bergler, René Witte, et al. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. *Reporter*, 1(5):4.

Chi-Un Lei, Ka Lok Man, and TO Ting. 2014. Using learning analytics to analyze writing skills of students: A case study in a technological common core curriculum course. *IAENG International Journal of Computer Science*, 41(3).

Manvi Mahana, Mishel Johns, and Ashwin Apte. 2012. Automated essay grading using machine learning. *Mach. Learn. Session, Stanford University*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Douglas R McCallum and James L Peterson. 1982. Computer-based readability indexes. In *Proceedings of the ACM'82 Conference*, pages 44–48. ACM.

G Harry McLaughlin. 1969. Smog grading: A new readability formula. *Journal of reading*, 12(8):639–646.

Robert Östling. 2013. Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.

Mark D Shermis and Jill C Burstein. 2002. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Malcolm Smith and Richard Taffler. 1992. Readability and understandability: Different measures of the textual complexity of accounting narrative. *Accounting, Auditing & Accountability Journal*, 5(4):0–0.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330.

Mann William and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281.