Different Flavors of GUM: Evaluating Genre and Sentence Type Effects on Multilayer Corpus Annotation Quality

Amir Zeldes
Department of Linguistics
Georgetown University
amir.zeldes@georgetown.edu

Dan Simonson
Department of Linguistics
Georgetown University
des62@georgetown.edu

Abstract

Genre and domain are well known covariates of both manual and automatic annotation quality. Comparatively less is known about the effect of sentence types, such as imperatives, questions or fragments, and how they interact with text type effects. Using mixed effects models, we evaluate the relative influence of genre and sentence types on automatic and manual annotation quality for three related tasks in English data: POS tagging, dependency parsing and coreference resolution. For the latter task, we also develop a new metric for the evaluation of individual regions of coreference annotation. Our results show that while there are substantial differences between manual and automatic annotation in each task, sentence type is generally more important than genre in predicting errors within our data.

1 Introduction

With the availability of increasingly diverse language resources and the viability of processing almost unrestricted Web data, domain adaptation and coverage of novel domains have become a major concern in NLP and corpus creation (see e.g. Daumé 2007, Finkel & Manning 2009, McClosky et al. 2010, Søgaard 2013). However, accuracy for both state of the art automatic tools and manual annotation of new tasks is typically reported on standard sources, typically newswire text, which often leads to overestimation of expected accuracy in both manual and automatic annotation. Manning (2011) points out that although we expect 97% accuracy from POS taggers on newswire, such a rate indicates an error every other sentence even within the training domain, and more in other domains or epochs. A major cause of problems in adaptation is the presence of unknown words from outside the training domain, which may be more influential than other aspects of the actual genre itself (cf. Plank 2011).

It has also been suggested that at least part of the source for these problems lies in less frequent kinds of utterances within and across domains, i.e. that domain adaptation may be folding in sentence type effects. For example, in an evaluation of the English Web Treebank, explicitly intended to expand the text types covered by reference Treebank data, Silveira et al. (2014:2898) remark that "[t]he most striking difference between the two types of data [Web and newswire] has to do with imperatives, which occur two orders of magnitude more often in the EWT." Specifically Silveira et al. found over 445 times more imperatives in EWT than in the Wall Street Journal corpus (Marcus et al. 1993). Despite this stark difference, there is remarkably little literature on sentence type as a factor in annotation quality or NLP tool performance. While sentence type is known to be important in computational models of language acquisition (see Frank et al. 2013), it has not been suggested that human annotators are affected by it. In the development of automatic annotation tools, explicit partitioning of sentence types for differential treatment is also rare (for an exception see Zhang et al. 2008 on machine translation).

Indeed, it is not clear whether sentence type is actually pertinent to annotation quality, especially for human annotators, who are generally able to understand most sentences without difficulty. The question we will be asking in this paper is

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 68–78,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

therefore whether sentence types are a better predictor of annotation quality than text type or genre, which is often postulated to be central without consideration of alternative explanations.[1]

2 Data

For our evaluation we will use the GUM corpus (Zeldes 2016)[2], a class-sourced, richly annotated multilayer corpus containing freely available texts from four different types: news articles from Wikinews, Wikimedia interviews, travel guides from Wikivoyage and how-to guides from wikiHow (abbreviated 'whow'). Each of these sources corresponds more or less to a different communicative intent, which lends itself to different types of sentences: news articles are narrative, telling about events, often in indicative past tense; travel guides are informational, giving modals of possibility and general truths about places; how-to guides are instructional, containing many imperatives and lists of ingredients; and interviews are conversational, often containing question and answer pairs or sequences. Interviews in particular could be expected to differ from the other types, due to differences between spoken and written language. The corpus contains 54 documents, totaling just over 44,000 tokens, as outlined in Table 1.

text type	source	texts	tokens
Interviews	Wikinews	14	12,661
News	Wikinews	15	9,402
Travel guides	Wikivoyage	11	9,240
How-tos (instructional)	wikiHow	14	12,776
Total		**54**	**44,079**

Table 1: Composition of the GUM corpus.

