Essay complexity evaluation

1. Historical background

- Ellis Page's *The imminence of grading essays by computer* (1966): "the technology as a tool for instructors who were burdened with hours of grading writing assignments."
- Project Essay Grade (**PEG**) (Ajay, Tillett, & Page, 1973). Although reported as stable, the project faced the lack of computer access and a time-consuming method for data entry.
- Automated Essay Evaluation (AEE) better known as a concept **since the 1990s**.
- **The main factors** facilitating the development of AEE: (a) the creation and widespread adoption of the Internet, (b) word processing software, and (c) the use of NLP in the evaluation of writing.
- Page and Petersen (1995) summarized their approach in their Phi Delta Kappan article, "The computer moves into essay grading: Updating the ancient test," at which point they introduced **the terminology of proxes and trins.** <...> trins represent the characteristic dimension of interest such as fluency or grammar whereas proxes (taken from approximations) are the observed variables with which the computer works. These are the features that a computer might extract from a text.
- At about the same time Vantage Learning released its first version of their **IntelliMetric** scoring engine (Elliot, 1999). The IntelliMetric scoring engine analyzes more than 300 semantic, syntactic, and discourse level features that fall into five major categories: focus and utility, development and elaboration, organization and structure, sentence structure, and mechanics and conventions.
- **Latent Semantic Analysis (LSA)** (Landauer, Foltz, & Laham, 1998; Landauer, Laham, & Foltz, 2003). LSA is a corpus-based statistical modeling approach that uses large corpora to model word usage. LSA generates information about statistical word usage in a document without regard to word ordering. This information for a single document is compared to other documents to determine the similarity between that source document and a large set of reference documents in terms of vocabulary use.
- The main problem of the AEE systems until recently: **lack of publicly-available AEE engines**, although strongly supported financially (see, e.g., *The Algorithm Didn't Like My Essay* in The NY Times).
- The contemporary ecological model for teaching and assessing writing is congruent with **the description given by Yancey (2012)** of the present state of writing assessment. Yancey sees the history of writing assessment as a series of phases, often overlapping: the period from 1950 to 1970 highlighted multiple-choice tests (Palmer, 1961); from 1970 to 1986, attention was focused on the holistic essay (White, 1985); and from 1986 to the present, portfolio assessment was viewed as the optimal evaluation method (Hamp-Lyons & Condon, 2000).

2. Some of the issues in ESL essay evaluation

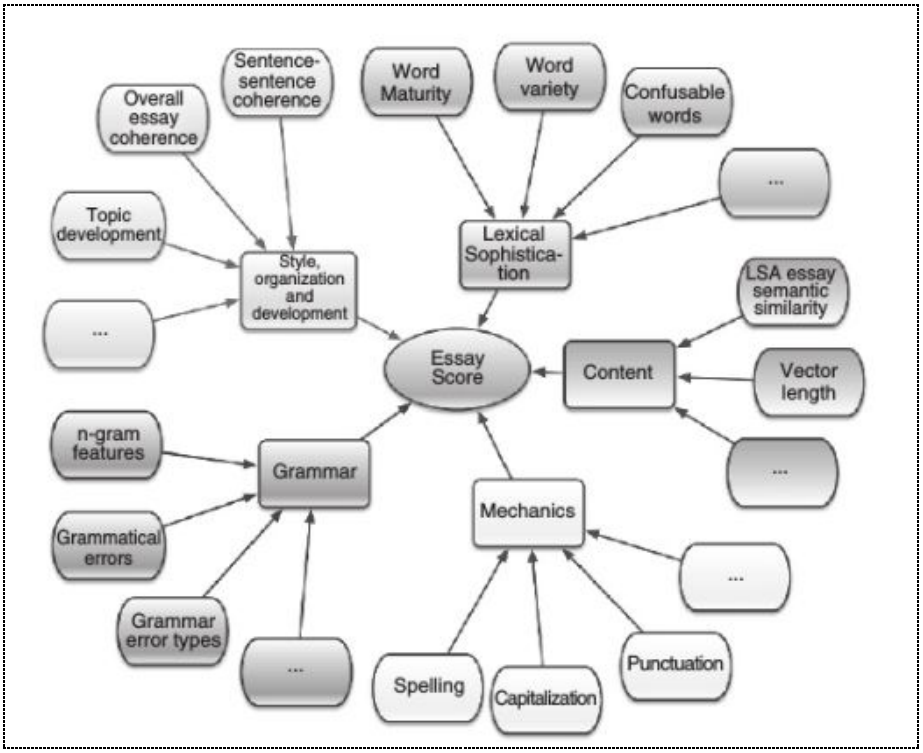
- Both Grimes and Warschauer (2010) and Chen and Cheng (2008) report that teachers and students believed that the scoring algorithm favored the traditional five-paragraph essay and that the scoring algorithms **discouraged creativity in writing**, suggesting that students wanting to get higher scores may be less likely to take risks in their writing.
- Crossley and McNamara (2011) found **significant differences between native and non-native** college-level writers on four word-based indices (hypernymy, polysemy, lexical diversity, and stem overlap), suggesting that these indices may be useful indicators of native vs. ESL writing. Similarly, Friginal and Weigle (2012) found that co-occurrence of specific language features such as agentless passives, attributive adjectives, and lack of other features such as second person pronouns, mental verbs, and that-complement clauses were associated with higher essay scores.
- As for the human side of the equation, it has been shown that failures in the implementation of educational technology are often due to **teacher resistance** rather than problems with technology (Curan, 2003, 2005, cited in Grimes & Warschauer, 2010).

