frontiers

Supplementary Material

Supplementary Table 1. Definitions of important writing features (38 out of 282), selected according to their local and global importance in essays of Table 7.

Feature	Description
1. word_count	Number of words
2. hdd42_aw *Lexical diversity*	The *D* index, derived from the hypergeometric distribution, computes the average type-token ratio (TTR) over multiple samples of 42 tokens randomly drawn from text (McCarthy and Jarvis, 2010).
3. lemma_mattr *Lexical diversity*	Moving average type-token ratio (TTR) with 50-word window (Crossley et al., 2016).
4. Ortho_N *Word neighbor information*	On average, the number of words that can be obtained by changing one letter of the original word while preserving the identity and positions of the other letters (Balota et al., 2007).
5. hdd42_fw *Lexical diversity*	The *D* index, derived from the hypergeometric distribution, computes the average type-token ratio (TTR) of function words over multiple samples of 42 tokens randomly drawn from text (McCarthy and Jarvis, 2010).
6. adjacent_overlap_verb_sent *Lexical overlap (sentence)*	Number of verb lemma types that occur at least once in the next sentence (Crossley et al., 2016).
7. KF_Freq_CW *Word frequency*	Word frequency refers to the number of times a content word occurs in a corpus of texts. Content words that are less frequent in a reference corpus (e.g., edifice, cuisine, egregious) are considered more sophisticated than content words that occur frequently (e.g., building, food, bad) (Kyle et al., 2018).
8. lexical_density_tokens *Lexical density*	Percentage of text tokens that are content words (Crossley et al., 2016).
9. ncomp_stdev *Noun phrase variety*	Dependents per nominal complement (standard deviation) (Kyle et al., 2018).
10. DC_C *Syntactic complexity*	Dependent clauses per clause (Kyle et al., 2018).

11. adjacent_overlap_pronoun_sent *Lexical overlap (sentence)*	Number of pronoun lemma types that occur at least once in the next sentence (Crossley et al., 2016).
12. poly_noun *Semantic network*	Average number of senses for nouns (Kyle et al., 2018).
13. all_logical *Connectives*	Ratio of logical connectives (e.g., actually, admittedly, after all) (Crossley et al., 2016).
14. Sv_GI *Word category: Action*	State verbs: 102 state verbs describing mental or emotional states (Crossley et al., 2017).
15. OG_N_FW *Word neighbor information*	Phonographic neighbors differ in one letter and one phoneme (e.g., shove and stone). Average number of phonographic neighbors among function words (homophones excluded) (Kyle et al., 2018).
16. hyper_verb_noun_s1_p1 *Semantic network*	Average hypernymy score for nouns and verbs (most frequent sense, first path) (Kyle et al., 2018).
17. Econ_GI *Word category: Economics, politics, and religion*	Economic: 502 words (269 in common with Econ@) that is used by the General Inquirer in disambiguating (e.g., abundance, account, account, accrue, acquisition) (Crossley et al., 2017).
18. dobj_per_cl *Clause complexity*	Direct objects per clause (Kyle et al., 2018).
19. grammar *Grammatical accuracy*	Number of grammatical errors (Crossley et al., 2019).
20. poly_adj *Semantic network*	Average number of senses per adjective (Kyle et al., 2018).
21. positive_adjectives_component *Sentiment analysis*	PCA component made of 9 indices from Lu Hui positive adjectives, Vader positive, General Inquirer positive adjectives, Laswell positive affect adjectives (Crossley et al., 2017).
22. det_pobj_deps_struct *Noun phrase complexity*	Determiners per object of the preposition (Kyle, 2016).
23. KF_Freq_AW *Word frequency*	Word frequency refers to the number of times a word occurs in a corpus of texts. Words that are less frequent in a reference corpus (e.g., edifice, cuisine, egregious) are considered more sophisticated than words that occur frequently (e.g., building, food, bad) (Kyle et al., 2018).

24. poss_pobj_deps_struct *Noun phrase complexity*	Possessives per object of the preposition (Kyle, 2016).
25. You_GI *Word category: Reference*	Number of (2nd-person) pronouns indicating another person is being addressed directly divided by the number of words in text (Crossley et al., 2017).
26. CP_T *Syntactic complexity*	Coordinate phrases per T-unit (Kyle, 2016).
27. adjacent_overlap_2_adj_sent *Lexical overlap (sentence)*	Number of adjective lemma types that occur at least once in the next two sentences (Crossley et al., 2016).
28. cl_ndeps_std_dev *Clause variety*	Dependents per clause (standard deviation) (Kyle, 2016).
29. LD_Mean_RT_SD_FW *Word recognition norms*	Standard deviation of mean lexical decision reaction time across all participants for function words (Kyle et al., 2018).
30. acad_construction_attested *Syntactic sophistication*	Percentage of constructions in text that are in reference corpus (COCA Academic) (Kyle et al., 2018).
31. ccomp_per_cl *Clause complexity*	Clausal complements per clause (Kyle, 2016).
32. McD_CD *Contextual distinctiveness*	Co-occurrence probability of word with 500 highly frequent context lemmas (within 5 unigrams to the left and right of the target lemma) (Kyle et al., 2018).
33. SENTENCE_FRAGMENT *Grammatical accuracy*	Number of sentence fragments (Crossley et al., 2019).
34. Submit_GI *Word category: Dominance, respect, money, and power*	Submit: 284 words connote submission to authority or power, dependence on others, vulnerability to others, or withdrawal (Crossley et al., 2017).
35. OG_N_H *Word neighbor information*	Number of phonographic neighbors; i.e., words differing in one letter and one phoneme (e.g., stove and stone); different from orthographic neighbors, which are formed by substituting one letter with another (e.g., stove and shove); includes homophones (Kyle et al., 2018).
36. hyper_noun_S1_P1 *Semantic network*	Average hypernymy score for nouns (most frequent sense, first path) (Kyle et al., 2018).

37. COCA_fiction_tri_2_MI *N-gram strength of association*	Mean Mutual Information (MI) score (item 1 = first bigram, item 2 = remaining word). The MI score represents the joint probability that two items will co-occur (Kyle et al., 2018).
38. WN_Mean_RT *Word recognition norms*	Word naming response time. Mean naming reaction time in milliseconds across all participants for each word (Kyle et al., 2018).