Document structure is annotated using TEI XML labels, and each text is annotated with POS tags and lemmas, dependency and constituent syntax, entities (using a subset of categories from OntoNotes, Hovy et al. 2006), information status (the scheme in Dipper et al. 2007), coreference, Rhetorical Structure Theory (Mann & Thompson 1988), and crucially, sentence type (see below). In this paper we will be concerned with:

1. POS tags – annotated manually using the extended Penn tag set used by the TreeTagger[3] (Schmid 1994)
2. Manually corrected Stanford Typed Dependencies (de Marneffe & Manning 2013)
3. Coreference annotation, including pronominal anaphora, lexical coreference and appositions (but not bridging, which is also annotated in the corpus).

Finally, the sentence type annotation layer supplies a kind of rough speech act or sentence mood, using an extended form of the SPAAC annotation scheme (Leech et al. 2003). The sentence types distinguished are given in Table 2.

tag	type	example
q	polar yes/no question	*Did she see it?*
wh	WH question	*What did you see?*
decl	declarative (indicative)	*He was there.*
imp	imperative	*Do it!*
sub	subjunctive (incl. modals)	*I could go*
inf	infinitival	*How to Dance*
ger	gerund-headed clause	*Finding Nemo*
intj	interjection	*Hello!*
frag	fragment	*The End.*
other	other predication or combination	*Nice, that!'* Or: *'I've had it, go!'* (decl+imp)

Table 2: Sentence type annotation in GUM.

Given genre metadata and sentence type annotations in the corpus, we would like to know which is a better predictor of errors on each layer.[4]

Our analyses of each data type will be addressed in separate experiments, similar in general configuration but adapted to the needs of the data type: POS tags in Section 4, dependencies in Section 5, and coreference in Section 6.

[1] An anonymous reviewer has pointed out that many other covariates of genre could be subjected to a similar treatment, in the vein of Biber's multidimensional analysis (see Biber 2009 for an overview), such as tense and other grammatical features. We agree completely: we are only beginning to understand the components of genre variation and how it interacts with annotation quality.

[2] The data is freely available under a CC license from http://corpling.uis.georgetown.edu/gum . We would like to thank the annotators, a current list of which is found at the same Web site.

[3] The tag set used by TreeTagger distinguishes forms of *be* (VB, 3rd person present VBZ,..) from *have* (VH, VHZ, ...) and other verbs (VV, VVZ, ...), as well as several punctuation tags and a special tag for *that* as a complementizer (IN/that). GUM also contains a second POS layer using the CLAWS5 tags (Garside & Smith 1997), which will not be evaluated here.

[4] An anonymous reviewer has asked about the decision to include the *sub* type as distinct from *decl*: this type was already in the existing annotation of GUM and was not added for this study. However modality is expressed syntactically e.g. via auxiliaries, ultimately influencing sentence structure, and semantic influence on humans should not be ruled out either.

3 Experimental setup

For each of the three tasks, POS tagging, dependency annotation, and coreference resolution, we first split the corpus into each of the four text types and collate responses from manual, automatic and gold annotation in GUM. Ordinarily, the manual annotation data released for a corpus is the same as the gold data – for this study we obtained uncorrected, single annotator versions of the data to approach annotation quality effects in an initial manually produced analysis.

Since GUM is a 'class-sourced' corpus, the unadjudicated annotations always represent work from relatively inexperienced student annotators, which was subsequently corrected by an experienced instructor. These corrections will be considered the 'gold' data for our evaluation.[5]

Once we have annotation graphs and labels from all three sources, we can easily compare manual and automatic annotation with the gold standard in each subcorpus. However comparisons across sentence types can be less straightforward: while POS tags can be evaluated in the different sentence types in isolation, coreference annotation cannot be easily evaluated while ignoring certain parts of the text. We therefore develop some extended metrics for the evaluation in Section 6. For all data sets, we keep track of the documents (and by proxy annotators) that contain each annotation as a random effect, and we will consider some competing independent variables, such as sentence length, as alternative explanations for annotation quality.