3. Currently used models and issues

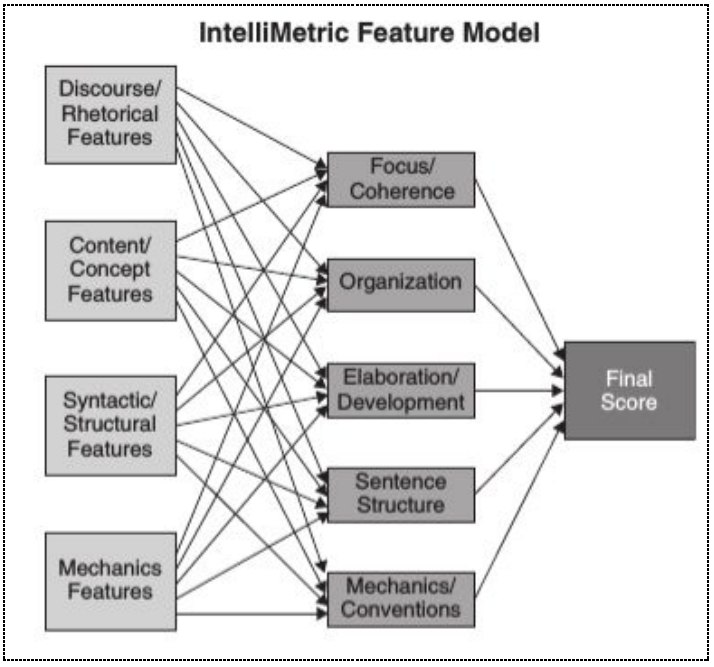
At the end of 2012: PEG, e-rater, IntelliMetric, CRASE, LightSIDE, AutoScore, IEA, Bookette, Lexile Writing Analyzer

Used now:

