Europarl Training corpus

	Spanish ↔ English	**French ↔ English**	**German ↔ English**
Sentences	1,259,914	1,288,901	1,264,825
Foreign words	33,159,337	33,176,243	29,582,157
English words	31,813,692	32,615,285	31,929,435
Distinct foreign words	345,944	344,287	510,544
Distinct English words	266,976	268,718	250,295

News Commentary Training corpus

	Spanish ↔ English	**French ↔ English**	**German ↔ English**	**Czech ↔ English**
Sentences	51,613	43,194	59,975	57797
Foreign words	1,263,067	1,028,672	1,297,673	1,083,122
English words	1,076,273	906,593	1,238,274	1,188,006
Distinct foreign words	84,303	68,214	115,589	142,146
Distinct English words	70,755	63,568	76,419	74,042

Language model data

	English	**Spanish**	**French**	**German**
Sentence	1,407,285	1,431,614	1,435,027	1,478,428
Words	34,539,822	36,426,542	35,595,199	32,356,475
Distinct words	280,546	385,796	361,205	558,377

Europarl test set

	English	**Spanish**	**French**	**German**
Sentences	2,000			
Words	53,531	55,380	53,981	49,259
Distinct words	8,558	10,451	10,186	11,106

News Commentary test set

	English	**Spanish**	**French**	**German**	**Czech**
Sentences	2,007				
Words	43,767	50,771	49,820	45,075	39,002
Distinct words	10,002	10,948	11,244	12,322	15,245