4 Part of speech tagging

4.1 Method

For the evaluation, we compare data from the annotators, who received only brief training, to three popular taggers: TreeTagger (TT, Schmid 1994), the Stanford Tagger (Toutanova et al. 2003) and Spacy (`https://spacy.io/`). Double corrected gold data was available for only 38,022 tokens, which are evaluated below. Since GUM was annotated using the TreeTagger's extended tag set, the most comparable evaluation will be between TT and human annotators. However, it is fairly straightforward to collapse the extended tag set into the more compact 36 tags used by the other taggers (*VVZ* and *VHZ* become *VBZ*, etc.), so that results for those taggers can be evaluated as well (though with somewhat less potential for errors, especially for the tag *IN/that*).

While our primary interest lies in gauging the relative influence of genre and sentence type, we would also like to consider some alternative explanations. Using mixed effects models from R's lme4 package, we will take individual document effects into account as a random effect. Mixed effects models (see Baayen 2008: 263-327 for an overview) allow us to assign some of the variance we see in the data to random effects, such as ostensibly unpredictable interpersonal variation between annotators, or the difficulty of particular documents: these factors are assumed to have a mean influence of 0 (since they are random), while positing individual intercepts for higher/lower baselines observable in our dependent variable (the error rate). Additionally, we also suspect that sentence length is a possible predictor of errors: for example, longer sentences may be more grammatically complex; or it could turn out that very short sentences (for example headings) lead to part of speech ambiguities. We therefore model length as a further fixed effect, which could be an alternative explanation for differences in error rates.

Although our null hypothesis must be an equal distribution of errors, we do not expect strong effects for text or sentence type in manual annotation, since tagging decisions are relatively local: trained annotators should be able to discern parts of speech even in heterogeneous sentences. Automatic taggers, by contrast, rely on the Markov assumption and learn tag distributions from chains of tokens, meaning that a greater influence of input type effects can be expected, especially in text types more dissimilar to newswire, on which the taggers are trained.

4.2 Results

Tables 3 and 4 give raw breakdowns of error frequencies across text and sentence types (asterisks designate significant predictors of error proportion in a simple linear model, for the annotation strategy in the respective row). The figures for TT are the most comparable to the manual fig-

[5] We have no doubt that the gold data also contains some errors, and that class-sourced data may be more erroneous than data obtained in other settings. But our premise is that manual annotation difficulties depending on genre and sentence types should still emerge in the comparison, especially since we will allow for document-by-document random effects.

	decl	frag	ger	imp	inf	other	q	sub	wh
Manual	93.87	94.70	**90.28***	93.14	93.20	95.34	**96.59++**	94.13	93.32
TT	**95.33+++**	90.46	**93.52+**	93.16	**79.59*****	90.60	93.34	94.63	92.30
Stanford	**95.21+++**	88.57	88.89	91.24	**78.91*****	90.50	**93.00+**	94.98	**93.09++**
Spacy	**94.43+++**	**87.81*****	**87.04*****	91.91	**82.99*****	89.94	94.37	94.38	94.11
Tokens	27,440	1,321	219	4,313	147	1,074	586	2,011	883

Table 4: Tagging accuracy by sentence type for manual and automatic annotation. Significance only indicated for deviations of more than 2% below the mean (with *) or above (with +).

ures, since the other two taggers are evaluated against the unextended tagset.[6]

Table 3 shows that genre effect sizes are modest for tagging. Manual annotation from scratch performs similarly to all of the taggers, and is only better for the how-to guides, which are the most accurate for humans, but worst for POS taggers. TT loses about 1% accuracy on this genre, while the other taggers lose about 2% accuracy; in other categories all three taggers are largely neck-and-neck, with Spacy surprisingly somewhat behind on news compared to other taggers.

