Lang Resources & Evaluation (2017) 51:581–612
DOI 10.1007/s10579-016-9343-x

ORIGINAL PAPER

The GUM corpus: creating multilayer resources in the classroom

Amir Zeldes[1]

Published online: 5 February 2016

Abstract This paper presents the methodology, design principles and detailed evaluation of a new freely available multilayer corpus, collected and edited via classroom annotation using collaborative software. After briefly discussing corpus design for open, extensible corpora, five classroom annotation projects are presented, covering structural markup in TEI XML, multiple part of speech tagging, constituent and dependency parsing, information structural and coreference annotation, and Rhetorical Structure Theory analysis. Layers are inspected for annotation quality and together they coalesce to form a richly annotated corpus that can be used to study the interactions between different levels of linguistic description. The evaluation gives an indication of the expected quality of a corpus created by students with relatively little training. A multifactorial example study on lexical NP coreference likelihood is also presented, which illustrates some applications of the corpus. The results of this project show that high quality, richly annotated resources can be created effectively as part of a linguistics curriculum, opening new possibilities not just for research, but also for corpora in linguistics pedagogy.

Keywords Multilayer corpora · Classroom annotation · Coreference · Information structure · Treebank · Parsing

1 Introduction

Among the trends in corpus linguistics in recent years, there are at least two developments that have promoted an explosion of complex language data becoming readily available: the advent of the age of multilayer annotations and the expansion

✉ Amir Zeldes
amir.zeldes@georgetown.edu

[1] Georgetown University, Washington, DC, USA

of the base of corpus creators and related software to allow for collaborative, distributed annotation across space and time. Corpora have grown progressively more complex and multifactorial, going beyond tagged, or even syntactically annotated treebanks to encompass multiple, simultaneous levels of analysis. For example, the Switchboard corpus (henceforth SWBD, Godfrey et al. 1992) and the Wall Street Journal corpus, (WSJ, see Marcus et al. 1993, both American English) have been repeatedly annotated to add information. Examples include coreference analysis or named entities (e.g. for WSJ in OntoNotes, Hovy et al. 2006, which was extended to include Mandarin Chinese and Modern Standard Arabic), phonetic and further disfluency annotation or prosody and ToBI breaks (SWBD, Calhoun et al. 2010), as well as discourse functional annotation (the RST Discourse Treebank based on WSJ, Carlson et al. 2001). For research on Continuous Speech Recognition, portions of the WSJ corpus were even read out loud and recorded (Paul and Baker 1992). Some corpora have been constructed as multilayer resources from the outset or shortly thereafter, such as the HCRC Map Task Corpus (Anderson et al. 1991, Scottish English), the ACE corpora (Mitchell et al. 2003, Mandarin, Arabic and English), the Potsdam Commentary Corpus (German, see Stede 2004; Stede and Neumann 2014) or the Manually Annotated Sub-Corpus of the Open American National Corpus (MASC, Ide et al. 2010, American English).

At the same time, the spread of corpus methodology next to theoretical approaches to linguistics, and the development of corpus and computational linguistics curricula, have meant that a growing number of projects are no longer carried out by a small group of experts within a funded project. Some recent projects have collected data with more substantial student participation over longer periods of time, and with a view to possible future expansions (e.g. learner corpora such as Falko for German, see Lüdeling et al. 2008; Reznicek et al. 2012; or historical corpora such as PROIEL, Haug et al. 2009, a comparative corpus of ancient Indo-European languages; and the RIDGES corpus of early German scientific texts, Krause et al. 2012). This trend has grown in particular in the creation of historical, philological corpus resources, in curricula where annotation tasks could be integrated into the classroom, and hiring external annotators is rarely an option. Expanding the corpus over multiple years with different annotators is then an attractive possibility [see e.g. the Homer Multitext project, Blackwell and Martin (2009) in Classics, but also RIDGES above for historical German].[1]

The present paper will describe a project of this nature, inspired by the resources cited above, and focused on contemporary English data for corpus and computational linguistics research: the Georgetown University Multilayer Corpus (GUM), collected as part of a linguistics curriculum at Georgetown University. Although

[1] Another alternative to student participation is crowdsourcing over platforms such as Amazon Mechanical Turk or CrowdFlower (see Sabou et al. 2014 for an overview of recent projects and some best practices). Here individuals with minimal or no training can carry out relatively simple tasks on a large scale. However the costs involved need to be covered, which is difficult to sustain for an open-ended corpus, and some more complex annotations, such as syntactic analysis, are difficult to find qualified persons to do. It is possible that the unavailability of crowdsourcing and other resources for older languages has contributed to the popularity of classroom annotation or 'class-sourcing' in these domains (I'm indebted to an anonymous reviewer for pointing the latter term out).

GUM is a small corpus by most standards, currently containing approx. 22,500 tokens,[2] it contains a very large amount of annotations (over 180,000), which allow for new types of queries, studies, and ultimately, research questions (cf. Ragheb and Dickinson 2013 for similar argumentation in rich annotation of L2 learner data). A main goal of this article is to evaluate the reliability of a corpus produced via student annotators with a background in linguistics but limited additional training on such a wide variety of annotation tasks, and to analyze the design and software decisions that facilitate projects of this sort. A secondary goal will be to show what we can learn by combining types of data that have existed in corpora for a while in new ways, and this will be illustrated by an example study on modeling coreferentiality in Sect. 4.

Corpus creation within the curriculum brings with it a unique set of challenges, including, but not limited to:

- Necessary compromises between pedagogical and research needs in selecting materials and annotation guidelines
- Time for training is limited by the nature of semester based teaching
- Integration of quality control and grading
- Complete turnover of annotation personnel each semester

To meet these challenges, frameworks must be developed to select documents that will interest students but come from an open ended, yet relatively homogeneous pool; to implement rapid teaching and feedback collection in learning each annotation task; to document decisions for later iterations or further corpus expansion; and to minimize friction with software interfaces, dedicating a maximum of time to the annotation tasks at hand. These topics will also be central to the discussion below.

With these goals and constraints in mind, the remainder of this article is structured as follows: Sect. 2 introduces the corpus design and discusses the choice of documents comprising the corpus, their usability for linguistic research, and corpus licensing and availability. Section 3 presents and evaluates the annotation layers produced for the GUM corpus as part of a linguistics curriculum: the subsections detail and contrast human annotator and NLP tool performance for part of speech tagging, document structure, syntactic annotation, entity tagging, information status annotation, coreference resolution, and finally Rhetorical Structure Theory analysis. Section 4 presents a case study illustrating the usefulness of the corpus by examining the likelihood of lexical NPs to have an antecedent in the text in a linear mixed effects model derived from the data. Section 5 concludes the discussion with a summary of lessons and best practices learned from the project and a discussion of prospective extensions to the data.

[2] These numbers represent the first round of documents from GUM, collected in 2014; at the time of writing, a second round is being processed which contains over 21,500 tokens from the 2015 iteration of the same course, bringing the total up to about 44,000 tokens (see more details below).

2 Corpus design

Selecting the right kinds of texts in correct proportions for a representative corpus has been a hotly debated topic for over two decades (Biber 1993; Crowdy 1993; Hunston 2008). Corpus design is of course intimately related to the research questions that a corpus is meant to answer (cf. Reppen 2010), and GUM is no exception. A major research interest behind the collection of the corpus is gathering information for discourse modeling and investigating the ways in which discourse referents are introduced, mentioned and referred back to, and how they enter into cohesive relations to form a larger discourse across text types and different communicative intentions.

While it is desirable for the corpus to cover a wide range of language types for both these and other research questions, for a corpus that is manually collected in the classroom, even reaching the size of corpora collected 50 years ago, such as the Brown corpus ($\sim$1 million tokens, Czech Hachek: Kučera and Francis 1967) is not possible. On the other hand, a small corpus does not mean that genre or text type variation should not be a goal: sampling from multiple language types substantially increases structural variability in the data [arguably the most important factor in representativeness, cf. Biber (1993: 243)] and allows for studies of language variation using metadata categories. Although there are certainly limitations to studies using small corpora, such as problems for studying specific areas of lexis and other infrequent phenomena [for example, recognizing associations between rare words like *jigsaw* and *puzzle*, cf. Sinclair (2004: 188–190)], many linguistic categories are frequent enough for differences in distribution to be apparent even in a small corpus, such as parts of speech, many syntactic phrase types and dependency functions, and even discourse entities, rhetorical functions and relationships between these, as we shall see below.

