

corpus	lines	en words	fr words
train	12.1M	304M	348M
mono	15.9M	—	406M
dev	6003	138k	155k
test	3003	71k	81k