	interview	**news**	**voyage**	**whow**
Manual	93.55	93.52	94.06	94.30*
TT	94.73	95.57*	95.21	93.44*
Stanford	94.50	95.78*	94.80	92.54***
Spacy	94.03	94.71	94.15	92.44***

Table 3: Tagging performance by genre, with significance in a simple linear model (*p<0.05;**p<0.001;***p<0.0001)

While a simple a linear model significantly correlates text type with performance at the 5% threshold for all annotation sources, only the slight differences in *whow* and *news* are significant predictors. Moreover, even before we consider a full multifactorial model, if we add document identity as a random effect in a mixed effects model with only genre as a fixed effect, the genre effect largely disappears, with the exception of the low *whow* performance by Spacy and the Stanford tagger. This suggests that most of what we are seeing is due to specific documents being more difficult for the taggers. In other words, humans and taggers do almost exactly as well across these text types.

Sentence type, by contrast, shows some stronger effect sizes, shown in Table 4. Since there are many sentence types, all rows are significant and very many values are significant at a 5% threshold; to improve readability significance is only indicated for deviations of 2% accuracy from the mean or more. Despite their significance, some of these are however based on very little data and should be interpreted with caution.

Gerunds, which are usually headings as in (1), are significantly worse for manual annotation, and infinitives as in (2) are worst for automatic tagging, but these are based on only 219 and 147 tokens respectively, so that results should be taken with a grain of salt despite their significance.

(1) *Hiring*/VVG *employees*/NNS

(2) *How*/WRB *to*/TO *Grow*/VV *Basil*/NN

Though the data is limited, the fact that these are mostly headings means it is possible that capitalization is causing problems in mistagging common nouns as proper nouns, which manual annotation is less susceptible to. Another possibility is that the shorter, more condensed sentence length makes these harder on account of missing function word cues (articles signaling nouns, etc.), meaning that length is a possible confound for the sentence type effect.

The remaining discrepancies are more certain, with about 87-90% accuracy in automatic tagging for *frag* and around 90% for the *other* type, based on much more data (1,321 and 1,074 tokens). For *frag*, we can suspect the reason is verbs: fragments lack a VP, which, assuming the verb can be recognized, would have a positive effect on tagging the surrounding arguments as nouns and their modifiers. For all three taggers, declaratives perform best by a wide margin, and as the gaps marked in bold show, other types are very substantially worse.

While these results are based only on sentence and text type separately, we can also check whether the sentence and text type effects are significant overall in a model that takes both into consideration, as well as the possible sentence length confound. Table 5 gives t-test values for the fixed effects in four mixed effects models including document identity as a random effect, and fixed predictors for text and sentence type,

[6] For all taggers we used the default English models supplied by their Web pages. The Stanford tagger model (*english-bidirectional-distsim*) did not distinguish tags for opening and closing quotation marks, which slightly boosts its accuracy.

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.

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 a subset of each document (e.g. interrogative sentences) without looking at other sentences.

More recently Martschat et al. (2015) introduced error analysis for mention pair types in the CORT system, which keeps track of each pair of mentions corresponding to a correct or incorrect linking decision in a mention-chain model.[9] For example, it is possible to diagnose precision or recall errors involving a pronominal anaphor with a common noun-headed antecedent, by counting correct and incorrect links of this type, in much the same way used by the MUC metric.

Building on Martschat et al.'s insights, we extend the MUC metric to features of single mentions involved in correct or incorrect links. We call this metric 'p-link', which stands for 'partitioned link score'. The basic idea is that a coreference failure (or success) has two equally responsible mentions in a consecutive mention-chain model. Each of the two mentions involved shares credit or blame for the classification decision. If a link partition is worth 1 precision or recall point, then involvement in a correct decision earns 0.5 points for the category that includes the mention at each end of the link.

Figure 1 illustrates this using the example from Pradhan et al. (2014), which has been extended with shading representing categories.