For specific text selection, I would like to argue that students in a classroom setting should be allowed to work on texts that interest them, rather than, for example, working on financial reporting simply because a corpus like WSJ offers the chance to extend an existing Treebank, whose language has also been studied in detail many times before. Selecting one's own text to work on creates a connection with the document, rather than having the content imposed from above. This is especially important if the text is to be analyzed repeatedly using different annotation schemes, in order to maintain interest.[3] Nevertheless, some restrictions must apply, including the realistic ability to complete the annotation project (not selecting texts that are too long/particularly difficult), the limitation to a fixed set of genres or sources to ensure consistency of the data (a corpus with five types and five texts each is more useful than a corpus of 25 disparate texts), and data should be available in digitized form to save time (effectively, texts usually come from the Web). A further important goal is to make the resulting data available in the public domain, which imposes certain copyright restrictions on data selection.

To implement these principles, the students working on the creation of the GUM corpus were allowed to pick any text of an appropriate length (some 500–1000

[3] The motivational effect of choosing one's own text is similar to Computer Assisted Language Learning tools that allow learners to work on a text of their own choosing in the target language, often from the Web (see the REAP project, http://boston.lti.cs.cmu.edu/reap/ and VIEW, http://sifnos.sfs.uni-tuebingen.de/VIEW/). I thank an anonymous reviewer for pointing this out.

Table 1 Documents in the GUM corpus

Text type	Source	Documents	Tokens
News (narrative)	Wikinews	6	5051
Interview (conversational)	Wikinews	7	6535
How-to (instructional)	wikiHow	7	6701
Travel guide (informative)	Wikivoyage	5	4369
Total		25	22,656

words based on a word-processor count),[4] on a subject of their choosing, from one of four text types: interviews, news articles, instructional texts and travel guides. These types were chosen to represent a variety of communicative purposes—conversational, narrative, instructional and informative—while relying on resources which could be made freely available under a Creative Commons license and could easily be expanded to more texts without complex digitization/recording efforts. Specifically, it was decided to draw on openly available Wiki resources, so that news and interview texts could be obtained from Wikimedia's Wikinews, instructional texts from wikiHow and travel guides from Wikivoyage. Table 1 summarizes the sources and their extent in the corpus.[5]

The documents from the Wikimedia foundation, including the News, Interview and Travel subcorpora are available under CC-BY (attribution) licenses, while wikiHow makes its texts available under a CC-BY-NC-SA license (non-commercial, share alike). GUM data is offered under the same licenses to the public, depending on the subcorpus in question, in multiple formats at: http://corpling.uis.georgetown.edu/gum. The corpus can also be searched online using ANNIS (Krause and Zeldes 2014), at: http://corpling.uis.georgetown.edu/annis.[6] The interface can be used to search through all annotation layers concurrently, and many of the annotations visualized in the sections below were rendered using this tool.

3 Implementing and evaluating classroom annotation

Evaluating a multilayer corpus in a uniform way is a difficult task because of the heterogeneity of the annotations involved. This is compounded by the fact that annotating in the classroom means that time constraints lead to different procedures

[4] Each of 21 students enrolled in the class selected a single text for annotation throughout the class. In one unusual case, a text which turned out to be too short after segmentation was supplemented by a second text of a similar length from the same genre. Three further texts were contributed, two by the instructor, and one by the teaching assistant; these were left out of the evaluation below. The course was open to both undergraduate and graduate students, but graduate students represented the majority of participants.

[5] The second round of data in 2015 adds 29 further documents from the same text types. See the corpus website for the latest data.

[6] ANNIS is an open source browser based platform for accessing multilayer corpora, originally developed at Potsdam University and currently in development at Humboldt University in Berlin and Georgetown University; see http://corpus-tools.org/annis for more information.

for different tasks: while it is feasible to annotate parts of speech fully manually, syntax trees can only be edited in a reasonable amount of time if parser output is corrected. As a result, the following sections each describe different procedures that were implemented to create each annotation layer, and how they were evaluated. In all cases, care should be taken when examining figures for accuracy: students were encouraged to discuss difficult cases in the classroom setting, meaning they were not working in the typical isolated manner used to evaluate e.g. inter-annotator agreement on a new annotation schema. While enforcing isolated work for annotators may be justified in a research project, it is directly at odds with the teaching philosophy at the core of the present project: that open discussion of annotation problems and difficult cases in class promotes learning Linguistics. The figures below should therefore be taken explicitly to represent an evaluation of the quality of data produced by classroom annotation as a holistic method, not the quality of work one can expect from students working in isolation.[7]

As noted above, students were allowed to pick any text they wanted from the four sources cited above, though the length restriction meant that some texts were not suitable (too short or too long). In some cases, a contiguous section of sufficient length was extracted from a longer text that a student was interested in, provided that it formed a coherent text in itself—this was especially the case in some of the longer travel guides (e.g. sufficient sections would be collected, and then some remaining part, such as 'accommodations' information was dropped), and also in the how-to guides (often if there were multiple distinct 'methods' to do something, only one or two would be included, and alternatives were omitted). The language for texts was required to be English, though small amounts of foreign words did occur naturally within the selections. Use of non-standard or erroneous language was marked up (using <sic> tags, see Sect. 3.2), and was overall infrequent. In one instance a poem was included in a text, and this was marked up as well (Sect. 3.2); the view taken in the corpus design was that variation of language types is enriching, and should not be avoided by excluding unusual types of English.

3.1 Tokenization and part of speech tagging

A fundamental task for any annotated corpus is tokenization and part of speech (POS) tagging, which are also the basis for many of the subsequent annotation layers. This also serves as a major part of the curriculum for working with corpora: gaining an understanding of POS tags and learning a specific tag set allows students to work with annotated data later on. Annotators initially applied the extended Penn Treebank tag set and were given access to the guidelines (Santorini 1990, extended

[7] An anonymous reviewer has commented on consultation of the instructor as a possible source of skewing for annotator accuracy. While it is true that some errors were certainly averted by discussion with the instructor during class, it is conversely very much not the case that there was time for the instructor or TA to advise students on the individual details of much of their annotations, given the size of the class and the corpus. Notwithstanding the degree to which the instructor or TA were able to directly reduce error rates, the data below should be taken as an evaluation of the quality of a 'class-sourced' corpus, which as we will see, contains errors nonetheless.

with tags from the TreeTagger's tag set, Schmid 1994).[8] Their texts were automatically tokenized using the TreeTagger's tokenizer and manually corrected by the students, who subsequently tagged the data. There were almost no corrections to tokenization, and tagging was done completely by hand, from scratch. Students used an XML aware text editor of their choosing, so that they would not have to change tools for the markup described in the next section; however the instructor recommended the freely available Notepad++ text editor for the Windows operating system (https://notepad-plus-plus.org/) or TextWrangler for Mac (http://www.barebones.com/products/textwrangler/), which were subsequently used by almost all participants (except those with access to high quality commercial software, e.g. oXygen, http://www.oxygenxml.com/). Output was then validated using a spreadsheet editor against possible tag values.[9]

For the evaluation of tagging quality, annotator performance was compared against the automatic output of the TreeTagger: cases where annotators agreed with the tagger were accepted as correct, and conflicts were adjudicated by the instructor or teaching assistant. Figure 1 charts POS tagging data from 21 annotators,[10] where the bottom bars are cases of agreement with the automatic tagger (eq), above which are stacked the cases where annotators made correct decisions missed by the tagger (+), cases of disagreement that could be defended either way (OK) and cases where student tags had to be corrected (err).

