

Dataset	# sentences	# tokens	
		English	Catalan
Training parallel (in-domain)	1,086,623	16,876,830	18,302,284
Training parallel (OpenSubs)	402,775	3,577,109	3,381,241
Training monolingual (in-domain)	5,306,055	-	100,426,922
Training monolingual (in-domain)	13,841,542	210,337,379	-
Training monolingual (web)	16,516,799	-	486,961,317
Development	2,000	34,562	38,114