Predictions of the Dependency Locality Theory (DLT) [3] have held in experiments using constructed stimuli [9], but effects have been weak or negative when applied broadly in naturalistic studies [e.g. [1]]. We use hand-corrected syntactic annotations of the Dundee eye-tracking corpus [4] to evaluate the possibility that this is due to errors in automatic dependency estimation.

Baseline model:
Sentence position, word length, length of preceding saccade, cumulative 5-gram probability, and total surprisal, with by-subject random slopes for each of these and random intercepts by word.

Exploratory data:
Every third sentence of Dundee.

Held-out data:
All Dundee sentences not in exploratory data.

Experiment 1 - DLT on Gold Dundee
We fit first-pass durations on held-out data using a baseline [7] (see above) and log-transformed DLT integration cost. Non-significant results (Table 2) show the negative effect found initially by [1] may have been due to automatic parser errors. Note that the predicted positive correlation with reading times is not observed, either.

Experiment 2: Broad-coverage variants of DLT
We then tested three broad-coverage modifications to DLT (right). BothMod most improved model fit on exploratory data (Table 1), so it was evaluated on the remainder of the corpus. Contrary to DLT predictions, the effect is significantly negative (Table 2).

DLT Variants tested
We used the following variants of DLT (modifications designed to better account for broad coverage phenomena):

1. The person that supervisors and co-workers caught stealing millions

2. The person that supervisors and co-workers caught stealing millions

3. Both VerbMod and CoordMod are applied together.

4. The person that supervisors and co-workers caught stealing millions

Follow-Up Study 1: Sentence Intercept
To control for the possibility of sentence-level effects, we added a random intercept by subject:sentid pair to the baseline. The negative effect persists on exploratory data (see Table 3).

Follow-Up Study 2: Amateur Novels Corpus
We ran follow-up study 1 on every 3rd sentence of the Amateur Novels Corpus [2] (fewer words but more subjects), again using hand-corrected syntactic annotations. Results are not significant for either DLT (original) or BothMod (see Table 4).

References