The 'OK' category mainly contains cases where the tagger's decisions represent form based decisions, whereas the annotator gives a more contextual or semantically motivated analysis. Often these cases involved non-standard uses of punctuation, as in (1), but in some cases there were real linguistic issues and ambiguities to discuss, as in (2) (the alternative POS tags are given after the slash separated by a pipe as *[student]|[tagger]*):

(1) *We have –/**SENT**|*: (hyphens mark cut off utterance, more like the tag 'SENT' than ':' for other punctuation)

(2) *a new obsession is taking **hold/RP|NN** on the internet.*
(unclear if 'take hold' is a separate phrasal verb or the same as 'take a hold' with NP object)

The proportion of 'true' annotator errors ('err') varies somewhat (mean 4.78 %, SD 2.4 %) but is of very similar magnitude to the proportion of improvements on the

[8] Main additions in the extended tag set are special tags for the verbs *be* (VB*) and *have* (VH*) versus lexical verbs (VV*), more tags for punctuation, and a special tag for the word *that* as a complementizer (IN/that). Especially the more fine-grained use of punctuation tags was useful for data from the Web, since there is a wide range of different symbols and functions. We also added a second layer of POS annotations using the UCREL CLAWS5 tag set (Garside and Smith 1997), to allow for comparison and combined searching. This layer has not yet been corrected or evaluated, but is available in the corpus online.

[9] Spreadsheet validation has been replaced in the most recent iteration of the course in favor of a Perl script which simultaneously controls XML markup guidelines and gives more verbose error messages. Students are instructed to use validation before submitting assignments to ensure that tag names are formally possible, thereby ruling out typos.

[10] Data from the one annotator working on two shorter texts has been collapsed into one group, here and below.

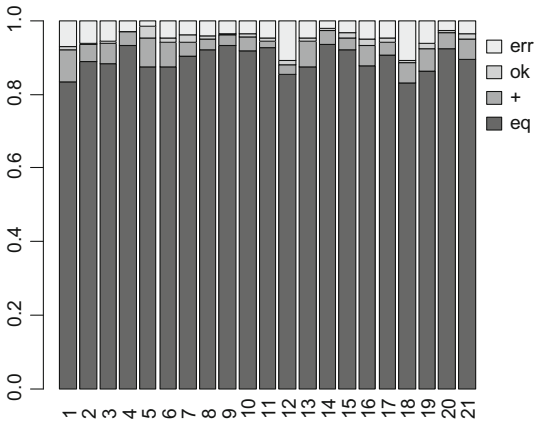

Fig. 1 Manual annotation performance adjudicated against the TreeTagger, broken down by annotator

tagger's output (mean 4.76 %, SD 1.79 %), meaning that adjudicating disagreements with the tagger is worthwhile, at least for these genres. This is even more crucial since tags are used for further steps in the pipeline, such as syntactic annotation. An evaluation of the agreed upon cases (student = tagger) remains outstanding, but these are expected to be, by and large, correct.

3.2 Structural TEI annotation

A more varied task and a good way to learn about XML markup from a pedagogical point of view is to annotate and preserve the document structure present in the original layout. For GUM we used TEI p5 markup (the Text Encoding Initiative format, Burnard and Bauman 2008; see http://tei-c.org) and chose a subset of structural elements through classroom discussion. When in doubt students were instructed to consult the TEI guidelines. Since texts in the corpus come from Web pages, it is relatively easy to train annotators to express elements already in the HTML code as XML elements, and an XML aware text editor helps to avoid errors. In group discussions within each text-type group, the subset of phenomena to be captured in TEI markup was decided on (i.e. joint discussions were held by the students who worked on interviews, travel guides, etc.). The final discussion including all groups in class led to the inclusion of the following 14 elements in the final corpus (Table 2).[11]

Some of these elements are specific to only one or two texts (lines of poetry occur in only one, but were marked up on the argument that some searches might wish to include or discard these, and further texts might have them in the future), though most are ubiquitous and relatively easy to agree on. The presence of <s> tags for

[11] Other elements were produced in keeping with TEI markup, including hierarchical divs for sections and subsections, but these were discarded from the merged corpus before it was indexed for searching, as they turned out to be rather inconsistent across genres semantically.

Table 2 TEI annotations in the GUM corpus

Element	Attributes	Description
figure	rend	Marks the position of a figure in the text and its rendering
head	rend	Marks a heading and its rendering
hi	rend	A highlighted section with a description of its rendering
incident	who	An extralinguistic incident (e.g. coughing), and the person responsible
item	n	Item or bullet point in a list, possibly with number
l	n	A line, e.g. in poetry, with its number
lg	n, type	A line group with the group's number and type (e.g. stanza)
list	rend, type	List of bullet points, with appearance and list type
p	rend	A paragraph and its rendering
quote		A quotation
ref	target	An external reference, usually a hyperlink, and its target
s		A main clause sentence span
sic		A section containing an apparent language error, thus in the original
sp	who	A section uttered by a particular speaker with a reference to that speaker

main sentence segmentation is important, since it is used as the basis for subsequent tasks, such as syntactic analysis.

The presence of the <sic> tag is also of particular interest, since it allows some evaluation of the 'messiness' of the text. Texts from the Web are sometimes criticized for their unclear provenance (cf. Lüdeling et al. 2007), and in some cases may be produced by non-native speakers, which is a concern. In a Wiki-based corpus, it is often impossible to speak of a single author, so that any hope of discussing attributes of a writer is not realistic to begin with. In practice, only one text showed multiple instances of apparently non-native usage of the sort in (3), whereas most <sic> tags indicate deviations that are minor typos (4), fillers or plausible native speaker errors (5) or naturally occurring disfluencies/non-standard language (6).

(3) *It is* <sic>*recommend*</sic> *that you use short words.*
(4) *the pictures of the muscles* <sic>etc.</sic> (for etc.)
(5) *it's really important to be able to get as* <sic>*a*</sic> *many international trips throughout the year* (superfluous *a* either falsely transcribed/spoken by a native speaker interviewee)
(6) *For the last year SNY has broadcast footage of me with my poems, so quite a few fans* <sic>*known*</sic> *about the "Mets Poet"* (interviewee is a native speaker)

In total, there were 14 sic tags, spanning 24 tokens in 6 of the 25 documents, giving a rough estimate of the relatively minor amount of problematic data with respect to grammaticality or correct orthography in the Wiki-based data.

3.3 Syntactic annotation

Annotating syntax is a highly complex task, and one at which linguistics students have an obvious advantage compared to annotators without linguistic training. While techniques such as crowd-sourcing may be used for other, potentially more intuitively resolvable tasks such as sentiment analysis (Hsueh et al. 2009), textual entailment or word sense disambiguation (Snow et al. 2008), creating, or even just correcting parses has not been treated using crowd methods before.[12] As a result, the possibility of applying distributed annotation to syntax in a linguistics curriculum is particularly exciting, and as I will discuss below, also of great value in teaching.

For the syntactic analysis in GUM, the sentence spans from the TEI annotation (<s> elements) were used as input units, and manually corrected tokenization and part of speech tags from the previous step were fed to a parser for better accuracy. Initially data was parsed automatically using the Stanford parser (Socher et al. 2013), and the resulting constituent trees were retained in the corpus. For manual correction, however, it was decided that dependency trees would be faster to correct and easier to teach, and the constituent trees were therefore converted to non-collapsed Stanford dependencies (using CoreNLP, see Manning et al. 2014) and these were corrected by each student using the collaborative online annotation interface Arborator (Gerdes 2013) and the Stanford guidelines (de Marneffe and Manning 2013).[13] The resulting constituent and dependency trees are searchable in tandem and visualized together in ANNIS to leverage as much syntactic information as possible (Fig. 2).

Using an online interface such as Arborator substantially facilitated training, since it only requires a browser for students to run and offers drag and drop functionality to edit dependency edges or alter labels. This allowed us to dedicate more time to discussing guidelines and difficult cases in class.

To evaluate accuracy, approximately 100 tokens (rounded to the next sentence) were taken from each annotator and corrected again manually by the instructor or TA, resulting in some 2280 tokens being checked by hand, or approximately 10 % of the corpus. Figure 3 shows the range of error rates in dependency attachment,

[12] Even for binary sentiment analysis (negative/positive), Hsueh et al. (2009: 30) report gold standard average accuracy of 0.974 for expert annotators, as opposed to 0.714 for Mechanical Turk annotators with minimal training. Although an evaluation attempting syntactic analysis via Mechanical Turk or Crowdflower remains outstanding, it is doubtful that complete parses could be obtained in this way, even when using many concurrent annotators. Simpler subtasks however, such as PP attachment disambiguation in new genres, have been shown to be possible using some semi-automatic support and quintuple annotation of the same data via Mechanical Turk (Jha et al. 2010), so that using crowdsourcing to improve specific aspects of parses in a multilayer corpus is very much a possibility.

