CS674, Spring 2002
Lecture Handout, 4/22/02

Examples drawn from Knight and Marcu (2000).

1. Good, average, and bad cases (figure 5 from Knight and Marcu, "decision" data removed). The baseline "produces compressions with highest word-bigram scores" (on average? as a (smoothed) bigram language model?)

Original:	Beyond the basic level, the operations of the three products vary widely.
Baseline:	Beyond the basic level, the operations of the three products vary widely.
Statistical:	The operations of the three products vary widely.
Human:	The operations of the three products vary widely.
Original:	Arborscan is reliable and worked accurately in testing, but it produces very large dxf files.
Baseline:	Arborscan and worked in, but it very large dxf.
Statistical:	Arborscan is reliable and worked accurately in testing, but it produces very large dxf files.
Human:	Arborscan produces very large dxf files.
Original:	Many debugging features, including user-defined break points and variable-watching and message-watching windows, have been added.
Baseline:	Debugging, user-defined and variable-watching and message-watching, have been.
Statistical:	Many debugging features, including user-defined points and variable-watching and message-watching windows, have been added.
Human:	Many debugging features have been added.

2. Experimental results for corpus Test (Table 1 of Knight and Marcu, Cmplg corpus and "decision" data removed). The average original sentence length was 21 words. Mean and standard deviations are given across four judges.

	Baseline	Statistical	Human
Compression	63.70%	70.37%	53.33%
Grammaticality	1.78 ± 1.19	4.34 ± 1.02	4.92 ± 0.18
Importance	2.17 ± 0.89	3.38 ± 0.67	4.24 ± 0.52