A Linguistic Feature Definitions

For the following definitions, if the denominator of the ratio is zero the result is treated as zero.

Feature Name	Definition
WRich10_S	Richness, 100 topics extracted from Wikipedia Dump
WClar10_S	Clarity, 100 topics extracted from Wikipedia Dump
WNois10_S	Noise, 100 topics extracted from Wikipedia Dump
Wtopc10_c	# of topics, 100 topics extracted from Wikipedia Dump
WRich20_S	Richness, 200 topics extracted from Wikipedia Dump
WClar20_S	Clarity, 200 topics extracted from Wikipedia Dump
WNois20_S	Noise, 200 topics extracted from Wikipedia Dump
WTopc20_c	# of topics, 200 topics extracted from Wikipedia Dump
BRich10_S	Richness, 100 topics extracted from WeeBit Corpus
BClar10_S	Clarity, 100 topics extracted from WeeBit Corpus
BNois10_S	Noise, 100 topics extracted from WeeBit Corpus
BTopc10_c	# of topics, 100 topics extracted from OneStop Corpus
BRich20_S	Richness, 200 topics extracted from WeeBit Corpus
BClar20_S	Clarity, 200 topics extracted from WeeBit Corpus
BNois20_S	Noise, 200 topics extracted from WeeBit Corpus
BTopc20_c	# of topics, 200 topics extracted from OneStop Corpus
AoA Kuperman	Mean age of acquisition of words (Kuperman database)
AoA Kuperman lemmas	Mean age of acquisition of lemmas
AoA Bird lemmas	Mean age of acquisition of lemmas, Bird norm
AoA Bristol lemmas	Mean age of acquisition of lemmas, Bristol norm
AoA Cortese and Khanna lemmas	Mean age of acquisition of lemmas, Cortese and Khanna norm
MRC familiarity	Mean word familiarity rating
MRC concreteness	Mean word concreteness rating
MRC Imageability	Mean word imageability rating
MRC Colorado Meaningfulness	mean word Colorado norms meaningfulness rating
MRC Pavio Meaningfulness	mean word Pavio norms meaningfulness rating
MRC AoA	Mean age of acquisition of words (MRC database)

Table 1: Existing Semantic Features

Feature Name	Definition
at_UEnti_C	average number of unique Entities per token (word)
LoCohPA_S	Local Coherence for PA score from Entity Grid
LoCohPW_S	Local Coherence for PW score from Entity Grid
LoCohPU_S	Local Coherence for PU score from Entity Grid
LoCoDPA_S	Local Coherence dist. for PA score from Entity Grid
LoCoDPW_S	Local Coherence dist. for PW score from Entity Grid
LoCoDPU_S	Local Coherence dist. for PU score from Entity Grid
ra_SSToT_C	ratio of SS transitions: total, count from Entity Grid
ra_SOToT_C	ratio of SO transitions: total, count from Entity Grid
ra_SXToT_C	ratio of SX transitions: total, count from Entity Grid
ra_SNToT_C	ratio of SN transitions: total, count from Entity Grid
ra_OSToT_C	ratio of OS transitions: total, count from Entity Grid
ra_OOToT_C	ratio of OO transitions: total, count from Entity Grid
ra_OXToT_C	ratio of OS transitions: total, count from Entity Grid
ra_ONToT_C	ratio of ON transitions: total, count from Entity Grid
ra_XSToT_C	ratio of XS transitions: total, count from Entity Grid
ra_XOToT_C	ratio of XO transitions: total, count from Entity Grid
ra_XXToT_C	ratio of XX transitions: total, count from Entity Grid
ra_XNToT_C	ratio of XN transitions: total, count from Entity Grid
ra_NSToT_C	ratio of NS transitions: total, count from Entity Grid
to_EntiM_C	total number of Entities Mentions
as_EntiM_C	average number of Entities Mentions per sentence
at_EntiM_C	average number of Entities Mentions per token (word)
to_UEnti_C	total number of unique Entities
as_UEnti_C	average number of unique Entities per sentence

Table 2: Existing Discourse-based features

Feature Name	Definition
to_TreeH_C	total parsed Tree Height of all sentences
as_TreeH_C	average parsed Tree Height per sentence
at_TreeH_C	average parsed Tree Height per token
to_FTree_C	total length of Flattened parsed Trees
as_FTree_C	average length of Flattened parsed Trees per sentence
at_FTree_C	average length of Flattened parsed Trees per token
nouns per word	number of nouns / number of words
proper nouns per word	number of proper nouns / number of words
pronouns per word	number of pronouns / number of words
conjunctions per word	number of conjunctions / number of words
adjectives per word	number of adjectives / number of words
verbs per word	number of verbs / number of words
adverbs per word	number of adverbs / number of words
modal verbs per word	number of modal verbs / number of words
prepositions per word	number of prepositions / number of words
interjections per word	number of interjections / number of words
personal pronouns per word	number of personal pronouns / number of words
wh-pronouns per word	number of wh-pronouns / number of words
lexical words per word	number of lexical words / number of words
function words per word	number of function words / number of words
determiners per word	number of determiners / number of words
VBDs per word	number of past tense verbs / number of words
VBPs per word	number of non-3rd person singular present verbs / number of words
VBZs per word	number of 3rd person singular present verbs / number of words
adjective variation	number of adjectives / numbers of lexical words

Table 3: Existing Syntactic features

Feature Name	Definition
TokSenM_S	total count of tokens * total count of sentence
TokSenS_S	sqrt(total count of tokens * total count of sentence)
TokSenL_S	log(total count of tokens)/log(total count of sent)
as_Token_C	average count of tokens per sentence
as_Sylla_C	average count of syllables per sentence
at_Sylla_C	average count of syllables per token
as_Chara_C	average count of characters per sentence
at_Chara_C	average count of characters per token
number of sentences	number of sentences
mean sentence length	number of words / number of sentences
number of characters	number of characters
number of syllables	number of syllables
SmogInd_S	Smog Index
ColeLia_S	Coleman Liau Readability Score
Gunning_S	Gunning Fog Count Score (New, US Navy Report)
AutoRea_S	Automated Readability Idx (New, US Navy Report)
FleschG_S	Flesch Kincaid Grade Level (New, US Navy Report)
LinseaW_S	Linear Write Formula Score

Table 4: Existing Lexical Features