[13] The decision to correct parser output, rather than annotate from scratch, was largely motivated by time constraints. Gerdes (2013: 94) suggests that student annotation of dependencies from scratch can achieve a rather low 79 % average attachment accuracy for French, also supporting the idea that parser correction is a good strategy. Although it would have been possible to use a native dependency parser, using constituent parses as a basis is known to produce good results (see Cer et al. 2010; I thank an anonymous reviewer for pointing this out), and also has the added advantage of representing constituent and dependency structure side by side in the merged corpus.

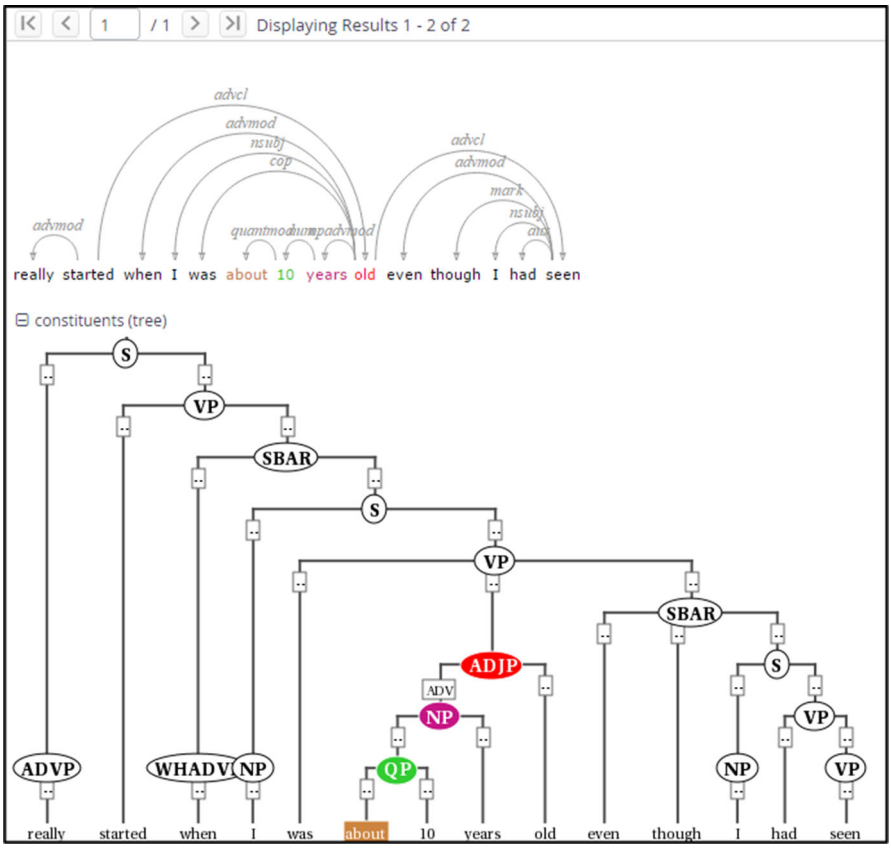

Fig. 2 Concurrent constituent and dependency annotations viewed in the ANNIS interface

labeling errors if no attachment error was made, and total proportion of tokens exhibiting any of these errors.[14]

The Stanford Parser version used for this data (3.4) has an unlabeled constituent parsing accuracy of around 90 % (Socher et al. 2013), and the newer version 3.5 (which was not available at the time) has unlabeled dependency attachment accuracy of around 92 % for Penn Treebank data and around 90 % for labeled dependencies (Chen and Manning 2014).[15] Given these accuracy rates, it is not surprising that the parser, represented by 'x' in Fig. 3, scores in the mid-80s, faring

[14] The reason for counting labeling errors only for correctly attached dependencies is that an incorrectly attached, but otherwise correct label is not seen as correctly ascertaining the function of the word in the sentence. Note that the numbers for the parser and students mean slightly different things: parser errors are those caught by either students or instructors, while student performance indicates how much correction is still required after the student pass.

[15] Similar accuracy has been achieved working from scratch by allowing 4–5 student annotators to work on the same sentences and taking a majority vote (see Gerdes 2013: 95 for collaborative F-scores of 0.91–0.92, working on French). However quadruple annotation in the context of GUM would have meant a very substantial reduction in corpus size, and also in the variability of data discussed in class.

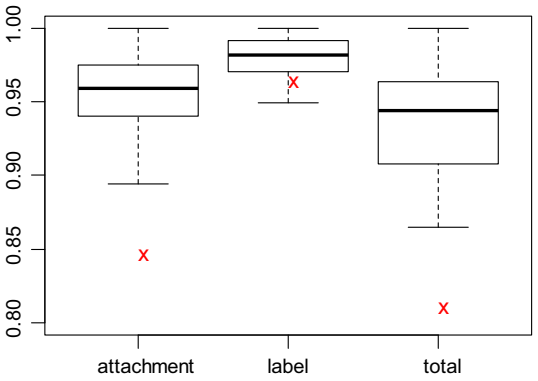

Fig. 3 Annotator and parser accuracy for dependencies. Parser performance is an 'x' below each box

considerably worse on the texts in this corpus, which are outside of its training domain. Since for many research questions texts other than the typical newswire type will be interesting, it's important to understand how usable parser output is, and what kinds of improvement are possible when allowing students with no previous treebanking training to correct parsing results. At the same time it's clear that even with syntax training, students can learn a lot from naturally occurring sentences. These will often feature phenomena that are not discussed in linguistics introductions and inevitably lead to interesting discussions about how non-canonical utterances fit into the annotation scheme. To give just a few examples: do captions of the sort "image: photographer" have a subject/predicate structure? How do URLs expanding on previous material relate to that material syntactically? What is the syntactic structure of dates—where is the head? Is "October 3rd" a kind of "3rd" like a nominal compound? What is the correct analysis for lists of ingredients, including complex quantifiers, quantities and appositions (measurements in the metric/imperial system…). Is "Method 2 to the rescue" a subject-predicate construction? These are all very real concerns for parsing unrestricted text across genres and they lead to genuinely open intellectual discussions about syntactic constructions that students previously hadn't thought about.

While Fig. 3 suggests that annotators were generally able to improve attachment accuracy from the mid-80s to the mid-90s as judged by the adjudicator, more modest improvements can be seen on correcting the label of a correct attachment: if the parser correctly selected the head, it also assigned the correct label quite frequently. With both error types taken as a whole, the proportion of tokens exhibiting no error was raised from around 80 % to somewhere in the low 90s, which can be seen as quite substantial. A related question is how the error rates interact with the different text types in the corpus. For parser performance, we would expect news texts to work best, while the unusual syntax of instructional texts (e.g. prevalence of imperatives) and more varied syntax of spoken interviews might contain more problems. Figure 4 shows the distribution of error rates across text types.

While the prediction that 'how to' articles from wikiHow will be difficult is borne out by the highest error rate, there is no significant difference in parser errors

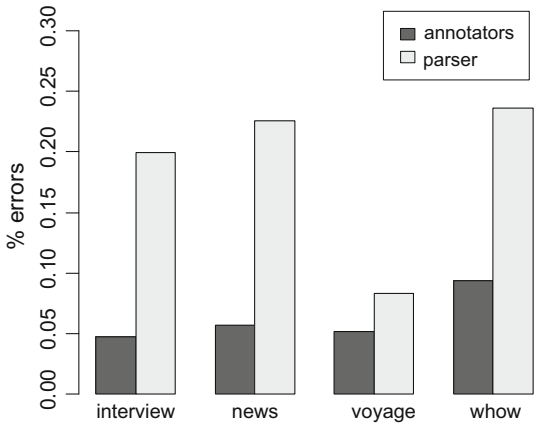

Fig. 4 Dependency parsing error rates and remaining annotator errors across text types

except, surprisingly, the very good performance of the parser on travel guides ('voyage', significantly fewer errors at $p = 1.452e{-}11$, $\chi^2 = 53.4748$). Annotator performance is also relatively homogeneous, with about 5 % errors present pervasively, but with significantly more errors for 'how to' (9.3 %, significant at $p = 0.002397$, $\chi^2 = 14.4099$). The latter are due at least in part to the presence of lists of steps or ingredients that are often difficult to analyze syntactically, whereas issues like presence of imperative syntax or difficult deixis (e.g. situational reference to parts or pictures) were not generally a problem for humans, at least as far as a qualitative evaluation can establish.

