Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	1,786,594		1,825,077		1,739,154		462,351	
Words	51,551,370	49,411,045	54,568,499	50,551,047	45,607,269	47,978,832	10,573,983	12,296,772
Distinct words	171,174	113,655	137,034	114,487	362,563	111,934	152,788	56,095

News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	132,571		115,562		136,227		122,754	
Words	3,739,293	3,285,305	3,290,280	2,866,929	3,401,766	3,309,619	2,658,688	2,951,357
Distinct words	73,906	53,699	59,911	50,323	120,397	53,921	130,685	50,457

United Nations Training Corpus

	Spanish ↔ English		French ↔ English	
Sentences	10,662,993		12,317,600	
Words	348,587,865	304,724,768	393,499,429	344,026,111
Distinct words	578,599	564,489	621,721	729,233

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	7,227,409	
Words	72,993,427	84,856,749
Distinct words	1,088,642	522,770

Europarl Language Model Data

	English	Spanish	French	German	Czech
Sentence	2,032,006	1,942,761	2,002,266	1,985,560	479,636
Words	54,720,731	55,105,358	57,860,307	48,648,697	10,770,230
Distinct words	119,315	176,896	141,742	376,128	154,129

News Language Model Data

	English	Spanish	French	German	Czech
Sentence	30,888,595	3,416,184	11,767,048	17,474,133	12,333,268
Words	777,425,517	107,088,554	302,161,808	289,171,939	216,692,489
Distinct words	2,020,549	595,681	1,250,259	3,091,700	2,068,056

News Test Set

	English	Spanish	French	German	Czech
Sentences	3003				
Words	75,762	79,710	85,999	73,729	65,427
Distinct words	10,088	11,989	11,584	14,345	16,922