- The E-rater® Automated Essay Scoring System
- The Intelligent Essay Assessor



- The IntelliMetric™ Automated Essay Scoring Engine



- West Virginia Writes™

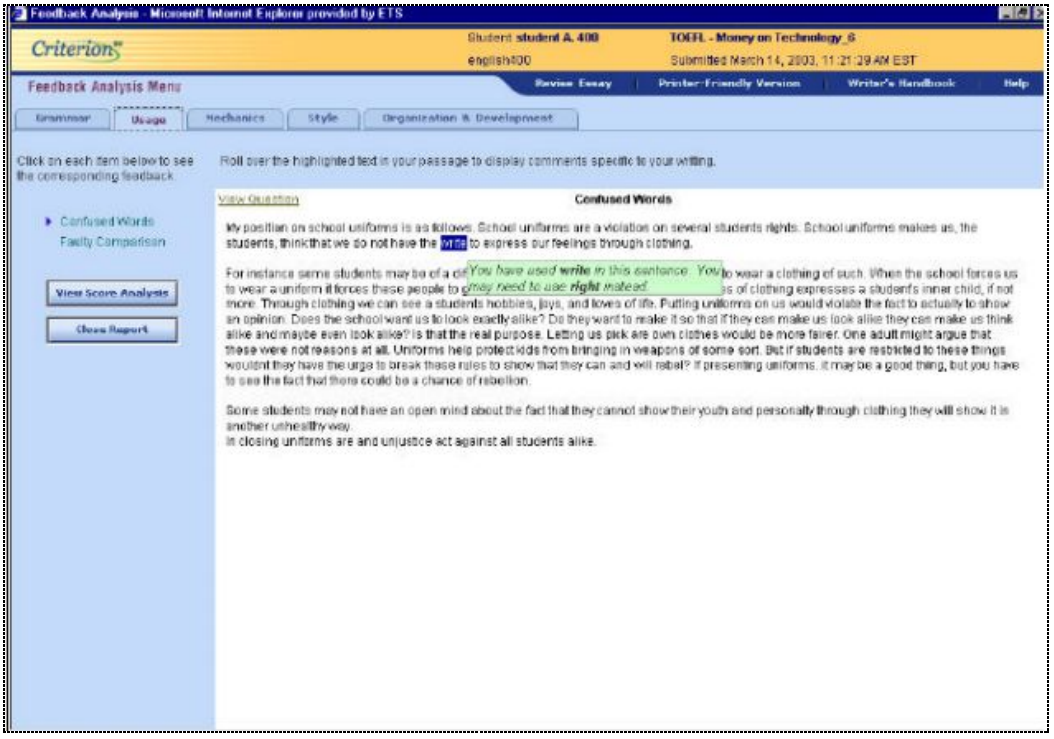
CTB AEE Grammar Error Detection

The CTB Bookette system has an embedded NLP-based grammar and style error detection tool. The system detects the following eight major types of common writing errors:

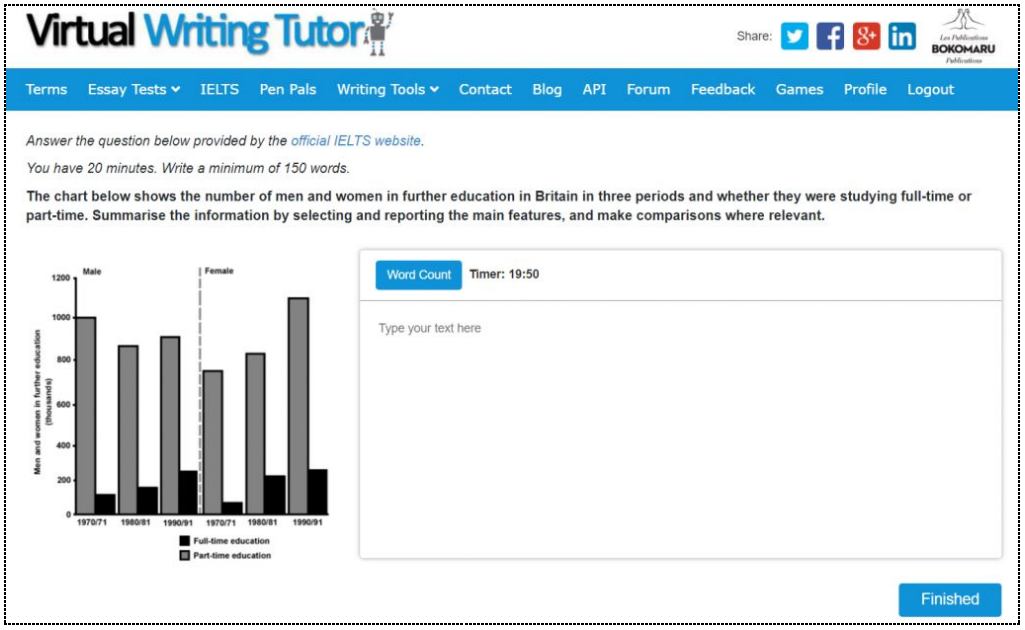
1. Agreement errors (e.g., The school for boys *were* established.).
2. Missing words (e.g., Beijing is *(the)* capital of China.).
3. Extra words (e.g., You should *be* try this.).
4. Wrong words (e.g., Listen to *she*!).
5. Confusion of similar words (e.g., then versus than).
6. Wrong word order (e.g., How you *are*?).
7. Punctuation errors and white space errors (i.e., additional space between words and sentences).
8. Spelling errors.

