Category	Feature	Short Description	Resource	Corr. ρ (avg.)	Corr. ρ (SD)
Discourse	ratio_referential	Ratio of referential tokens to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.1278	0.09
Length	doc_num_sents	Number of sentences per document / text	SpaCy, Stanza,	0.1819	0.18
	doc_num_tokens	Number of tokens per document / text	SpaCy, Stanza,	0.5041	0.22
	num_characters	Number of characters per document / text	SpaCy, Stanza, TSEval	0.5224	0.19
	num_characters_per_sentence	Number of characters per sentence	SpaCy, Stanza, TSEval	0.5301	0.17
	num_characters_per_word	Number of characters per word	SpaCy, Stanza, TSEval	0.3895	0.16
	num_sentences	Number of sentences per document / text	SpaCy, Stanza, TSEval	-0.0324	0.11
	num_syllables_in_sentence	Number of syllables in document / text	SpaCy, Stanza, pyphen, TSEval	0.525	0.19
	num_syllables_per_sentence	Number of syllables per sentence	SpaCy, Stanza, pyphen, TSEval	0.4634	0.23
	num_syllables_per_word	Number of syllables per word	SpaCy, Stanza, pyphen, TSEval	0.3924	0.16
	num_words	Number of tokens per document / text	SpaCy, Stanza, TSEval	0.479	0.18
	num_words_per_sentence	Number of tokens per sentence	SpaCy, Stanza, TSEval	0.4863	0.16
Lexical	average_pos_in_freq_table	Average frequency rank of tokens in FastText embeddings	SpaCy, Stanza, TSEval	-0.0907	0.22
	lexical_complexity_score	Lexical complexity based on ranks of FastText embeddings	SpaCy, Stanza, FastText, TSEval	0.0649	0.13
	type_token_ratio	Type-token-Ratio	SpaCy, Stanza, TSEval	-0.3364	0.16
	max_pos_in_freq_table	Maximum frequency rank of tokens in FastText embeddings	SpaCy, Stanza, TSEval	0.2014	0.07
Psycholinguistic	concreteness	Concreteness of words based on MEGAHR and FastText-Embeddings	MEGA.HR crossling	-0.4254	0.24
	imagebility	imagebility of words based on MEGAHR and FastText-Embeddings	MEGA.HR crossling	-0.3962	0.24
Readability	sentence_fkgl	Flesch-Kincaid-Grading-Level, designed for English	SpaCy, Stanza, TSEval	0.4738	0.19
	sentence_fre	Flesch-Reading Ease, designed for English	SpaCy, Stanza, TSEval	-0.3976	0.19
Syntactic	avg_distance_betweeen_verb_particle	Average distance between verb and particle based on dependency tree	SpaCy, Stanza,	0.0385	
	avg_distance_betweeen_words	Average distance between words based on dependency tree	SpaCy, Stanza,	0.3934	0.17
	max_distance_betweeen_verb_particles	Maximum distance between verb and particle based on dependency tree	SpaCy, Stanza,	0.0385	
	max_distance_betweeen_words	Maximum distance between words based on dependency tree	SpaCy, Stanza,	0.2109	0.14
	check_if_head_is_noun	Whether the head of the dependency tree is a noun	SpaCy, Stanza, TSEval	-0.1147	0.13
	check_if_head_is_verb	Whether the head of the dependency tree is a verb	SpaCy, Stanza, TSEval	0.0954	0.13
	check_if_one_child_of_root_is_subject	Whether a child of a root is a subject (not a verb)	SpaCy, Stanza, TSEval	0.0144	0.15
	check_passive_voice	Whether a sentence is in passive voice	SpaCy, Stanza, TSEval		
	average_length_NP	Average length of noun phrase in tokens	SpaCy, Stanza, TSEval	0.3485	0.13
	average_length_VP	Average length of verb phrase in tokens	SpaCy, Stanza, TSEval	0.4324	0.16
	avg_length_PP	Average length of prepositional phrase in tokens	SpaCy, Stanza, TSEval	0.1406	0.07
	parse_tree_height	Depth or height of the dependency tree	SpaCy, Stanza, TSEval	0.4657	0.15
	ratio_clauses	Ratio of tokens associated to a clause to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.1918	0.12
	ratio_of_coordinating_clauses	Ratio of tokens associated to a coordinating clause to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.2503	0.16
	ratio_of_subordinate_clauses	Ratio of tokens associated to a subordinating clause to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.2361	0.14
	ratio_prepositional_phrases	Ratio of tokens associated to a prepositional phrase to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.1797	0.14
	ratio_relative_phrases	Ratio of tokens associated to a relative clause to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.3262	0.15
	is_non_projective	Whether a dependency tree is non projective	SpaCy, Stanza, TSEval	0.0883	0.13
Morphosyntactic	ratio_Abbr_Yes	Ratio of nouns which are an abbreviation to all nouns	SpaCy, Stanza, UniversalDependencies	-0.0685	0.07
	ratio_Case_Abe	Ratio of nouns in abessive case to all nouns	SpaCy, Stanza, UniversalDependencies	0.2394	
	ratio_Case_Acc	Ratio of nouns in accusative case to all nouns	SpaCy, Stanza, UniversalDependencies	0.0658	0.29
	ratio_Case_Ben	Ratio of nouns in benefactive case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Case_Cau	Ratio of nouns in causative case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Case_Cmp	Ratio of nouns in comparative case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Case_Cns	Ratio of nouns in considerative case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Case_Com	Ratio of nouns in comitative case to all nouns	SpaCy, Stanza, UniversalDependencies	0.1293	
	ratio_Case_Dat	Ratio of nouns in dative case to all nouns	SpaCy, Stanza, UniversalDependencies	0.1822	0.09
	ratio_Case_Dis	Ratio of nouns in distributive case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Case_Equ	Ratio of nouns in equative case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Case_Erg	Ratio of nouns in ergative case to all nouns	SpaCy, Stanza, UniversalDependencies		

Table 20: Overview of all 100 features, including correlation coefficient with CEFR level across all languages.

