tions for the data, such as sentence length or individual document difficulty.

	manual		parser		
	attach	**label**	**attach**	**label**	**tok**
interv.	88.1	89.2	80.2	83.2	1405
news	89.9	90.5	80.9	82.5	1222
whow	87.0	87.5	80.7	82.1	1371
voyage	88.4	90.4	82.0	**87.1+**	1058
decl	93.6	94.8	87.0	90.3	3588
frag	**89.3*****	**89.0*****	**76.0*****	**72.1*****	337
sub	85.7	89.3	82.1	89.3	28
q	**100+**	100	86.3	87.7	73
imp	93.6	95.3	86.4	88.4	361
other	**87.3*****	**88.0*****	**70.6*****	**76.6*****	299
inf	100	93.1	96.6	89.7	29
wh	**88.0***	90.4	80.7	84.3	83

Table 6: Parser and corrector accuracies.

The four mixed-effects models summarized in Table 7 show that while sentence type survives, genre is no longer significant. Moreover, sentence length was disruptive only for humans (in contrast to Ravi et al.'s data, though that study did not include sentence type as a predictor).

	manual		automatic	
	label	**attach**	**label**	**attach**
length	-1.62	**-3.02****	1.70	-1.42
news	1.08	-0.13	-0.36	-0.34
voyage	0.93	-0.43	1.31	0.03
whow	-0.16	-0.76	0.25	-0.06
frag	**-4.48*****	**-5.15*****	**-7.09*****	**-5.34*****
imp	0.23	-0.17	-0.15	-0.24
inf	-0.19	0.90	0.27	1.03
other	**-3.85****	**-2.31***	**-5.71*****	**-4.84*****
q	1.29	0.28	-0.55	-1.59
sub	-1.01	-1.63	0.14	-0.69
wh	-1.29	**-2.23***	-1.06	**-2.07***

Table 7: t values from mixed effects models for parsing accuracy using sentence type, genre and length, with document random effects.

The most striking sentence type predictor is *wh*, though it is based on little data. As length has been factored in, these are cases where length is not a sufficient predictor of the observed error rate. Upon closer inspection, *wh* sentences are shorter overall – about 10 tokens on average – while declaratives are 21 tokens on average but similarly difficult. Both types are dense in the syntactic content that can lead to errors while easy to catch categories, such as trivial modifiers, are more rare - see the dearth of easy modifier functions despite complex syntax in examples (3–5).

(3) *What analysis did you perform on the specimens and what equipment was used?*

(4) *What are the startup costs involved?*

(5) *Why run for president?*

The type *frag* was a strong predictor of error. Many instances of *frag* in the data were more complex than a simple NP, such as captions for image credit (6), dates (7), NPs with foreign word heads (8) or potentially ambiguous NPs (9), among many other short bits of language with little else available to contextualize them.

(6) *Image: Mathias Krumbholz.*

(7) *Tuesday, September 1, 2015*

(8) *Beauveria bassiana on a cicada in Bolivia.*

(9) *Clothing supply closet*

Imperatives were not a strong predictor of error; this is surprising given Silvera et al. (2014)'s characterization of imperatives being an essential difference between newswire and non-newswire text. While lacking an overt subject, imperatives were largely syntactically conventional. Omitting the subject relation did not create difficulties for the parser or annotators.

6 Coreference resolution

6.1 Method

Domain adaptation in coreference resolution has been discussed often, both in the context of multiple text types within standard reference corpora (e.g. conversation, newswire and Web subcorpora in datasets such as the ACE corpus, see Yang et al. 2012) or novel domains that are not included in most reference corpora, such as Biomedical NLP (Apostolova et al. 2012, Zhao & Ng 2014). Such studies suggest a genre or text type effect for coreference; sentence type effects, by contrast, have not yet been studied.

Pradhan et al. (2014) give a detailed overview and reference implementation of evaluation metrics for coreference resolution, including the MUC, B^3 and CEAF scores, which are averaged to produce the standard CoNLL score. The metrics focus on correct links between postulated entities, correct mention recognition, and correct entity recognition across mentions (see Pradhan et al. for details and references). Using the metrics on subcorpora of genres is unproblematic: scores can be reported for each subcorpus. However for sentence types, we encounter problems: the metrics were designed for the evaluation of entire running documents and cannot be applied directly to parts of documents, since we will not be running systems or manually annotating only