- LightSIDE − Open Source Machine Learning for Text...
- The Writing-Pal...
- Criterion™ Online Essay Evaluation; article (2003)
  "Criterion has two complementary applications: E-rater, an automated essay scoring system and Critique Writing Analysis Tools, a suite of programs that detect errors in

grammar, usage, and mechanics, that identify discourse elements in the essay, and that recognize elements of undesirable style."



- Virtual Writing Tutor including Automated IELTS Essay Evaluation



+ 6 Essay rating tools on the Virtual Writing Tutor

- MY Access for schools
- PEG Writing® – MI Write

4. Overviews

- ➢ Some of the models contrasted (The Handbook of Automated Essay Evaluation, 313−343).
- ➢ An overview in the dissertation by K. Zupanc:

AEE System	Attr. Types	Methodology	Prediction Model
PEG [Page]	Style	Statistical	multiple linear regression
PS-ME [Mason and Grove-Stephenson]		NLP	linear regression
e-rater [Burstein et al.]	Style & Content	NLP	linear regression
IntelliMetric [Schultz]			multiple mathematical models
Bookette [Rich et al.]			neural networks
OzEgrader [Fazal et al.]			
CRASE [Lottridge et al.]			machine learning
AutoScore [Shermis and Hamner]			statistical model
Lexile [Smith et al.]			Lexile measure
Ranked-based AEE [Chen et al.]			learning to rank
Multi-classifier Fusion AEE [Bin and Jian-Min]			ensemble classifiers
BETSY [Rudner and Liang]		Statistical	Bayesian networks
SEAR [Christie]			linear regression
Neural Essay Assessor [Taghipour and Ng]		Deep learning	recurrent neural networks
AES using NN [Alikaniotis et al.]			
AEG using MN [Zhao et al.]			memory networks
LightSIDE [Mayfield and Rosé]	Content	Statistical	machine learning
IEA [Foltz et al.]		LSA, NLP	
Semantic-tree-based AEE [Chali and Hasan]		LSA, tree kernel functions	cosine similarity
GLSA based AEE [Islam and Hoque]		GLSA	
Markit [Williams and Dreher]		NLP, PMT	linear regression
SAGrader [Brent et al.]	Semantic	FL, SN	rule-based expert systems
OBIE-based AEE [Gutierrez et al.]		OIE, DL	/
SAGE		OIE, NLP	random forest

Accuracy comparison of various systems from the literature and results from the Kaggle competition.

System	Avg. acc.	rank
SAGE	0.8325	1
Sollers & Gxav*	0.8014	2
SirGuessalot & PlanetThanet & Stefan*	0.7986	3
VikP & jman*	0.7978	4
Efimov+Berengueres*	0.7956	5
@ORGANIZATION*	0.7947	6
PEG [Page, 1994]	0.7888	7
Martin*	0.7857	8
cs224u*	0.7828	9
jackpot (Jason)*	0.7826	10
e-rater [Burstein et al., 2013a]	0.7656	11
IntelliMetric [Schultz, 2013]	0.7588	12
CRASE [Lottridge et al., 2013]	0.7494	13
LightSIDE [Mayfield and Rosé, 2013]	0.7494	14
Ranked-based [Chen et al., 2012]	0.7363	15
AutoScore [Shermis and Hamner, 2013]	0.7325	16
IEA [Foltz et al., 2013]	0.7344	17
Bookette [Rich et al., 2013]	0.6981	18
Lexile [Smith et al., 2014]	0.6331	19

* Results were obtained from the leader board of AES competition on Kaggle website[5].