Category	Feature	Short Description	Resource	Corr. ρ (avg.)	Corr. ρ (SD)
Morphosyntactic	ratio_Case_Ess	Ratio of nouns in essive case to all nouns (relevant for ET)	SpaCy, Stanza, UniversalDependencies	0.1196	
	ratio_Case_Gen	Ratio of nouns in genitive case to all nouns	SpaCy, Stanza, UniversalDependencies	0.3528	0.12
	ratio_Case_Ins	Ratio of nouns in instrumental case to all nouns (relevant for CZ)	SpaCy, Stanza, UniversalDependencies	0.2204	0.02
	ratio_Case_Nom	Ratio of nouns in nominative case to all nouns	SpaCy, Stanza, UniversalDependencies	-0.0286	0.13
	ratio_Case_Par	Ratio of nouns in partitive case to all nouns	SpaCy, Stanza, UniversalDependencies	0.2252	
	ratio_Case_Tem	Ratio of nouns in temporal case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Case_Tra	Ratio of nouns in translative case to all nouns (relevant for ET)	SpaCy, Stanza, UniversalDependencies	0.5228	
	ratio_Case_Voc	Ratio of nouns in vocative case to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Definite_Com	Ratio of complex nouns to all nouns (relevant for AR)	SpaCy, Stanza, UniversalDependencies		
	ratio_Definite_Cons	Ratio of nouns in construct state to all nouns (relevant for AR)	SpaCy, Stanza, UniversalDependencies		
	ratio_Definite_Def	Ratio of definite nouns to all nouns	SpaCy, Stanza, UniversalDependencies	0.0569	
	ratio_Definite_Ind	Ratio of indefinite nouns to all nouns	SpaCy, Stanza, UniversalDependencies	-0.059	
	ratio_Foreign_Yes	Ratio of nouns which are foreign to all nouns	SpaCy, Stanza, UniversalDependencies		
	ratio_Mood_Cnd	Ratio of verbs with conditional mood to all verbs	SpaCy, Stanza, UniversalDependencies	0.2172	0.11
	ratio_Mood_Imp	Ratio of verbs with imperative mood to all verbs	SpaCy, Stanza, UniversalDependencies	-0.0243	0.1
	ratio_Mood_Ind	Ratio of verbs with indicative mood to all verbs	SpaCy, Stanza, UniversalDependencies	-0.0782	0.1
	ratio_Mood_Jus	Ratio of verbs with jussive mood to all verbs	SpaCy, Stanza, UniversalDependencies	0.0798	
	ratio_Mood_Qot	Ratio of verbs with quotative mood to all verbs	SpaCy, Stanza, UniversalDependencies	0.1015	
	ratio_Mood_Sub	Ratio of verbs with subjunctive mood to all verbs	SpaCy, Stanza, UniversalDependencies	0.2076	0.13
	ratio_Number_Dual	Ratio of nouns in dual number to all nouns	SpaCy, Stanza, UniversalDependencies	0.0811	
	ratio_Number_Plur	Ratio of nouns in plural number to all nouns	SpaCy, Stanza, UniversalDependencies	0.2426	0.19
	ratio_Number_Sing	Ratio of nouns in singular number to all nouns	SpaCy, Stanza, UniversalDependencies	0.0016	0.19
	ratio_Polarity_Neg	Ratio of negative verbs to all verbs	SpaCy, Stanza, UniversalDependencies	0.2287	0.19
	ratio_Polarity_Pos	Ratio of positive verbs to all verbs	SpaCy, Stanza, UniversalDependencies	0.2199	
