A Detailed IAA Analysis

Individual annotators. Five annotators took part in this study. All are computational linguistics researchers with advanced training in linguistics. Their involvement in the development of the scheme falls on a spectrum: Annotator A was the leader of the project and lead author of the guidelines. Annotator B was the second most active figure in guidelines development for an extended period, but took a break of several months in the period when the guidelines were finalized (prior to the pilot study). Annotator C was involved in the later stages of guidelines development. Annotator D was involved only at the very end of guidelines development, and primarily learned the scheme from reading the annotation manual. Annotator E was not involved in developing the guidelines and learned the scheme solely from reading the manual (and consulting with the guidelines developers for clarification on a few points). Annotators A, B, and C are native speakers of English, while Annotators D and E are nonnative but highly fluent speakers.

Table 7 shows that agreement rates of individual pairs of annotators range between 71.8% and 78.7% for roles and between 74.1% and 88% for functions. This is high for a scheme with so many labels to choose from. Interestingly, there is not an obvious relationship in general between annotators' backgrounds (native language, amount of exposure to the scheme) and their agreement rates. It is encouraging that Annotators D and E, despite recently learning the scheme from the guidelines, had similar agreement rates to others.

Common confusions. In figure 3 we visualize labels confused by annotators in chapters 4 and 5 of *The Little Prince* (§5), summed over all pairs of annotators. The red and blue lines correspond to the local semantic groupings of categories in the hierarchy. Confusions happening within the triangles closest to the diagonal are therefore more expected than confusions farther out in the matrix. As discussed in §5, most disagreements actually do fall within these clusters (of varying granularity), indicating the scheme's robustness.

		B	C	D	E	avg	plr
A	role	78.2	74.1	78.7	74.5	76.4	86.1
	fxn	81.5	84.3	88.0	81.5	83.8	90.3
B	role		73.1	74.5	71.8	74.4	82.9
	fxn		77.3	81.0	74.1	78.5	83.8
C	role			73.6	72.7	73.4	80.1
	fxn			83.3	80.6	81.4	88.0
D	role				73.1	75.0	84.7
	fxn				81.0	83.3	91.7
E	role					73.0	83.3
	fxn					79.3	86.1

Table 7: Pairwise interannotator agreement rates, each annotator's average agreement rate with others ("avg"), and each annotator's rate of agreeing with the label chosen by the plurality of annotators ("plr"). Tokens for which there is no plurality (6 for both role and function) are included and counted as disagreement for all annotators. Figures are exact label match percentages.

The three most frequently confused scene roles are Agent/Originator (***his*** *report*, under Participant), Gestalt/Whole (*the soil* ***of*** *that planet*, Gestalt is the parent of Whole), and Theme/Topic (*I am not at all sure* ***of*** *success*, Theme is the parent of Topic). The three most frequently confused functions are Gestalt/Possessor (***your*** *planet*, Gestalt is the parent of Possessor), Theme/Topic, and Locus/Manner (*the astronomer had presented it ...* ***in*** *a great demonstration*, both are children of Circumstance).

B Features of the Feature-rich Model

For each of the neighboring words of the word or phrase to be classified (as described in §6.3), we extracted indicator features for:

1. the lowercased word, capitalization, and universal and extended POS tags,
2. the word being present in WordNet,
3. WordNet synsets for the first and all senses,
4. the WordNet lemma and lexicographer file name,
5. part, member, and substance holonyms of the word,
6. Roget thesaurus divisions of the word, if it exists,
7. any named entity label associated with the word,
8. its two and three letter character prefixes and suffixes, and
9. common affixes that produce nouns, verbs, adjectives, spatial or temporal words, and gerunds.

C Hyperparameters for the Neural Model

Table 8 presents the hyperparameters used by the neural system, for each of the four settings.