5. Alternative modern approaches

- A Neural Approach to Automated Essay Scoring (2016)
  "Traditional automated essay scoring systems rely on carefully designed features to evaluate and score essays. The performance of such systems is tightly bound to the quality of the underlying features. <...> In this paper, we develop an approach based on recurrent neural networks to learn the relation between an essay and its assigned score, without any feature engineering. <...>The results show that our best system, which is based on long short-term memory networks, outperforms a strong baseline by 5.6% in terms of quadratic weighted Kappa, without requiring any feature engineering."
- Semantics-based automated essay evaluation (2018)
  Among the problems stated: Weak consideration of text semantics; Reproduction of biased human grades; Lack of standards and good practice (due to predominance of proprietary systems).
  SAGE − Semantic Automated Grader for Essays

6. AEE assessment criteria

1. **Validity and Reliability of AEE** (The Handbook of Automated Essay Evaluation, 181−195). General issues.
2. **Scaling and Norming for AEE** (The Handbook of Automated Essay Evaluation, 196−216). General issues.
3. **Grammatical Error Detection in AES and Feedback** (The Handbook of Automated Essay Evaluation, 251−264) − directly applicable to our models.

Average ranks of 50 most relevant attributes (1-25) within all 136 attributes across all 9 datasets. The ranks were obtained using the forward attribute selection. Group abbreviations stand for: syntax-linguistic (S-L), syntax-content (S-C), coherence-spatial (C-S), coherence-networks (C-N), and consistency (C).

attribute	*group*	*average rank*
1. number of words	S-L	16.44
2. number of different words	S-L	23.33
3. score point level for max cos. sim. over all score points	S-C	27.33
4. number of sentences	S-L	32.11
5. number of tokens	S-L	33.78
6. number of spellchecking errors	S-L	36.78
7. Geary's *C*	C-S	38.33
8. local centrality (WTLS_CC)	C-N	38.52
9. Clark's and Evans' distance to nearest neighbour	S-L	38.67
10. pattern cosine	S-C	39.11
11. type-token-ratio	S-L	39.89
12. number of genitive markers	S-L	41.22
13. number of characters	S-L	41.33
14. cosine similarity with source text	S-C	42.00
15. entropy of the link betw. centr. distr. (ENTR_BCL)	C-N	41.27
16. index (minimum distance/maximum distance) (Euclid)	C-S	42.89
17. average word length	S-L	43.22
18. number of verbs - past tense	S-L	43.89
19. average distance between neighbouring points (Euclid)	C-S	44.22
20. number of long sentences	S-L	44.89
21. min distance between points and centroid (cos)	C-S	47.11
22. number of predeterminers	S-L	50.44
23. number of particles	S-L	50.67
24. Getis's *G*	C-S	50.78
25. relative distance	C-S	51.11

Lexical sophistication and grammar attributes.

lexical sophistication	*grammar*
1. number of characters,	29. number of different PoS tags
2. number of words,	30. height of the tree presenting sentence structure,
3. number of long words,	31. correct verb form,
4. number of short words,	32. number of grammar errors,
5. most frequent word length,	*Number of each PoS tag*
6. average word length,	33. coordinating conjunction,
7. number of sentences,	34. numeral,
8. number of long sentences,	35. determiner,
9. number of short sentences,	36. existential there,
10. most frequent sentence length,	37. preposition/subordinating conjunction,
11. average sentence length,	38. adjective,
12. number of different words,	39. comparative adjective,
13. number of stopwords,	40. superlative adjective,
Readability measures [Dubay, 2007; Smith and Jönsson, 2011]	41. ordinal adjective or numeral,
14. Gunning Fox index,	42. modal auxiliary,
15. Flesch reading ease,	43. singular or mass common noun,
16. Flesch Kincaid grade level,	44. plural common noun,
17. Dale-Chall readability formula,	45. singular proper noun,
18. automated readability index,	46. plural proper noun,
19. simple measure of Gobbledygook,	47. preposition,
20. LIX,	48. participle,
21. word variation index,	49. predeterminer,
22. nominal ratio,	50. genitive marker,
Lexical diversity [Mellor, 2011]	51. personal pronoun,
23. type-token-ratio,	52. possessive pronoun,
24. Guiraud's index,	53. adverb,
25. Yule's K,	54. comparative adverb,
26. the D estimate,	55. superlative adverb,
27. hapax legomena - number of words occurring only once in a text,	56. particle, "to" as preposition or infinitive marker,
28. advanced Guiraud,	57. verb - base form,
	58. verb - past tense,
	59. verb - gerund/present participle,
	60. verb - past participle,
	61. verb - 3rd person sing. present,
	62. wh-determiner,
	63. wh-pronoun,
	64. wh-adverb.

