Table 9: Alignment Scores by SWA

NYT-RSS	CLwritten	CLspoken
-3.061 (2000)	-1.882 (1629)	0.450 (4110)

10 Why T3 fails

It is interesting and worthwhile to ask what caused T3, heavily clad in ideas from the recent machine learning literature, to fail on NYT-RSS, as opposed to the 'CLwritten' and 'CLspoken' corpora on which T3 reportedly prevailed compared to other approaches (Cohn and Lapata, 2009). The CLwritten corpus comes from written sources in the British National Corpus and the American News Text corpus; the CLspoken corpus comes from transcribed broadcast news stories (cf. Table 7).

We argue that there are some important differences between the NYT-RSS corpus and the CLwritten/CLspoken corpora that may have led to T3's poor record with the former corpus.

The CLwritten and CLspoken corpora were created with a specific purpose in mind: namely to assess the compression-by-deletion approach. So their authors had a very good reason to limit gold standard compressions to those that can be arrived at only through deletion; annotators were carefully instructed to create compression by deleting words from the source sentence in a way that preserves the gist of the original sentence. By contrast, NYT-RSS consists of naturally occurring compressions sampled from live feeds on the Internet, where relations between compression and its source sentence are often not as straightforward. For instance, to arrive at a compression in NYT-RSS in Table 7 involves replacing *race* with *her plans to unseat senator Hillary Rodam Clinton*, which is obviously beyond what is possible with the deletion based approach.

In CLwritten and CLspoken, on the other hand, compressions are constructed out of parts that appear *in verbatim* in the original sentence, as Table 7 shows: thus one may get to the compressions by simply crossing off words from the original sentence.

To see whether there is any significant difference among NYT-RSS, CLwritten and CLspoken, we examined how well gold standard compressions are aligned with source sentences on each of the corpora, using SWA. Table 9 shows what we found. Parenthetical numbers represent how many pairs of compression and source are found in each corpus. A larger score means a tighter alignment between gold standard compression and its source sentence: we find in Table 9 that CLspoken has a source sentence more closely aligned with its compression than CLwritten, whose alignments are more closely tied than NYT-RSS's.

Figure 4 (found in the previous page) shows how SWA alignment scores are distributed over each of the corpora. CLwritten and CLspoken have peaks at around 0, with an almost entire mass of scores concentrating in an area close to or above 0. This means that for the most of the cases in either CLwritten or CLspoken, compression is very similar in form to its source. In contrast, NYT-RSS has a heavy concentration of scores in a stretch between -5 and -10, indicating that for the most of time, the overlap between compression and its source is rather modest compared to CLwritten and CLspoken.

So why does T3 fails on NYT-RSS? Because NYT-RSS contains lots of alignments that are only weakly related: in order for T3 to perform well, the training corpus should be made as free of spurious data as possible, so that most of the alignments are rated over and around 0 by SWA. Our concern is that such data may not happen naturally, as the density distribution of NYT-RSS shows, where the majority of alignments are found far below 0, which could raise some questions about the robustness of T3.

11 Conclusions

This paper introduced the model free approach, GST/g, which works by creating compressions only in reference to dependency structure, and looked at how it compares with a model intensive approach T3 on the data gathered from the Internet. It was found that the latter approach appears to crucially rely on the way the corpus is constructed in order for it to work, which may mean a huge compromise.

Interestingly enough, GST/g came out a winner on the particular corpus we used, even outperform-