Table 4: Overview of the 57 traditional and new features investigated (grouped into feature sets)

Feature set	Size	Subtypes	Example/Description
Syntactic complexity	16	Length of production unit Subordination Coordination Particular structures	e.g. mean length of clause e.g. clauses per sentences e.g. coordinate phrases per clause e.g. complex nominals per clause
Lexical richness	12	Lexical density Lexical diversity Lexical sophistication	e.g. ratio contents words / all words e.g. type token ratio e.g. words on General Service List
Register-based n-gram frequency	25	Spoken ($n \in [1, 5]$) Fiction ($n \in [1, 5]$) Magazine ($n \in [1, 5]$) News ($n \in [1, 5]$) Academic ($n \in [1, 5]$)	measures of frequencies of n-grams of order 1-5 from five language registers from the Corpus of Contemporary American English (Davies, 2008)
Information theory	3	$\text{Kolmogorov}_{\text{Deflate}}$ $\text{Kolmogorov}_{\text{Deflate Syntactic}}$ $\text{Kolmogorov}_{\text{Deflate Morphological}}$	measures use Deflate algorithm and relate size of compressed file to size of original

Table 5: Mean scores of top-ten most discriminative language features across CEFR levels

	Feature	A1	A2	B1	B2	C1	C2
1	bigramFiction	38.10	41.78	46.82	50.10	47.65	51.25
2	bigramAcad	28.78	23.93	34.01	40.87	43.56	47.03
3	ANC	0.28	0.25	0.23	0.24	0.24	0.26
4	MLWc	4.28	4.08	4.13	4.37	4.53	4.47
5	MLWs	1.35	1.28	1.31	1.40	1.43	1.40
6	bigramSpoken	42.24	40.90	50.90	56.43	55.82	58.24
7	unigramAcad	110.49	97.77	119.87	134.51	142.79	153.87
8	trigramFiction	4.10	4.72	5.53	5.89	5.13	5.14
9	BNC	0.47	0.47	0.42	0.44	0.44	0.45
10	bigramNews	40.15	38.40	47.30	52.15	53.26	55.48
11	unigramFiction	134.15	123.01	141.03	152.10	161.14	171.13
12	NGSL	0.20	0.18	0.16	0.15	0.15	0.16
13	LD	0.51	0.52	0.51	0.51	0.53	0.53
14	unigramSpoken	137.01	123.94	144.18	156.25	166.00	174.82
15	trigramNews	3.36	3.11	3.96	4.40	4.44	4.50
16	unigramMag	137.09	123.52	143.35	155.42	166.28	175.47
17	unigramNews	135.81	122.51	142.05	153.86	164.79	173.61
18	bigramMag	33.44	31.43	38.54	41.67	43.40	44.12
19	KolDef	0.99	0.96	1.00	1.03	1.06	1.08
20	KolDefMor	0.94	0.94	0.94	0.94	0.94	0.94

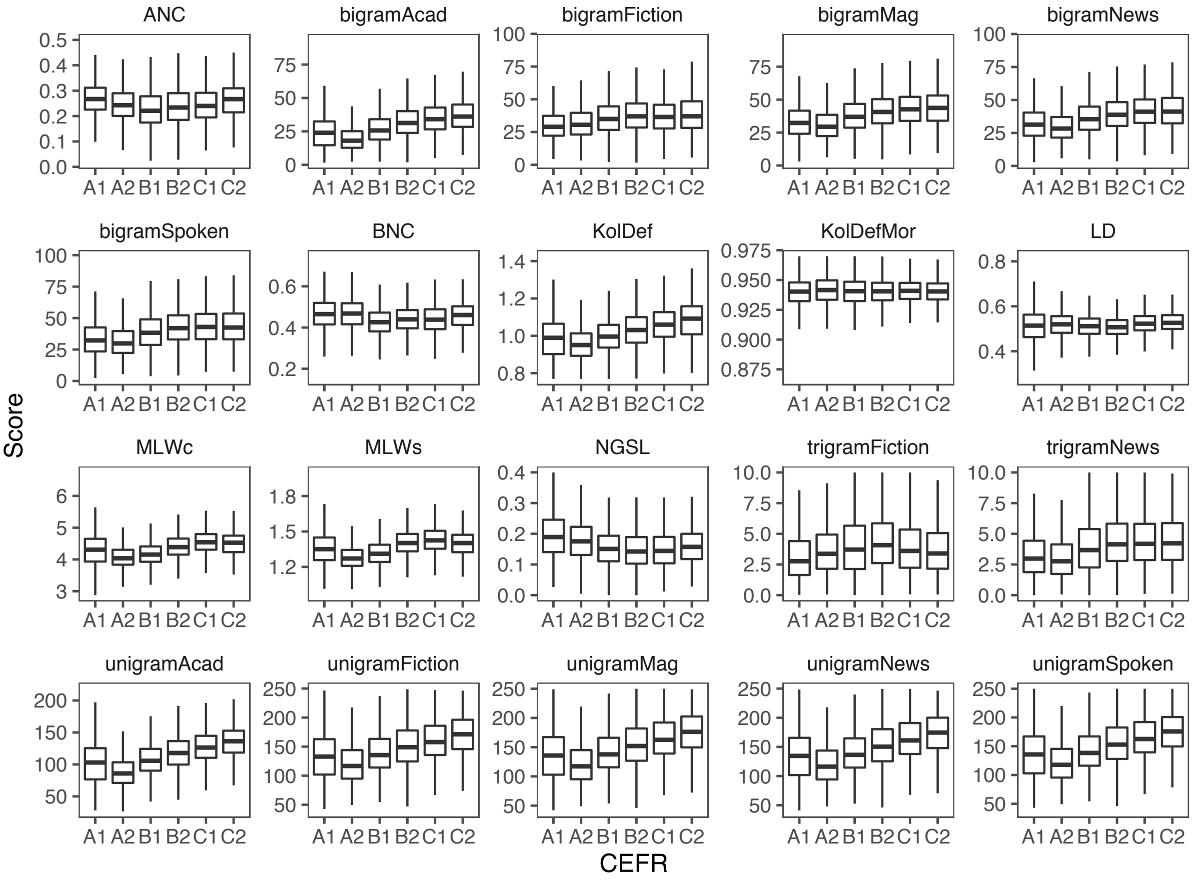

Figure 3: Distribution of scores of top-ten most discriminative language features across CEFR levels.