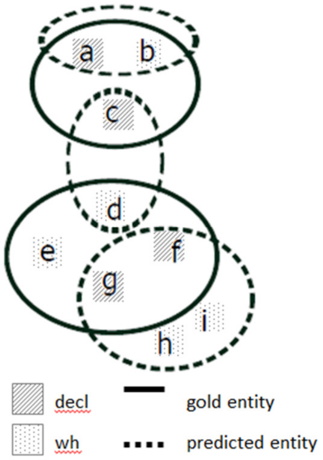

Figure 1: Gold (solid) and predicted (dashed) entities, with mentions in two categories distinguished by shading.

The solid oval represent two gold entities, with mentions {a,b,c} and {d,e,f,g}. Dashed ovals give three predicted entities, with mentions {a,b}, {c,d} and {g,f,h,i}. Note that mention e is not in any predicted entity, and h+i are not in the gold data. Pradhan et al.'s implementation of the MUC metric tallies the partitions with respect to gold and predicted mentions, such that a predicted link a+b is a correct positive (since a+b are in the same gold entity), c+d is a false positive, and the absence of predicted b+c is a false negative.

The p-link score builds on this by counting 0.5 points of correct positive, correct negative, etc. for each mention, such that points accrue for the respective category of that mention. The metric is a direct extension of Pradhan et al.'s definitions for recall (R) and precision (P):

$$p\text{-}link_{R,\pi} = \frac{\sum_{i=1}^{N_k}(|K_i^\pi| - p(K_i^\pi))}{\sum_{i=1}^{N_k}(|K_i^\pi| - 1)}$$

$$p\text{-}link_{P,\pi} = \frac{\sum_{i=1}^{N_r}(|R_i^\pi| - p'(R_i^\pi))}{\sum_{i=1}^{N_r}(|R_i^\pi| - 1)}$$

where K_i is the i^{th} entity in the key (gold) data (and Ri is correspondingly the i^{th} response entity); $|K_i^\pi|$ is the weighted partition magnitude within entity i, i.e. the number of instances of a mention from partition type π being either the source or target of a coreference link, multiplied by the weight 0.5 (since source and target may be of different types, and each is worth 'half a link'); and $p(K_i^\pi)$ is the set of elements of type π obtained by intersecting the key entities with the response entities, with each mention again being worth 0.5 points for its respective type π.[10]

Thus for the example in Figure 1, declaratives get 0.5 points for their correct involvement in a+b, but none for the missing link with c, and 1 point for their involvement in the correct g+f (since both are *decl*). The total possible links for declaratives in Figure 1 are worth 2 points (0.5 for a+b, 0.5 for b+c and 1 for g+f), so that *decl* scores a recall of 1.5/2 or 0.75 in this example. Indeed, only 1 of 4 *decl* link endpoints is missed in this example. We have implemented the p-link metric as an extension to Pradhan et al.'s original code, and our code is freely available.[11]

To test whether genre or sentence type has more influence on p-link, we evaluate manual and automatic coreferencer output, using a con-

[9] This approach assumes a 'mention-pair' model, in which each anaphor is linked to its antecedent in a chain. By contrast, 'mention-cluster' or 'entity-mention' models (see Rahman & Ng 2011) focus on entities as clustered groups of mentions referring to the same entity.

[10] Although we assign anaphors and antecedents equal weights of 0.5, other weights are conceivable.

[11] Code available at: `https://github.com/amir-zeldes/reference-coreference-scorers`.

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.

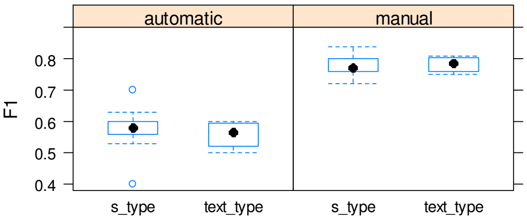

Figure 2: Box plots for p-link F-scores by partition using manual and automatic annotations.