In sum these results suggest that parser performance on even relatively standard English Web genres is rather limited,[16] but correction by student annotators with very limited experience can bring accuracy up to the 95 % area that is often expected for corpus resources, all while learning the basics of building and using syntactically annotated corpora and dependency grammar.

3.4 Entities, coreference and information structure

On top of the morpho-syntactic annotation layers, many research questions as well as practical applications require semantic information regarding the discourse referents in texts, when they are initially introduced in the text, how they are referred back to, and what licenses and allows the resolution of their realization as pronouns, definite or indefinite NPs (see Sect. 4; for theoretical and typological background see Givón 1983; Lyons 1977: 177–196; Grosz et al. 1995, among others). To expose this information, GUM contains entity type annotation, coreference annotation and information structural annotations coding the information status of each referent at each mention.

[16] Especially as opposed to blogs, chat or Twitter data. This confirms the criticalness of creating manually annotated gold standards for Web genres and new social media for domain adaptation of automatic parsing; see Silveira et al. (2014).

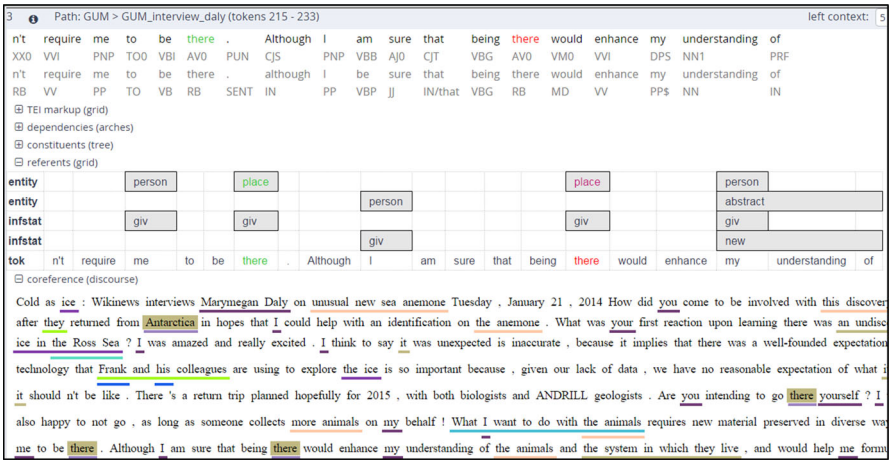

Fig. 5 Entity, information status and co-reference annotation in ANNIS. Two hits for '*there*' co-refer, and go back to '*Antarctica*' as shown in a full document view. The document view underlines coreferent entities and colors them when clicked, while the grid above shows all entities in the search result. (Color figure online)

For entity types we used a collapsed version of the OntoNotes annotation scheme (Weischedel et al. 2012), which was reduced to 11 types, both in order to facilitate learning and to prevent very sparse categories in the relatively small corpus (see similar considerations in the next section as well). However, in order to allow the study of all discourse referents in the texts, the selection of markables was substantially expanded over OntoNotes to include all referential NPs (and some non-NPs, see below), and not just coreferring or named entities; this focus necessitated some changes. The annotation scheme contains the categories below, which relate to the OntoNotes scheme as follows (Table 3).

With the exception of OBJECT and ABSTRACT, all other GUM categories are taken from the OntoNotes scheme, but are applied also to non-named and non-coreferring entities (e.g. an unnamed "some country" is still a PLACE, even if it is not subsequently mentioned). Most collapsed categories are a straightforward simplification of the OntoNotes guidelines, such as not distinguishing different kinds of locational entities or time expressions (though dates are specifically annotated on the TEI layer, making them distinct). The categories OBJECT and ABSTRACT were introduced in order to cover a variety of non-named entity NPs, and were found to be useful for subsuming otherwise rare categories, such as LAW and DISEASE or PRODUCT and WORK OF ART. The categories ORDINAL and CARDINAL are not annotated, and instead where a more complete NP is absent, they are annotated as the respective type of the referent being counted. For example, an 'MMPI form' (a kind of personality test) was annotated as an OBJECT, so a subsequent mention as 'one' is annotated in the same way:

(7) *[a mini-MMPI]*$_{\text{OBJECT}}$ *that I developed… if you'd like to see [one]*$_{\text{OBJECT}}$

Table 3 GUM entity type annotation scheme compared with OntoNotes

GUM	Subsumes OntoNotes
PERSON	
PLACE	GPE, LOCATION, FACILITY
ORGANIZATION	NORP, ORGANIZATION
OBJECT	PRODUCT, WORK OF ART and all other concrete objects
EVENT	
TIME	DATE, TIME
SUBSTANCE	
ANIMAL	
PLANT	
ABSTRACT	LANGUAGE, DISEASE, LAW and all other abstractions
QUANTITY	PERCENT, MONEY and all other quantities
(variable)	ORDINAL, CARDINAL

Although the scheme diverged from OntoNotes in the respects listed above, students were given the OntoNotes guidelines as a point of reference, both for pedagogical value and as a first authority on questionable cases.

For information status, the three way distinction 'giv(en)', 'acc(essible)', 'new' and the markable selection guidelines were taken over from Dipper et al. (2007). Following the guidelines, 'giv' translates to something already mentioned in the preceding discourse, while 'acc' refers to entities that require no introduction (e.g. generic, such as 'the sun', deictic/indexical, such as 'here', 'I', etc.) and 'new' to entities not introduced or inferred from previous discourse. No finer grained distinctions of activation (e.g. optional given-active/inactive in Dipper et al.) were made, and annotators were instructed to leave idiom NPs (e.g. '*on [the other hand]*') completely unannotated, rather than assigning the category 'idiom' from Dipper et al's guidelines. Markables were allowed to be nested, and were annotated for every referential NP, pronouns (including adverbial 'there'), and even clauses, but only when these were referred back to by a pronoun, for the benefit of coreference annotation.

The coreference annotation scheme was based on OntoNotes in terms of linking guidelines, such as non-annotation of subject-predicate coreference by default (no coreference within '*[John] is [a teacher]*'), inclusion of possessives, and the distinction between co-referring and appositional NPs, but the scheme was extended with additional types of coreference, such as bridging and cataphora, following the German TüBa-D/Z scheme (Telljohann et al. 2012). Five types are distinguished in total, given in Table 4.

The bridging type distinguishes entities that are immediately accessible in context by virtue of a previous mention of a related referent (e.g. part-whole: *[a car]* <-bridge-*[the wheels]*), but not when an explicit possessive appears (in which case the possessive is anaphoric to the previous NP: *[I]* <-*ana*-*[my] hands*; we also annotate the nesting referent *[my hands]*, but it is not considered bridging to *[I]*). Coreference, entity types and information status were annotated simultaneously

Table 4 Coreference annotation types in GUM

Type	Direction	Description
ana	Back	Anaphoric pronoun
cata	Forward	Cataphoric pronoun
appos	Back	Apposition
bridge	Back	Bridging
coref	Back	All other types of coreference (e.g. nominal re-mention)

Table 5 F-scores for markable and coreference detection and label accuracy for coreference relations, entity types and information status

	Precision	Recall	F-score		Accuracy
markables	0.9533	0.9419	0.9476	*coref rel*	0.9129
coreference	0.9185	0.8113	0.8616	*entity type*	0.9209
				infstat	0.9542

using WebAnno (Yimam et al. 2013). Figure 5 illustrates the final searchable result in ANNIS.

To evaluate annotation accuracy on these tasks, the entire data was reviewed by the instructor or TA, and errors were counted in terms of missing markable, incorrect markable, wrong value for information status or entity type, missing coreference edge or incorrectly annotated coreference. Mean F-scores for correct markable detection and for coreference detection are given alongside average accuracy rates in Table 5.