	ratio_Tense_Fut	Ratio of verbs in future tense to all verbs	SpaCy, Stanza, UniversalDependencies	0.2023	0.07
	ratio_Tense_Imp	Ratio of verbs in imperfect to all verbs	SpaCy, Stanza, UniversalDependencies	0.2096	0.13
	ratio_Tense_Past	Ratio of verbs in past tense to all verbs	SpaCy, Stanza, UniversalDependencies	0.2782	0.17
	ratio_Tense_Pqp	Ratio of verbs in pluferfect to all verbs	SpaCy, Stanza, UniversalDependencies		
	ratio_Tense_Pres	Ratio of verbs in present tenst to all verbs	SpaCy, Stanza, UniversalDependencies	-0.1113	0.16
	ratio_Voice_Act	Ratio of verbs in active voice to all verbs	SpaCy, Stanza, UniversalDependencies	0.0225	0.18
	ratio_Voice_Mid	Ratio of verbs in middle voice to all verbs	SpaCy, Stanza, UniversalDependencies	0.0682	
	ratio_Voice_Pass	Ratio of verbs in passive voice to all verbs	SpaCy, Stanza, UniversalDependencies	0.3196	0.17
	ratio_mwes	Ratio of multi-word expressions to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.1171	0.13
	ratio_named_entities	Ratio of multi-word expressions to all tokens based on SpaCy pipeline	SpaCy, Stanza, TSEval		
	ratio_of_adjectives	Ratio of adjectives to all tokens	SpaCy, Stanza, TSEval	0.172	0.14
	ratio_of_adpositions	Ratio of adpositions to all tokens	SpaCy, Stanza, TSEval	0.2342	0.13
	ratio_of_adverbs	Ratio of adverbs to all tokens	SpaCy, Stanza, TSEval	0.07	0.19
	ratio_of_auxiliary_verbs	Ratio of auxiliary verbs to all tokens	SpaCy, Stanza, TSEval	0.0034	0.13
	ratio_of_conjunctions	Ratio of conjunctions to all tokens	SpaCy, Stanza, TSEval	0.1429	0.13
	ratio_of_determiners	Ratio of determiners to all tokens	SpaCy, Stanza, TSEval	0.2459	0.16
	ratio_of_function_words	Ratio of function words to all tokens based on dependency tree relations	SpaCy, Stanza, TSEval	0.238	0.16
	ratio_of_interjections	Ratio of interjections to all tokens	SpaCy, Stanza, TSEval	-0.201	0.15
	ratio_of_nouns	Ratio of nouns to all tokens	SpaCy, Stanza, TSEval	-0.0509	0.17
	ratio_of_numerals	Ratio of numerals to all tokens	SpaCy, Stanza, TSEval	-0.005	0.21
	ratio_of_particles	Ratio of particles to all tokens	SpaCy, Stanza, TSEval	0.143	0.16
	ratio_of_pronouns	Ratio of pronouns to all tokens	SpaCy, Stanza, TSEval	-0.0809	0.14
	ratio_of_punctuation	Ratio of punctuation marks to all tokens	SpaCy, Stanza, TSEval	-0.2226	0.16
	ratio_of_symbols	Ratio of symbols to all tokens	SpaCy, Stanza, TSEval	0.0542	0.11
	ratio_of_verbs	Ratio of verbs to all tokens	SpaCy, Stanza, TSEval	-0.0954	0.14
	verb_noun_ratio	How many verbs occur per noun? The higher the value (the more verbs), the easier the text	SpaCy, Stanza, TSEval	0.0501	0.21

Table 21: Overview of all 100 features, including correlation coefficient with CEFR level across all languages. Part II.