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.

er). Sentence length is negatively correlated with manual precision and automatic recall (longer is harder), though there is no effect on manual precision. This can be explained by long sentences making human annotators miss mentions, but not resolve them incorrectly; the coreferencer, by contrast, prefers close antecedents, meaning long sentences offer more close competitors.

In terms of the partitions, none of the text type effects are significant, but several of the sentence types survive: fragments are still hard for the coreferencer, above and beyond prediction based on pronominality, sentence length and genre, but not for humans. Imperatives, by contrast, are significantly easier for everyone. These typically refer to at-issue, non-subject, lexical NPs, since imperatives have no overt subject. The imperatives in the data, typically instructions in how-to and travel guides are often adjacent to lexical remention of the same entities, making them easy to resolve via lexical identity (11). Fragments, by contrast, and especially very short ones that the model expects to be easy, sometimes corefer via synonyms, perhaps to deliberately avoid remention after headlines, as in (12). This makes them easy for humans, but difficult for the machine.

(11) *Read below for more of* [*the interview*] *in full.* [*Interview*] …

(12) [*Superstars*]
Each collection donated by the Andy Warhol Photographic Legacy Program holds Polaroids of [*well-known celebrities*]

Finally, the coreferencer is more likely than usual to get wh-referents right, beyond the positive effects of pronominality and short length. This suggests that wh-questions too have comparatively simple mention structure and tend to mention lexical NPs that are likely to recur verbatim or with identical heads, rather than more roundabout references (e.g. 13).

(13) - *What is [Heaven Sent Gaming]?*
- *[Heaven Sent Gaming] is basically me and Isabel*

7 Conclusion

The results from our data set indicate that, across the board, sentence type variation is a better predictor of annotation quality than genre. Although it is obvious that there are more sentence types than genres in our study, this result is not obvious: many patterns of style and vocabulary are specific to genres such as travel guides or interviews, and sentence types are cross-classified across all text types. There are more imperatives in how-to and travel guides, and more questions in interviews, but these types are attested in all genres, and the multifactorial models consistently choose sentence type with no remaining added effect for genre. Additionally, even a coarse binary factor such as pronominality can survive in a multifactorial model that finds sentence type significant, but not genre.

It should be noted that the genres surveyed here are not very distant: We are certain that adding Computer Mediated Communication (e.g. Twitter data) as a further text type would radically alter our results. However, given the scope of differences in annotation quality across sentence types, we would also expect to see strong effects of sentence type within and across more disparate genres, such as CMC data of various kinds.

A practical implication of this study is that it may be worth redoubling annotation quality control on sentence types known to be problematic for a certain task. As we have seen, these can vary between manual and automatic annotation, the automatic tool used, and the task itself. It is also clear that, as noted by Silveira et al. (2014), we are in great need of more diverse annotated datasets, and especially ones containing underrepresented sentence types, such as imperatives, questions and non-canonical sentences.

Acknowledgments

We would like to thank Shuo Zhang for helping to process the coreference data, the participants of the course LING-367 Computational Corpus Linguistics for contributing their annotations, and three anonymous reviewers for valuable comments on earlier versions of this paper. For a current list of contributors to GUM, see `http://corpling.uis.georgetown.edu/gum`.

References

Emilia Apostolova, Noriko Tomuro, Pattanasak Mongkolwat and Dina Demner-Fushman. 2012. Domain Adaptation of Coreference Resolution for Radiology Reports. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*. Montreal, 118–121.

R. Harald. Baayen. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Douglas Biber. 2009. Multi-Dimensional Approaches. In Anke Lüdeling & Merja Kytö (eds.), *Corpus*

Linguistics. An International Handbook. Vol. 2. Berlin: Mouton de Gruyter, 822–855.

Danqi Chen and Christopher D. Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Doha, Qatar, 740–750.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of ACL 2007*. Prague, Czech Republic, 256–263.

