

Type	Dataset	# Sentences
High-resource: German↔English		
Training	Common Crawl + Europarl v7 + News Comm. v12	4,356,324
Dev	Newstest 2015+2016	5,168
Test	Newstest 2017	3,004
Mono-DE	News Crawl 2016	26,982,051
Mono-EN	News Crawl 2016	18,238,848
Low-resource: Tagalog↔English		
Training	News/Blog	50,705
Dev/Test	News/Blog	491/508
Dev/Test*	Bible	500/500
Sample*	Bible	61,195
Mono-TL	Common Crawl	26,788,048
Mono-EN	ICWSM 2009 blog	48,219,743
Low-resource: Swahili↔English		
Training	News/Blog	23,900
Dev/Test	News/Blog	491/509
Dev/Test*	Bible-NT	500/500
Sample*	Bible-NT	14,699
Mono-SW	Common Crawl	12,158,524
Mono-EN	ICWSM 2009 blog	48,219,743