It is clear that coreference suffers more from low recall than precision compared to markable detection: annotators were more likely to disregard or not notice a relation than to connect entities against the guidelines' instructions.[17] It is also clear, however, that the guidelines are not always decidable, and much of the learning process as part of the curriculum was recognizing and discussing problematic cases: is an athlete talking about "*we*" in fact referring to her teammates as a set of 'persons', or to the team as an 'organization'? Does "*the Internet*" have an always accessible, generic information status in the same way as "*the sun*"? For bridging especially there are very many relations that might be considered to supply an indirect link where opinions differ: for example, in a sequence such as "[*an arrogant person*] *is less likely than* [*others*]…", one might take a word like 'others' to always bridge back to an antecedent (what it is 'other' or different from), or one might consider this to be part of the meaning of 'other' and assign the phenomenon to a different level of the analysis. In the context of the classroom, these are not

[17] For a very similar result in an annotation experiment with 14 students using two different interfaces doing coreference in somewhat different data (including Twitter), see Jiang et al. (2013), who report an F-Score of 0.83 in coreference pairs for people and locations, using the better of the two interfaces. Jiang et al. also remark on high human precision but low recall, the opposite of current NLP.

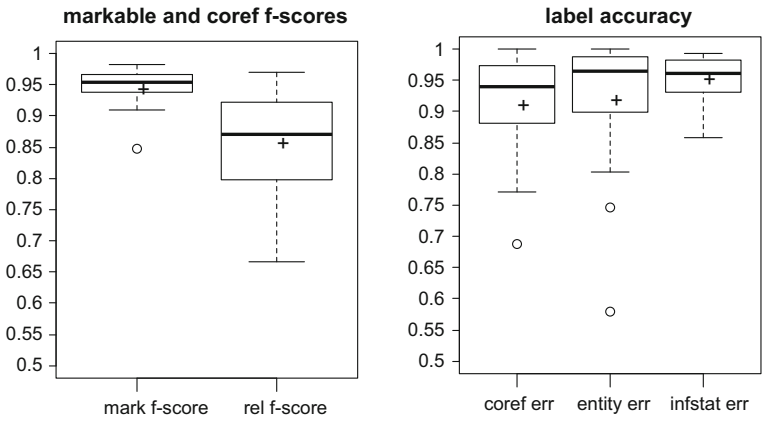

Fig. 6 Range of F-scores and accuracy rates for the data in Table 5

necessarily nuisances but sources of fruitful discussion, and an impetus for linguistic research on these topics.

Despite these complex issues, the adjudicated accuracy of coreference annotations is rather high, with performance on par with state of the art projects: Hirschman et al. (1998) reported low 80 % for both precision and recall of coreference in the MUC scheme, while in OntoNotes, average agreement with adjudicated results is at 91.8 % (Hovy et al. 2006, see Lüdeling et al. 2016, for an overview). Information status accuracy is also good, with over 0.95. For a rough comparison, note that for the three-way scheme used here, Ritz et al. (2008) report a range of $\kappa = 0.6$–0.8, while Nissim (2006) reports two-way given/new annotation performance of $\kappa = 0.902$ on Switchboard, and Riester et al. (2007) find $\kappa = 0.78$ for a six-way scheme for German (see Lüdeling et al., to appear for discussion). Although κ values cannot be reported for the single annotators of this layer, and as mentioned above, annotators were not working in isolation but in a classroom, the infrequency of corrections suggests satisfactory performance (cf. Jiang et al. 2013 for similarly successful results using crowdsourcing techniques and student annotators).

Some of the scores are more consistent than others, and some results may also be affected by outliers, which is not surprising considering the size and heterogeneous background of the group of annotators. Figure 6 gives the distribution of errors and F-scores across annotators for recognition of a markable or coreference relation of some kind on the left, and the assignment of the correct label type (coreference link type, entity type and information status) on the right.

Entity annotation in particular suffers from some outliers, and manual inspection reveals that the two documents with worst performance relate to a strong preference to categorize entities as concrete that were later adjudicated to be abstract, or the other way around. Intangible events and concepts, or abstracts that may have a physical realization (e.g. forms of writing as a process vs. end product) may be judged either way to some extent, meaning that guidelines must be refined on this point.

Table 6 Performance of CoreNLP dcoref on a test document from GUM

	All	No synonyms	No events	No events/synonyms
precision	0.6363	0.6363	0.6511	0.6511
recall	0.3835	0.3943	0.4117	0.4242
F-score	0.4786	0.4869	0.5045	0.5137

To get a rough idea of how annotators compare with state-of-the-art NLP on coreference resolution, one document was annotated using CoreNLP's deterministic 'dcoref' system (Lee et al. 2013) and compared to the manual annotation. In order to provide maximal compatibility with training data, the 'news' text type was selected. Lee et al. report F-scores of around 60 % on several non-gold annotated datasets, such as MUC, OntoNotes and CoNLL data, but the data for which the system is built does not match up to the GUM scheme exactly, since, for example, bridging is not meant to be recognized. For this reason bridging relations were discarded from the evaluation. The system is also not optimized for event level coreference going back to non-nominal markables (though it does occasionally produce these in the presence of pronominal anaphors) and cannot be expected to reliably identify coreference of synonymous lexical NPs (i.e. 'different names for the same thing'). Table 6 gives F-scores for the system on the test data when these factors are considered or ignored.

The F-score is in the area of 50 %, depending on the data included in the evaluation, and not including bridging. Out of 18 non-bridging referent chains in the document, spanning 68 markables, the system was able to identify 11, but also found 5 spurious chains, and one of the correct chains was split into two. These results show that while automated systems do an admirable job of finding a large part of coreference relations in an unseen document, very substantial gains can be seen when human annotators are involved, even with relatively brief training.

3.5 Discourse annotation with RST

Annotating the discourse structure of a document in terms of the rhetorical effect a speaker/writer aims to achieve on the recipient of the text is a complex, but potentially very rewarding task that has been implemented in corpus studies primarily within the framework of Rhetorical Structure Theory (RST, Mann and Thompson 1988; Taboada and Mann 2006; see Stede 2008 for criticism and some suggestions on complementing RST with other multilayer annotations for discourse research). RST postulates that documents can be segmented into so-called "elementary discourse units" (EDUs), which are non-overlapping, contiguous spans of text that relate to other EDUs, or groups of EDUs, to form structures of argumentation. For example one EDU can 'justify' the utterance of another EDU, give 'evidence' for its veracity, 'concede' apparent contradictions, etc. The exact number and types of these relations vary between implementations of RST (see below).

The first task of an RST analysis is the segmentation of text into EDUs, which generally correspond to sentences or smaller units, such as clauses, or in some

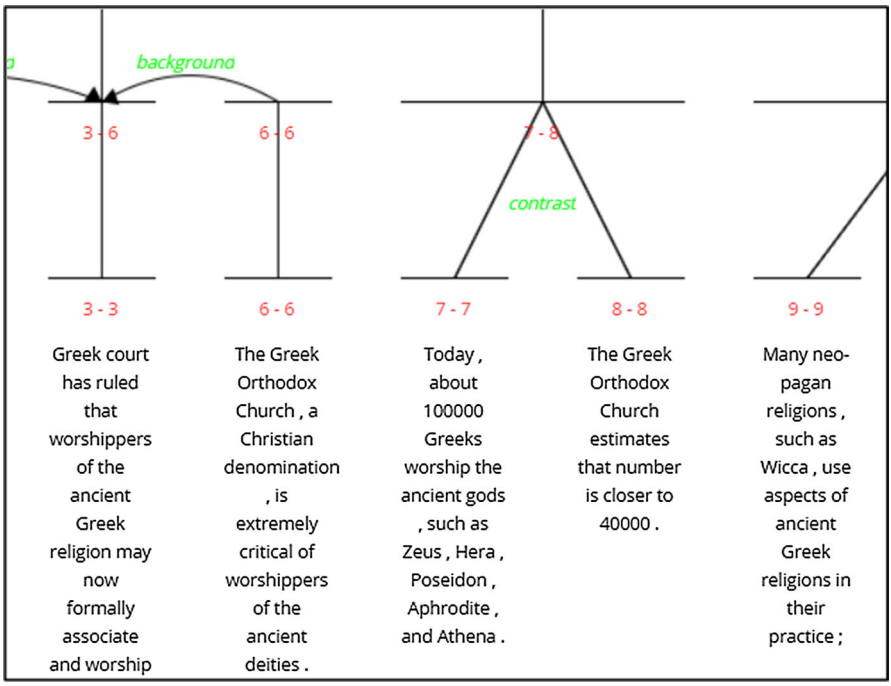

Fig. 7 RST visualization for a document in ANNIS

approaches even prepositional phrases (e.g. 'because of X' may be used to supply a reason, much like a whole adverbial clause). For GUM, we chose to follow the definitions used by the RST Discourse Treebank (Carlson et al. 2001), the largest RST corpus available, making the sentence spans used in the previous sections the maximal units for EDUs, but instructing annotators to segment adverbial subordinate clauses (but not subject or object clauses), as well as infinitival clauses (e.g. infinitive to express purpose) as EDUs. The corpus contains a total of 2663 RST nodes (including EDU groups), or about 107 nodes per document, which may be considered fairly large documents in terms of RST annotation projects. Segmentation and the subsequent analysis of relations were carried out using RSTTool (O'Donnell 2000).[18] Figure 7 shows part of an RST analysis for a document from the news section in ANNIS. As the figure shows, there are two major structural types of relations: *satellite-nucleus* relations, e.g. a satellite can provide 'background' information for a nucleus (curved arrows), or *multi-nuclear* relations, such as 'contrast', in which both EDUs (or sets of EDUs in a tree) are equally important, and stand in a symmetric relationship.