Stefanie Dipper, Michael Götze and Stavros Skopeteas (eds.). 2007. Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure. *Interdisciplinary Studies on Information Structure* 7.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical Bayesian Domain Adaptation. In *Proceedings of NAACL-HLT 2009*. Boulder, CO, 602–610.

Stella Frank, Sharon Goldwater and Frank Keller. 2013. Adding Sentence Types to a Model of Syntactic Category Acquisition. *Topics in Cognitive Science* 5(3):495–521.

Roger Garside and Nicholas Smith. 1997. A Hybrid Grammatical Tagger: CLAWS4. In Roger Garside, Geoffrey Leech and Tony McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, 102–121.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York: ACL, 57–60.

Mohammad Khan, Markus Dickinson and Sandra Kübler. 2013. Towards Domain Adaptation for Parsing Web Data. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*. Hissar, Bulgaria, 357–364.

Matthew Lease and Eugene Charniak. 2005. Parsing Biomedical Literature. In Robert Dale, Kam-Fai Wong, Jian Su and Oi Yee Kwong (eds.), *Proceedings of IJCNLP 2005*. Berlin: Springer, 58–69.

Geoffrey Leech, Tony McEnery and Martin Weisser. 2003. *SPAAC Speech-Act Annotation Scheme*. University of Lancaster, Technical Report.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3):243–281.

Christopher D. Manning. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing, 12th International Conference (Proceedings of CICLing 2011)*. Tokyo, 171–189.

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics* 19(2):313–330.

Marie-Catherine de Marneffe and Christopher D. Manning. 2013. *Stanford Typed Dependencies Manual*. Stanford University, Technical Report.

Sebastian Martschat, Thierry Göckel and Michael Strube. 2015. Analyzing and Visualizing Coreference Resolution Errors. In *Proceedings of NAACL-HLT 2015*. Denver, CO, 6–10.

David McClosky, Eugene Charniak and Mark Johnson. 2010. Automatic Domain Adaptation for Parsing. In *Proceedings of NAACL 2010*. Los Angeles, CA, 28–36.

Barbara Plank. 2011. *Domain Adaptation for Parsing*. PhD Thesis, University of Groningen.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng and Michael Strube. 2014. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the ACL*. Baltimore, MD, 30–35.

Altaf Rahman and Vincent Ng. 2011. Narrowing the Modeling Gap: A Cluster-Ranking Approach to Coreference Resolution. *Journal of Artificial Intelligence Research* 40(1):469–521.

Sujith Ravi, Kevin Knight and Radu Soricut. 2008. Automatic Prediction of Parser Accuracy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*. Honolulu, 887–896.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK, 44–49.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauery and Christopher D. Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland, 2897–2904.

Anders Søgaard. 2013. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. (Synthesis Lectures on Human Language Technologies.) San Rafael: Morgan & Claypool.

Kristina Toutanova, Dan Klein, Christopher D. Manning and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2003)*. Stroudsburg, PA: ACL, 252–259.

Jian Bo Yang, Qi Mao, Qiao Liang Xiang, and Ivor W. Tsang, Kian Ming A. Chai and Hai Leong Chieu. 2012. Domain Adaptation for Coreference Resolution: An Adaptive Ensemble Approach. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing* *and Computational Natural Language Learning (EMNLP 2012)*. Jeju Island, Korea, 744–753.

Amir Zeldes. 2016. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*.

Amir Zeldes and Shuo Zhang. 2016. When Annotation Schemes Change Rules Help: A Configurable Approach to Coreference Resolution beyond OntoNotes. In: *Proceedings of the NAACL-HLT 2016 Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*. San Diego, CA, 92-101.

Jiajun Zhang, Chengqing Zong and Shoushan Li. 2008. Sentence Type Based Reordering Model for Statistical Machine Translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, 1089–1096.

Shanheng Zhao and Hwee Tou Ng. 2014. Domain Adaptation with Active Learning for Coreference Resolution. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis, EACL 2014*. Gothenburg, 21–29.