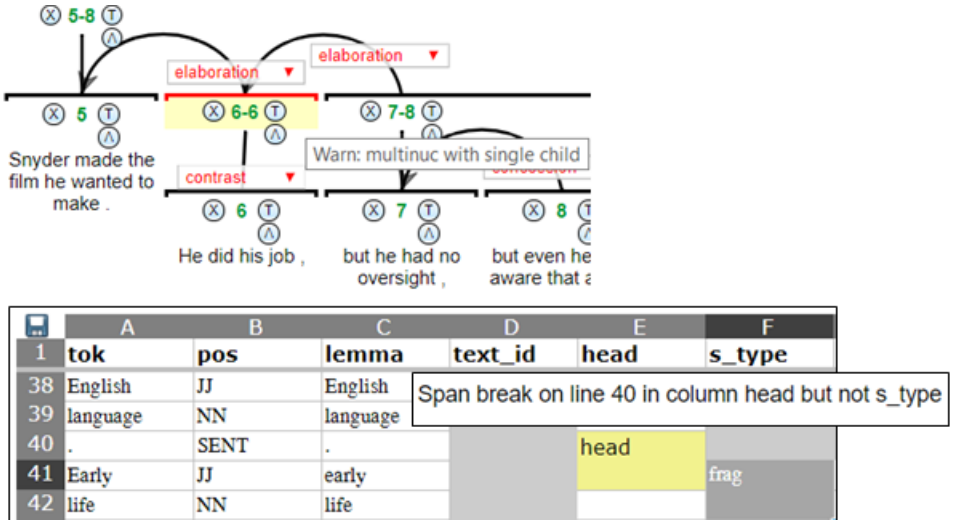
Validating and Merging a Growing Multilayer Corpus – the Case of GUM

This paper reports on expanding a class-sourced, richly-annotated and freely available corpus of English Web genres called GUM (Georgetown University Multilayer corpus, Zeldes 2017). Expanding the existing corpus of news, interviews, how-to guides and travel guides, we add four new genres: academic writing, biographies, fiction, and reddit discussions. These are annotated by students in the classroom using multiple online annotation tools to add TEI-XML structural markup, rough speech act information, POS tagging, dependencies, entity and coreference annotations, and discourse parses in Rhetorical Structure Theory (Mann & Thompson 1988).



	A	B	C	D	E	F
1	tok	pos	lemma	text_id	head	s_type
38	English	JJ	English			
39	language	NN	language			
40	.	SENT	.		head	
41	Early	JJ	early			frag
42	life	NN	life			

Span break on line 40 in column head but not s_type

Figure 1. Annotation tool warnings – top: discourse parse with single contrast node; bottom: sentence type 'frag' breaks in middle of heading annotation.

In the talk, we discuss challenges in maintaining high quality annotations in new genres across a wide range of annotation types. We present error-highlighting strategies in annotation tools (Figure 1) and validating merging tools (Figure 2), which collate and compare dependency annotations, POS tags, sentence types, discourse parses and more (cf. Dickinson & Meurers 2003 on catching errors in individual layers).

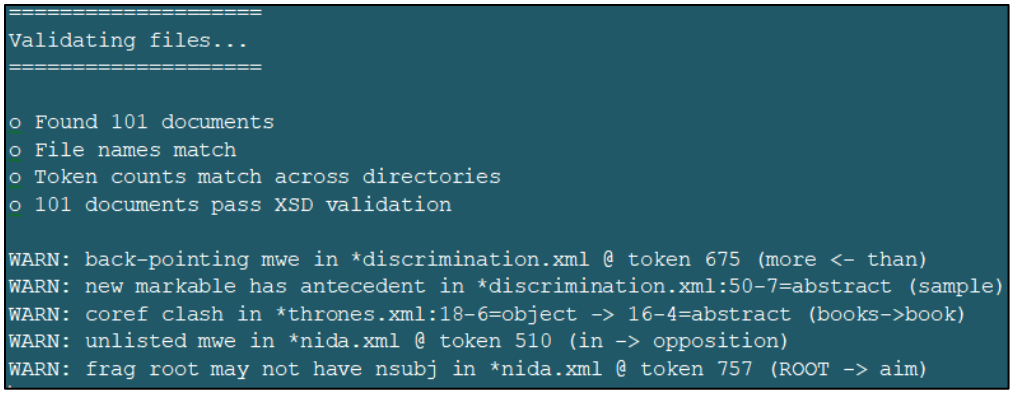


Figure 2. Merge validation output.

For example, our tools:

- Verify closed class configurations, e.g. ‘mwe’ annotations linking unlisted multiword expressions
- Check entity-type identity across coreference links
- Compare sentence and phrase borders in discourse parses, dependency parses and sentence type annotation
- Rule out implausible combinations, e.g. imperatives cannot dominate a subject function

We evaluate using an older corpus version created without these facilities. Results show that despite careful manual curation, per 10,000 tokens the merging tools catch an additional:

- 20 tagging errors
- 8 lemmatization errors
- 34 dependency errors
- 6 sentence type/border errors
- 11 coreference/entity errors
- 3 discourse parsing errors

These include errors preventing valid merging of multilayer data from different annotation tools, and are vital to maintaining high corpus quality.

References

Dickinson, Markus & W. Detmar Meurers (2003), Detecting Errors in Part-of-Speech Annotation. In: *Proceedings of EACL 2003*. Budapest, Hungary, 107–114.

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

Zeldes, Amir (2017), The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation* 51(3), 581–612.