Corpus Name	Lang Code (ISO 638-1)	Format	Category	Size	Annotation Method	Expert Annotators	Distinct L1	Inter-Annotator Agreement	CEFR Coverage	License	Resource
`cambridge-exams`	en	document-level	reference	331	n/a	n/a	n/a	n/a	A1-C2	CC BY-NC-SA 4.0	Xia et al. (2016)
`elg-cefr-en`	en	document-level	reference	712	manual	3	n/a	n/a	A1-C2, plus	CC BY-NC-SA 4.0	Breuker (2022)
`cefr-sp`	en	sentence-level	reference	17,000	manual	2	n/a	$r = 0.75, 0.73$	A1-C2	CC BY-NC-SA 4.0	Arase et al. (2022)
`elg-cefr-de`	de	document-level	reference	509	manual	3	n/a	n/a	A1-C2	CC BY-NC-SA 4.0	Breuker (2022)
`elg-cefr-nl`	nl	document-level	reference	3,596	manual	3	n/a	n/a	A1-C2, plus	CC BY-NC-SA 4.0	Breuker (2022)
`icle500`	en	document-level	learner	500	manual	28	ur, pa, bg, zh, cs, nl, fi, fr, de, el, hu, it, ja, ko, lt, mk, no, fa, pl, pt, ru, sr, es, sv, tn, tr	Rasch $\kappa = -0.02$	A1-C2, plus	CC BY-NC 4.0	Thwaites et al. (2024), Granger et al. (2009)
`cefr-asag`	en	paragraph-level	learner	299	manual	3	fr	Krippendorf $\alpha = 0.81$	A1-C2	CC BY-NC-SA 4.0	Tack et al. (2017)
`merlin-cs`	cs	paragraph-level	learner	441	manual	multiple	hu, de, fr, ru, pl, en, sk, es	n/a	A2-B2	CC BY-SA 4.0	Boyd et al. (2014)
`merlin-it`	it	paragraph-level	learner	813	manual	multiple	hu, de, fr, ru, pl, en, sk, es	n/a	A1-B1	CC BY-SA 4.0	Boyd et al. (2014)
`merlin-de`	de	paragraph-level	learner	1,033	manual	multiple	hu, de, fr, ru, pl, en, sk, es	n/a	A1-C1	CC BY-SA 4.0	Boyd et al. (2014)
`hablacultura`	es	paragraph-level	reference	710	manual	multiple	n/a	n/a	A2-C1	CC BY NC 4.0	Vásquez-Rodríguez et al. (2022)
`kwiziq-es`	es	document-level	reference	206	manual	multiple	n/a	n/a	A1-C1	CC BY NC 4.0	Vásquez-Rodríguez et al. (2022)
`kwiziq-fr`	fr	document-level	reference	344	manual	multiple	n/a	n/a	A1-C1	CC BY NC 4.0	Original
`caes`	es	document-level	learner	30,935	computer-assisted	multiple	pt, zh, ar, fr, ru	n/a	A1-C1	CC BY NC 4.0	Vásquez-Rodríguez et al. (2022)
`deplain-web-doc`	de	document-level	reference	394	manual	2	n/a	Cohen $\kappa = 0.85$	A1,A2,B2,C2	CC-BY-SA-3, CC-BY-4, CC-BY-NC-ND-4, save_use_share	Stodden et al. (2023)
`deplain-apa-doc`	de	document-level	reference	483	manual	2	n/a	Cohen $\kappa = 0.85$	A2-B1	CC-BY-SA-3, CC-BY-4, CC-BY-NC-ND-4, save_use_share	Stodden et al. (2023)
`deplain-apa-sent`	de	sentence-level	reference	483	manual	2	n/a	n/a	A2-B2	By request	Stodden et al. (2023)
`elle`	et	paragraph-level, document-level	learner	1,697	manual	2	n/a	n/a	A2-C1	CC BY 4.0	Allkivi et al. (2024), Vajjala and Rama (2018)
`efcamdat-cleaned`	en	sentence-level, paragraph-level	learner	406,062	manual	n/a	br, zh, tw, ru, sa, mx, de, it, fr, jp, tr	n/a	A1-C1	Cambridge	Geertzen et al. (2013), Shatz (2020) Huang et al. (2017)
`beast2019-w&i`	en	sentence-level	learner	3,600	manual	multiple	n/a	n/a	A1-C2	Cambridge	Bryant et al. (2019), Yannakoudakis et al. (2018)
`peapl2`	pt	paragraph-level	learner	481	manual	n/a	zh, en, es, de, ru, fr, ja, it, nl, tet, ar, pl, ko, ro, sv	n/a	A1-C2	CC BY SA NC 4.0	Martins et al. (2019)
`cople2`	pt	paragraph-level	learner	942	manual	n/a	zh, en, es, de, ru, fr, ja, it, nl, tet, ar, pl, ko, ro, sv	n/a	A1-C1	CC BY SA NC 4.0	Mendes et al. (2016)
`zaebuc`	ar	paragraph-level	learner	214	manual	3	en	Unnamed $\kappa = 0.99$	A2-C1	CC BY SA NC 4.0	Habash and Palfreyman (2022)
`readme`	ar, en, fr, hi, ru	sentence-level	reference	9,757	computer-assisted	2	n/a	Krippendorf $\kappa = 0.67, 0.78$	A1-C2	CC BY SA NC 4.0	Naous et al. (2024)
`apa-lha`	de	document-level	reference	3,130	n/a	n/a	n/a	n/a	A2-B1	Public	Spring et al. (2021)
`learn-welsh`	cy	document-level, sentence-level, discourse-level	reference	1,372	manual	n/a	n/a	n/a	A1-A2	Public	Original

Table 22: The UNIVERSALCEFR-FULL directory of dataset information reporting full details of properties of corpora included in the main collection.