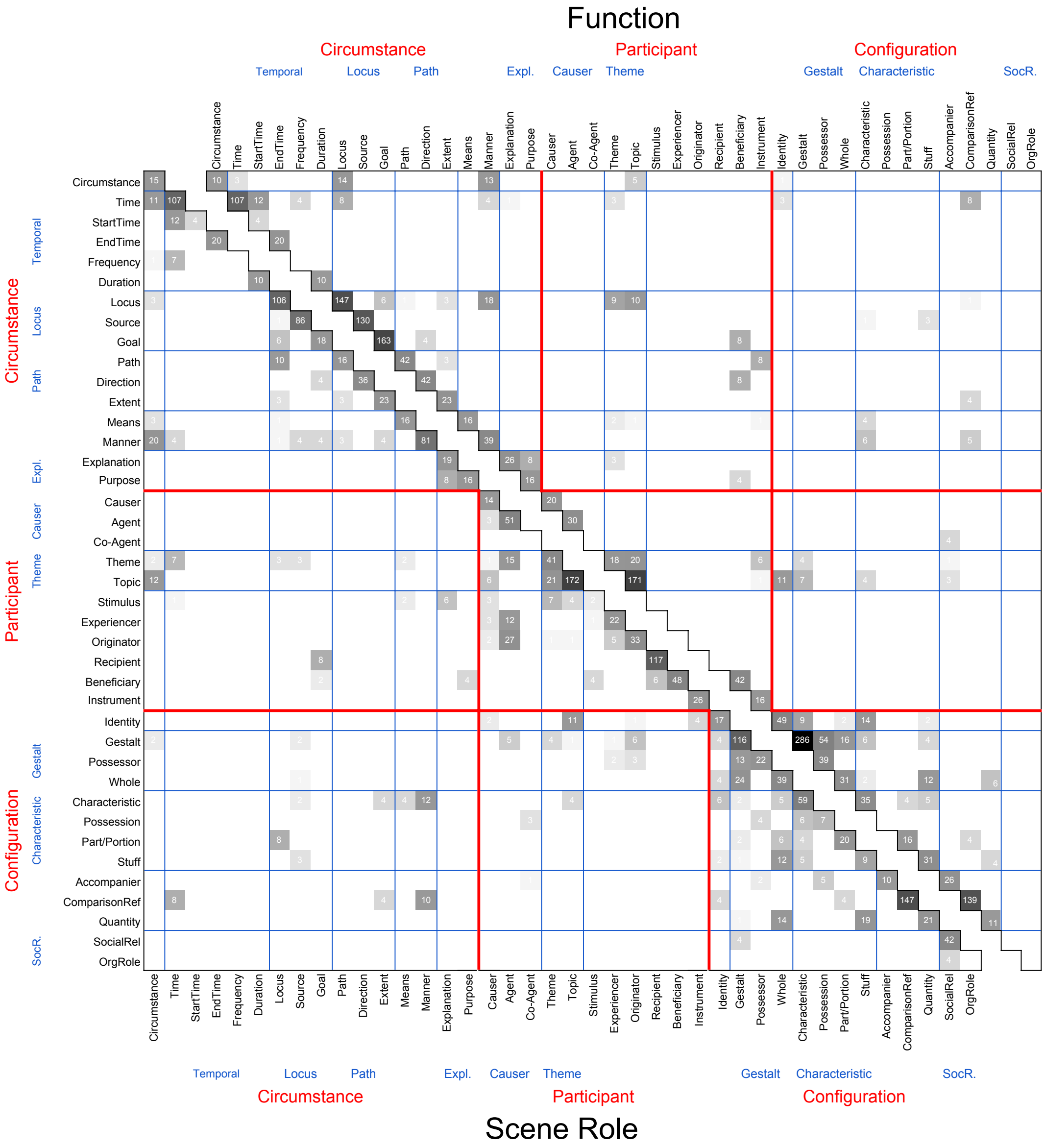

Figure 3: Confusion matrices for role (bottom/left) and function (top/right) labels, summed across all annotator pairs.

Hyperparameter	Auto ID/Auto Prep.	Auto ID/Gold Prep.	Gold ID/Auto Prep.	Gold ID/Gold Prep.
External Word2vec embd. dimension	300	300	300	300
Token internal embd. dimension	50	100	10	10
Update token Word2vec embd.?	No	No	No	No
Update lemma Word2vec embd.?	Yes	Yes	Yes	No
MLP layer dimension	80	80	100	100
MLP activation	tanh	tanh	relu	relu
BiLSTM hidden layer dimension	80	100	100	100
MLP Dropout Prob.	0.32	0.31	0.37	0.42
LSTM Dropout Prob.	0.45	0.24	0.38	0.49
Learning rate	0.15	0.15	0.15	0.15
Learning rate decay	0	0	10^{-4}	0
POS embd. dimension	5	25	25	5
UD dependencies embd. dimension	5	25	10	25
NER embd. dimension	5	5	10	5
GOVOBJ-CONFIG embd. dimension	3	3	3	3
LEXCAT embd. dimension	3	3	3	3

Table 8: Selected hyperparameters of the neural system for each of the four settings. With the exception of the external Word2vec embeddings dimension (which is fixed), the parameters were tuned using random grid search on the development set.