The inventory of relations used in implementations of RST varies widely, though often a distinction is made between a large fine-grained set and a coarser set used to group finer relations into more easily distinguishable clusters. Marcu et al. (1999)

[18] For the latest iteration of GUM, a new browser-based interface called rstWeb has been developed, which is being used for annotating RST online and facilitating collaborative annotation and adjudication over a server. See http://corpling.uis.georgetown.edu/rstweb/info/ for details.

annotated data from WSJ, the Brown corpus learned section and the MUC7 coreference corpus. They report using some 70 different relations (50 available for satellite-nucleus relations and 23 multi-nuclear types) with internal clusters, including an 'other-relation' option for difficult cases. Their reduced set at the highest level of cluster abstraction contains only 16 relations. Carlson et al. (2001) describe a development of the same data set having 53 satellite-nucleus relation types and 25 multi-nuclear relations, and state that higher abstraction levels were "often preferred" by annotators, with the highest level of the taxonomy containing the same 16 classes, such as a general 'cause' relation not distinguishing sub-types of cause, such as 'volitional-cause', 'non-volitional-cause' etc.

For projects in other languages, inventories have been substantially smaller, with a more or less flat taxonomy based directly on Mann and Thompson (1988)'s seminal paper on RST. For German, the Potsdam Commentary Corpus (Stede 2004) has 31 relations (including different topological subtypes, such as two types of 'evaluation' for use as a nucleus or satellite, but not including purely structural devices, such as the 'span' relation to group multiple units in a tree). Redeker et al. (2012) used 32 relations for a corpus of 80 written Dutch texts, including multiple variants for multi-nuclear and satellite-nucleus relations (e.g. 'restatement' can be either).

Going up to 70 relations seemed unrealistic for the brief course time available for GUM, which meant that an inventory more similar to the higher level clusters, or that used by Stede or Redeker et al. had to be preferred. Additionally, the same sparseness considerations that motivated reducing the entity inventory in the previous section apply to RST as well: adding rare categories could literally contribute singleton values to a corpus of this size. In determining the specific relations to be used, both the frequency of occurrences for relations in other corpora was considered (e.g. 'unless' is amongst the least frequent in previous corpora, and was therefore a prime candidate for dropping) and the similarity and potential for confusion with other relations. All causal relations of the type 'volitional-cause', 'non-volitional-cause', and 'reason' relations were reduced to one core 'cause' relation similarly to Carson et al.'s highest taxonomic level, despite some differences in the underlying definitions. Similar multi-nuclear relations such as 'list', 'joint', 'conjunction' and 'disjunction' were also reduced to the semantically most underspecified 'joint'. Table 7 gives the list of 20 relations that were used and taught to annotators in the space of 2 weeks of course sessions (counting multi-nuclear and satellite-nucleus variants of 'restatement' as distinct).

To build the annotation graph and assign the relations we used Carlson et al.'s (2001: 5) 'Style 2' practice of first segmenting large parts of the text (at least whole paragraphs, often the entire text) and then building subtrees for locally coherent fragments, which were joined together at higher levels to form the macrostructure of the document.[19]

[19] 'Style 1', which consists of immediately building relations for incrementally segmented texts, proved slower, matching Carlson et al.'s notion that it is less suitable for texts of substantial length (GUM documents average 64.12 EDUs, somewhat above the RST Discourse Treebank with 56.59 EDUs, cf. Carlson et al. 2001: 7).

Table 7 RST relations used in the GUM corpus

Relation	Structure	Relation	Structure
Antithesis	Satellite-nucleus	Motivation	Satellite-nucleus
Background	Satellite-nucleus	Preparation	Satellite-nucleus
Cause	Satellite-nucleus	Purpose	Satellite-nucleus
Circumstance	Satellite-nucleus	Result	Satellite-nucleus
Concession	Satellite-nucleus	Solutionhood	Satellite-nucleus
Condition	Satellite-nucleus	Contrast	Multinuclear
Elaboration	Satellite-nucleus	Joint	Multinuclear
Evaluation	Satellite-nucleus	Sequence	Multinuclear
Evidence	Satellite-nucleus	Restatement	Multinuclear or satellite-nucleus
Justify	Satellite-nucleus		Satellite-nucleus

Evaluating accuracy for RST annotations in terms of segmentation, satellite/nucleus attachment, and relation labels is complicated, for reasons that have been reviewed in the literature (see Marcu et al. 1999 for an overview). Put briefly, for segmentation it is possible that segments from multiple annotators do not overlap neatly. Because graphs may not match even in the units of the tree hierarchy, and not just label/attachment mismatches, each combination of segmentation borders identified by any annotator must be considered, strongly skewing measures such as kappa. For attachment, an evaluation of (binary) branching structures is possible if no empty transitions are assumed by enumerating all possible trees as viable options.

However the addition of optional levels of hierarchy complicates this evaluation, since any number of levels may be added to the tree as a form of disagreement. As an example, consider analyses A and B in Fig. 8.

There is a substantial labeling and attachment disagreement in the figure about the ‘cause’ (A, above) versus ‘result’ relationship (B, below) in two analyses of the same text. However there is also a topological disagreement about how many levels of hierarchy should appear above the contrast in units 7–8 (an extra level in B) and above EDU 3 (“Greek court…” extra level in A). Labeling disagreements can in turn be measured either for a predetermined given tree structure (an unrealistic stipulation), or they must somehow factor in differences dictated by the graph topology. Finally, it is not clear whether differences stemming from ‘leaf’ level changes should be considered more or less important than macroscopic differences in document structure.

Because of these difficulties and the fact that time constraints only allowed one annotation pass and correction by the instructor, it was decided to evaluate differences in terms of edit operations between the annotator’s version and the corrected version based on the guidelines. This is by no means an optimal metric for evaluating the quality of RST annotation as a task in general (a key concern for Marcu et al. 1999), but does give an idea of how variable annotator performance was overall. Figure 9 gives the range of accuracy in terms of instructor corrections

