Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	1,965,734		2,007,723		1,920,209		646,605	
Words	56,895,229	54,420,026	60,125,563	55,642,101	50,486,398	53,008,851	14,946,399	17,376,433
Distinct words	176,258	117,481	140,915	118,404	381,583	115,966	172,461	63,039

News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	157,302		137,097		158,840		136,151	
Words	4,449,786	3,903,339	3,915,218	3,403,043	3,950,394	3,856,795	2,938,308	3,264,812
Distinct words	78,383	57,711	63,805	53,978	130,026	57,464	136,392	52,488

United Nations Training Corpus

	Spanish ↔ English		French ↔ English	
Sentences	11,196,913		12,886,831	
Words	318,788,686	365,127,098	411,916,781	360,341,450
Distinct words	593,567	581,339	565,553	666,077

10^9 Word Parallel Corpus

	French ↔ English	
Sentences	22,520,400	
Words	811,203,407	668,412,817
Distinct words	2,738,882	2,861,836

CzEng Training Corpus

	Czech ↔ English	
Sentences	14,833,358	
Words	200,658,857	228,040,794
Distinct words	1,389,803	920,824

Europarl Language Model Data

	English	Spanish	French	German	Czech
Sentence	2,218,201	2,123,835	2,190,579	2,176,537	668,595
Words	59,848,044	60,476,282	63,439,791	53,534,167	14,946,399
Distinct words	123,059	181,837	145,496	394,781	172,461

News Language Model Data

	English	Spanish	French	German	Czech
Sentence	51,827,706	8,627,438	16,708,622	30,663,107	18,931,106
Words	1,249,883,955	247,722,726	410,581,568	576,833,910	315,167,472
Distinct words	2,265,254	926,999	1,267,582	3,336,078	2,304,933

News Test Set

	English	Spanish	French	German	Czech
Sentences	3003				
Words	73,785	78,965	81,478	73,433	65,501
Distinct words	9,881	12,137	11,441	14,252	17,149