figurable rule-based coreferencer called xrenner (Zeldes & Zhang 2016).[12] The tool can be set up to produce GUM's annotation scheme. The same data subset as for POS tagging was doubly corrected, and is used below.

6.2 Results

Table 8 gives p-link precision and recall for manual (double corrected) and automatic coreference resolution in the genre vs. sentence type partitions. The results show that differences between genres are comparatively small: although humans fare best on news and travel guides and worst on interviews, their performance is rather comparable, with a range of only .06 F1 points.

	manual			automatic		
	R	**P**	**F1**	**R**	**P**	**F1**
interview	0.67	0.86	0.75	0.59	0.60	0.60
news	0.74	0.90	0.81	0.53	0.56	0.54
voyage	0.77	0.83	0.80	0.51	0.49	0.50
whow	0.71	0.86	0.77	0.60	0.58	0.59
decl	0.72	0.86	0.78	0.56	0.57	0.56
frag	0.75	0.88	0.81	0.45	0.37	0.40
ger	0.68	0.86	0.76	0.59	0.59	0.59
imp	0.66	0.87	0.75	0.61	0.59	0.60
inf	0.65	0.80	0.72	0.46	0.63	0.53
other	0.79	0.91	0.84	0.54	0.58	0.56
q	0.67	0.86	0.76	0.62	0.65	0.63
sub	0.69	0.88	0.77	0.61	0.56	0.58
wh	0.71	0.91	0.80	0.66	0.75	0.70

Table 8: Partitioned precision and recall p-link scores.

Recall is universally lower than precision, suggesting that many cases of lexical coreference ('different names for the same thing') are left out by annotators with only minimal training (as we will see below, pronouns were overwhelmingly resolved correctly). The automatic coreferencer, by contrast, has the easiest time with interviews and how-to guides, due to two simple facts: the long chains of 'I' and 'you' boost scores in interviews, and the how-to guides tend to refer to the main subject of the guide repeatedly by name, making a lexical matching strategy work well. The range of F1 scores is within .1 points, larger but still modest.

Sentence types, by contrast, show much greater variance, with F1 scores ranging 0.72-0.84 for manual annotation and 0.40-0.70 for the coreferencer. Figure 2 plots the ranges of values.

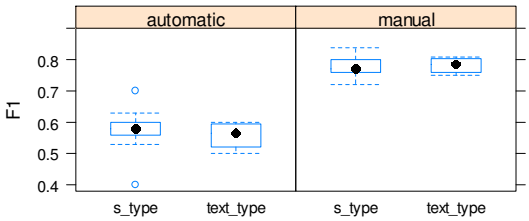

Figure 2: Box plots for p-link F-scores by partition using manual and automatic annotations.

It is clear that sentence types are more spread out, but for automatic annotation this is also due to two outliers: wh-questions as in (10), which do well, possibly due to a simpler information structure and fewer 'confusing' adjuncts, and fragments, which do badly for the coreferencer, possibly because of coreference via synonyms (see e.g. 12 below).

(10) *then circumstances allowed* [*her*] *to attend the exhibit. Why did* [*she*] *so badly want to attend?*

It is however possible that sentence types are more spread out because they form more categories, and some of the smaller ones may distort the skew of F1 scores. We would therefore like to know whether a model given both types of partitions would find either or both significant in predicting errors. Again we control for length (*imp* and *frag* are also short), but also for pronominality, since some sentence types may include more pronouns, for which recall is higher for both human and machine. Table 9 gives t values and significance for 4 mixed effects models predicting precision and recall errors, allowing for different error-rate intercepts for each document.

	manual		automatic	
	recall	**precision**	**recall**	**precision**
length	**-2.16***	-0.28	**-6.55*****	**-4.53*****
news	1.61	1.73	1.11	0.58
voyage	-1.29	1.90	-0.82	-1.08
whow	-0.79	1.50	0.17	0.11
frag	**2.02+**	1.46	**-5.69*****	**-3.95*****
ger	0.53	-1.45	-0.56	0.28
imp	**2.25+**	-1.13	**2.27+**	**3.00++**
inf	-0.98	-0.54	-0.37	-0.39
other	1.42	**2.88++**	-0.82	-0.51
q	-1.45	-0.38	-0.12	-1.57
sub	-0.82	-1.39	-0.15	-0.58
wh	1.72	1.52	**3.71++**	**2.69++**
pron	**11.96+++**	**14.56+++**	**17.71+++**	**21.38+++**

Table 9: t-values for mixed effects models of precision and recall for manual and automatic annotation.

All models predict highly significant positive scores for pronominality (i.e. pronouns are easi-

[12] The tool is open source and freely available at: `http://corpling.uis.georgetown.edu/xrenner`.