(continued) Mechanics and content attributes.

mechanics	*content*
65. number of spellchecking errors,	68. cosine similarity with source text,
66. number of capitalization errors,	69. score point level for maximum cosine similarity over all score points,
67. number of punctuation errors,	70. cosine similarity with essays that have highest score point level,
	71. pattern cosine [Attali, 2011],
	72. weighted sum of all cosine correlation values [Attali, 2011].

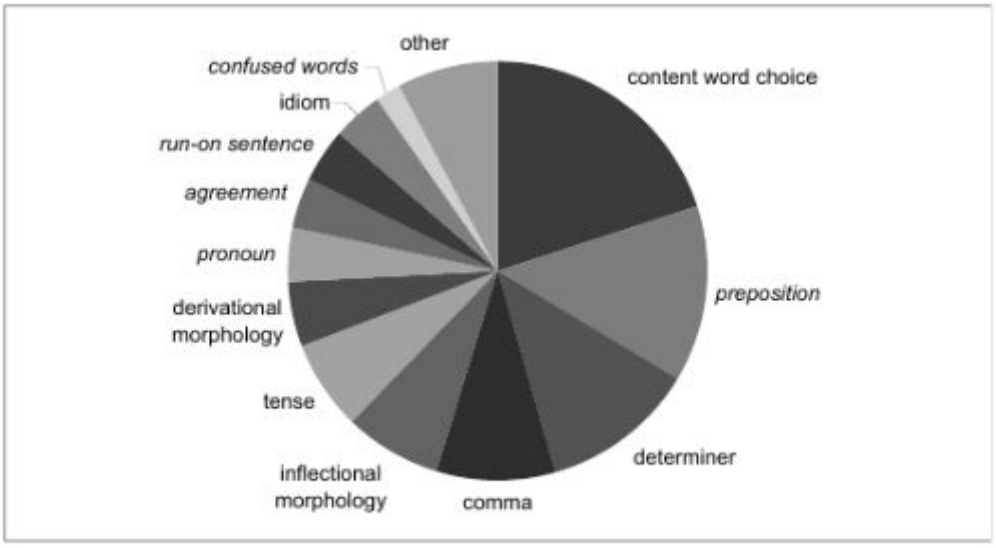


Figure 15.1 Learner errors from *Cambridge Learners Corpus*.

Spelling mistakes amount to almost half of the mistakes in the Cambridge Learner Corpus. How should they be treated?

Syntactic errors versus usage errors? (The Handbook of Automated Essay Evaluation, 251−252)

Grammar-based techniques or statistically-based techniques?

Annotator training?

Text cohesion analysis?

Text quality versus writing skill?

Discourse analysis (criteria: The Handbook of Automated Essay Evaluation, 304−306)?

Check out the sources

1. An overview of the field (journalistic)
2. Semantics-based automated essay evaluation (dissertation, 2018)
3. Pedagogical practices and perceived learning effectiveness in EFL writing classes (see the references!)
4. Feedback from Automated Essay Evaluation Systems: A Review of Selected Research (see the references!)
5. https://www.ets.org/toefl/score-users/scores-admissions/interpret
6. https://www.sketchengine.eu/cambridge-learner-corpus/
7. The Handbook of Automated Essay Evaluation