	stage = 2:raw text	stage = 3:deduped&cleaned	
null knc_Arab taq_Tfng	* CD • • • • • • • • • • • • • • • • • •		
dyu_Latn sat_Olck tzm_Tfng kon_Latn mni_Beng		***	
kam_Latn nus_Latn cjk_Latn umb_Latn		* * * * * * * * * * * * * * * * * * *	
fon_Latn bjn_Arab kmb_Latn kas_Arab			
kas_Deva kea_Latn knc_Latn tum_Latn	##		
ace_Arab lua_Latn mai_Deva dzo_Tibt	+3 • • • • • • • • • • • • • • • • • • •	#	
kik_Latn dik_Latn tso_Latn bho_Deva kbp_Latn		3+	
fij_Latn mos_Latn asm_Beng ssw_Latn		x + (0	
ewe_Latn bod_Tibt nso_Latn lao_Laoo	# • • • • • • • • • • • • • • • • • • •	→	
azb_Arab shn_Mymr war_Latn smo_Latn			
ory_Orya nya_Latn quy_Latn san_Deva	# (		
bam_Latn kac_Latn lug_Latn fuv_Latn plt_Latn	+ × · · · · · · · · · · · · · · · · · ·	3+ (0	
taq_Latn sot_Latn tsn_Latn luo_Latn		X +3 +0 = = = = = = = = = = = = = = = = = =	
kmr_Latn bak_Cyrl mri_Latn pbt_Arab	# • • • • • • • • • • • • • • • • • • •		
tpi_Latn snd_Arab sag_Latn ydd_Hebr	+		
fao_Latn bem_Latn ckb_Arab pan_Guru hne_Deva	+30		
tuk_Latn kir_Cyrl kin_Latn uig_Arab	+39 +39 +30 (#: (= 0	+15 = 1	
twi_Latn lij_Latn min_Latn khm_Khmr			
awa_Deva tat_Cyrl lus_Latn lin_Latn			
run_Latn jav_Latn crh_Latn guj_Gujr ban_Latn	X40     0     0     0       X40     0	3+ (=) = =	
pag_Latn ibo_Latn yue_Hant epo_Latn	+ 1130 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	** III	
bug_Latn Itg_Latn ceb_Latn ace_Latn		* (* 0	group • cc_m
wol_Latn kan_Knda ayr_Latn szl_Latn bjn_Latn	+3 (		<ul><li>cc_n</li><li>cc_o</li><li>cc_r</li><li>ia_archivebot</li></ul>
tgk_Cyrl sna_Latn gle_Latn gla_Latn	30	\$\tag{\tag{\tag{\tag{\tag{\tag{\tag{	<ul><li>ia_m</li><li>ia_n</li><li>ia_o</li><li>ia_survey</li></ul>
oci_Latn tel_Telu gaz_Latn npi_Deva		* • • • • • • • • • • • • • • • • • • •	style cc ia_archivebot la ia_survey
mya_Mymr yor_Latn mar_Deva hau_Latn tir_Ethi	+0 × = = 0 = = = = = = = = = = = = = = =	+3	
cym_Latn scn_Latn khk_Cyrl mal_Mlym			
swh_Latn ltz_Latn hye_Armn lmo_Latn		+ 6x	
urd_Arab kaz_Cyrl isl_Latn mkd_Cyrl ilo_Latn	+: • • • • • • • • • • • • • • • • • • •	+ • • • • • • • • • • • • • • • • • • •	
kat_Geor mag_Deva kab_Latn fur_Latn	3+9 • • • • • • • • • • • • • • • • • • •		
vec_Latn bel_Cyrl zul_Latn srp_Cyrl prs_Arab		+; • • • • • • • • • • • • • • • • • • •	
glg_Latn tgl_Latn pap_Latn als_Latn	# • • • • • • • • • • • • • • • • • • •	3+ •) = = = = = = = = = = = = = = = = = =	
lim_Latn afr_Latn tam_Taml sun_Latn		30 = = = = = = = = = = = = = = = = = = =	
amh_Ethi hat_Latn grn_Latn eus_Latn		34D	
azj_Latn ast_Latn xho_Latn sin_Sinh uzn_Latn			
srd_Latn hin_Deva nno_Latn lvs_Latn	143		
mlt_Latn slv_Latn bos_Latn hrv_Latn		+3 × · · · · · · · · · · · · · · · · · ·	
heb_Hebr cat_Latn ben_Beng lit_Latn tha_Thai	+30		
est_Latn bul_Cyrl kor_Hang nob_Latn	+36		
slk_Latn som_Latn ukr_Cyrl fin_Latn	+30	+++(*) = = (**) ++++++++++++++++++++++++++++++++++	
dan_Latn hun_Latn ron_Latn ces_Latn	+x3		
zsm_Latn ell_Grek ara_Arab vie_Latn zho_Hant		3+3 +	
swe_Latn tur_Latn ind_Latn pes_Arab			
pol_Latn ita_Latn por_Latn jpn_Jpan		**3	
nld_Latn fra_Latn spa_Latn deu_Latn zho_Hans		3	
rus_Cyrl eng_Latn			
	0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 chars	0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 chars	
1	chars	chars	