as well as length. Each column gives values for a different tagger or manual annotation.[7]

	Manual	TT	Stanford	Spacy
length	-0.38	-1.10	1.46	0.78
news	0.26	0.81	1.29	0.87
voyage	-0.05	0.00	-0.58	-0.52
whow	0.50	-1.30	-1.74	-1.84
frag	0.86	**-6.60*****	**-8.23*****	**-8.41*****
ger	-1.28	-1.60	**-4.22*****	**-4.52*****
imp	**-2.26***	**-2.60****	**-5.01*****	**-2.65****
inf	-0.61	**-8.00*****	**-7.90*****	**-5.24*****
other	0.55	**-5.80*****	**-4.90*****	**-5.17*****
q	1.29	**-2.10***	**-2.08***	-0.08
sub	-0.06	0.31	1.13	0.94
wh	0.55	-3.9	**-2.28***	-0.30

Table 5: t values for mixed effects models with document, genre, sentence and length effects (significant values bold).

The effects disappear almost entirely for manual annotation, suggesting document or annotator specific factors. The significant result for *imp* is related to the positive coefficient of *whow*, which is collinear with the presence of *imp* (r^2=-0.285).[8]

Results for the taggers remain highly significant and entirely restricted to sentence types: the model consistently chooses sentence type over genre, despite the presence of the length predictor, which is somewhat correlated with imperatives (0.16) and fragments (0.20). The overall picture emerging from these results is that sentence type is more influential than genre, and that effects in manual annotation are modest. For taggers, *decl* is much better than any other type.

5 Dependency parsing

5.1 Method

Of the three tasks examined in this paper, we expect the most marked input effects for syntactic parsing. Parsing is not only well known to be affected by genre and domain (Lease & Charniak 2005, Khan et al. 2013), as well as sentence length (Ravi et al. 2008), but it is also directly related to sentence type, since the unit of annotation is the sentence, and local problems in a parse can disrupt accuracy throughout each clause.

Unlike POS tagging, dependency annotations in GUM represent manually corrected output from the Stanford Parser (see Chen & Manning 2014; V3.5 was used). While the entire corpus was corrected by student annotators, only 4,872 tokens were corrected a second time by an experienced instructor. Although this is a small dataset, we choose to use it rather than the whole corpus both because it is more reliable, and because this allows us to evaluate human errors in the initial correction. Our results for manual annotation therefore apply to the task of parser correction, and not to annotation from scratch.

Here too, we consider text and sentence type, but also sentence length, as well as individual document effects. Our null hypothesis is an equal distribution of errors among all partitions. We suspect a stronger effect for sentence length, since long distance dependencies are likelier in long sentences and may be more difficult for humans and automatic parsing, by opening up more opportunities for actual and apparent ambiguities. Sentence type may also have a strong effect, especially for types underrepresented in parser training data (i.e. the Penn Treebank, Marcus et al. 1993). This is expected for imperatives and non-canonical clauses, whereas the *decl* and *sub* types are expected to perform best.

5.2 Results

Table 6 gives accuracy by genre and sentence type for dependency label and attachment. The types *intj* and *ger* have been dropped, since they were represented by fewer than 10 tokens in the doubly corrected data. Token counts in each partition are included for the remaining categories.

As expected, humans improved on the parser in all cases. Genre is only significant for *voyage*, and only in parser label assignment. More pronounced negative effects can be seen for *frag* and *other*, which carry over from parser to manual correction. Smaller effects for the question types can be observed, but are based on few tokens.

Although the results confirm the expected good performance on *decl* and lower importance of genre, imperatives emerge as unproblematic and only *frag* and *other* stand out. At the same time, it is possible there are alternative explana-

[7] Note that *decl* and *interview* represent the intercept for sentence and text type, meaning figures for other types represent deviations from these values.

[8] An anonymous reviewer has asked about other genre/type correlations in our data: beyond *imp*+*whow*, the more distant second is *wh* questions in the *interview* subcorpus: although the coefficient for *wh* is not significantly collinear in the model, these two category combinations together are responsible for almost 50% of the chi squared residuals for sentence type versus genre (*imp*+*whow*: 41.1%, *wh*+*interview*: 8.2%). Since *imp* forms 32.8% of the *whow* data but only 11.3% of all data, there is some potential for conflation between results for *imp* in *whow* and *whow* as a whole, whereas for interviews, *wh* is only 6.8% of the data – a very significant proportional deviation from the average of 2.3%, but still modest in absolute terms.