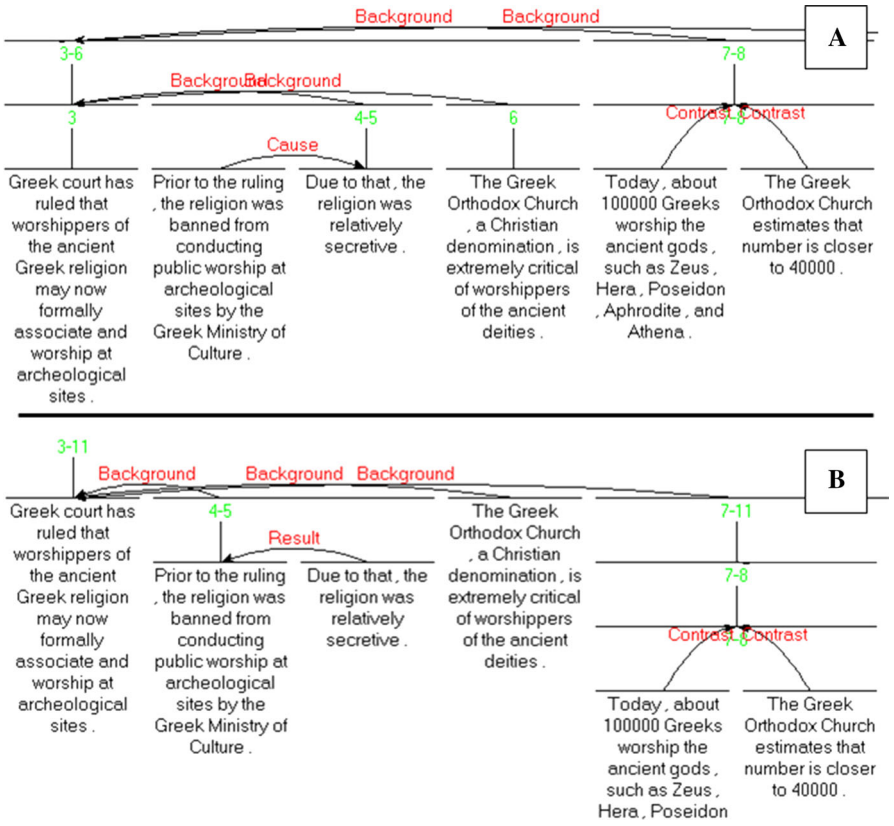

Fig. 8 Examples of relation/attachment disagreement: 'cause' (from EDU 4 to 5) in **a** but 'result' (5–4) in **b**; and graph-topological disagreement: addition of a hierarchical layer above EDU 3 ("Greek court…) in **a**, versus addition of a layer above the contrast in EDUs 7–8 in **b**

for different aspects of the data, normalized to the number of EDUs in each document.

On average, 4.07 % EDUs were added to those included by annotators, i.e. about 1 EDU needed to be added for every 25 EDUs initially submitted (but recall that sentence <s> spans were taken as default EDUs). In only one case was an EDU segmentation viewed as redundant and was removed (not plotted). Additions of levels to the tree mandated by the guidelines, including a full span encompassing a section after a heading (and not just the heading attached to the head segment of the section), were relatively frequent, proportional to 7.72 % (i.e. close to 8 added grouping spans per 100 segments). The most important part of the analysis is however the attachment and labeling performance, with attachment accuracy of 87.22 % and labelling accuracy of 86.58 % as compared to the 'gold standard' after instructor adjudication. These numbers should be taken with a grain of salt, since RST analyses are often controversial at the level of fine details: the 'gold standard' itself cannot be viewed as similarly reliable to equivalent resources for syntax or even coreference.

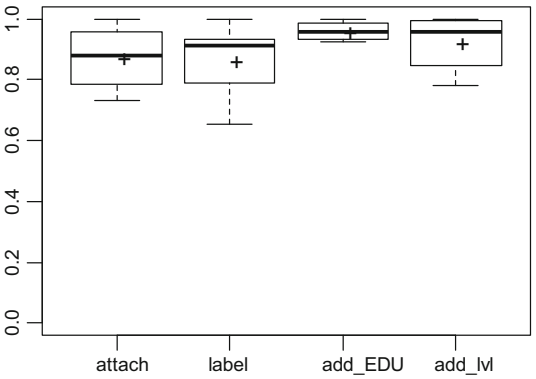

Fig. 9 Proportion of edited types on annotator RST analyses in terms of attachment change, label change, addition of EDUs or addition of a hierarchical span or multi-nuclear level to the document tree

Although the numbers here cannot be compared with the results in Marcu et al. (1999), they do show that the signal to noise ratio in student annotation of RST is reasonably high, and of course there are as yet no NLP resources to compare to as a baseline. While it is impossible to aspire to results of the quality found in the RST Discourse Treebank outlined above within one course, it should be noted that the Discourse Treebank was annotated by "more than a dozen people on a full or part-time basis over a 1 year time frame" (Carlson et al. 2001: 8), for 385 documents (about 176,000 tokens). For a project like GUM, on the order of 1/8 the size of that corpus, only a couple of weeks of course time could be spent learning and applying guidelines, so that the modest results above are at least encouraging, and will hopefully be useful.

4 Case study: characterizing coreferring lexical NPs

To illustrate the kind of studies made possible by a corpus like GUM, in this section we will look at the characteristics of full lexical NPs (non-pronouns) which have antecedents in the text. The issue of identifying coreferring lexical NPs is of considerable practical importance, since for tasks such as coreference resolution it is not obvious for a given lexical NP whether antecedents should be looked for. Whereas a phrase such as 'her' is automatically expected to have an antecedent in the text, a phrase such as 'the president' may or may not be resolvable to a particular person mentioned elsewhere in the text. It is therefore interesting to ask what the characteristics of a lexical NP are that increase the likelihood of antecedents being present, and the GUM corpus allows us to gather several sources of information for this purpose.[20]

Previous approaches have generally used definiteness, proper noun status and grammatical function (e.g. Ritz 2010; Durrett and Klein 2013; Lee et al. 2013). Intuitively, lexical NPs that do have an antecedent should be definite, though proper nouns, which are by nature definite, may or may not signal previous mention. Entity type can also be related: people are likely candidates for subsequent mentions in a

[20] In a similar vein, Recasens et al. (2013) attempt to characterize singleton phrases, i.e. phrases that *do not* have an antecedent in the text. This study is complementary to their findings (see below).

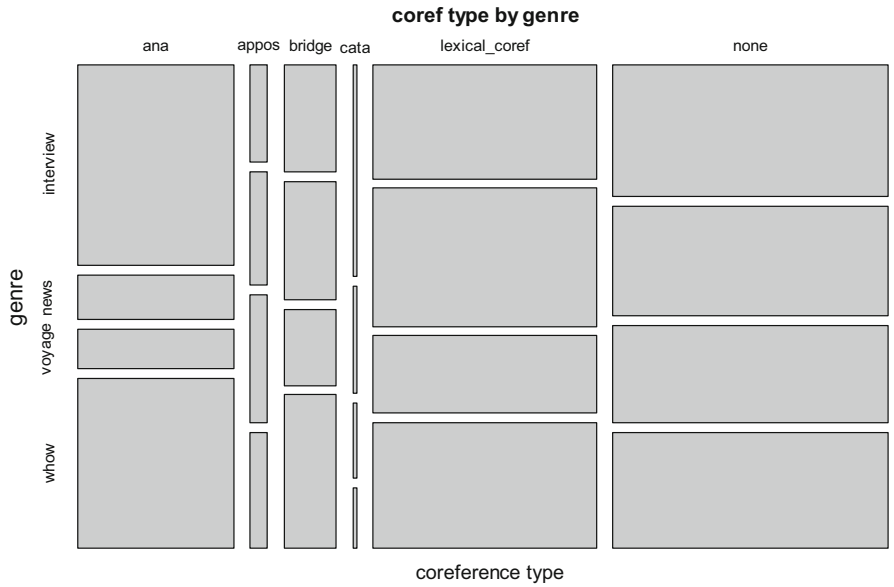

Fig. 10 Coreference type proportions across genres. Bar widths represent the relative amount of data they are based on

text, whereas dates are not as likely to be discussed repeatedly, and entity types are used by coreference resolution systems to match properties with antecedent candidates (Lee et al. 2013). Grammatical function, especially in the context of chains of coreferring subjects in a narrative chain (cf. Chambers and Jurafsky 2009), may also be relevant, though it should be noted that recurring lexical NP subjects are perhaps less likely, since we would expect them to be replaced by a pronoun before long. Other factors that are readily available in the corpus and may be relevant are the length of each referent (longer phrases contain more information and suggest entities not previously discussed), number [plural number may be less specific and somewhat less prone to coreference in context, as also found by Recasens et al. (2013)], and the rhetorical function of the clause containing the referent, which was not available to previous studies (some relations, such as 'circumstance' are likely to contain marginal or 'backgrounded' referents that will not be subsequently discussed). Position in the text may also have an influence, as most referents are expected to be introduced early. Finally, it is possible that the different genres in the corpus correlate with different patterns of coreference. Whether this is the case and to what extent is a matter that has not yet been studied in depth.

To answer the last question first, it appears that the genres behave quite differently across all coreference types, but more similarly for coreferring lexical NPs (i.e. excluding 'ana', 'cata' and 'bridge'). Figure 10 shows that the voyage and news texts cluster together in dispreferring anaphora, and that how-to guides are more likely to exhibit bridging than others. For comparison, in voyage texts appositions are more prevalent than bridging, the opposite of how-to texts. This is likely to be due at least

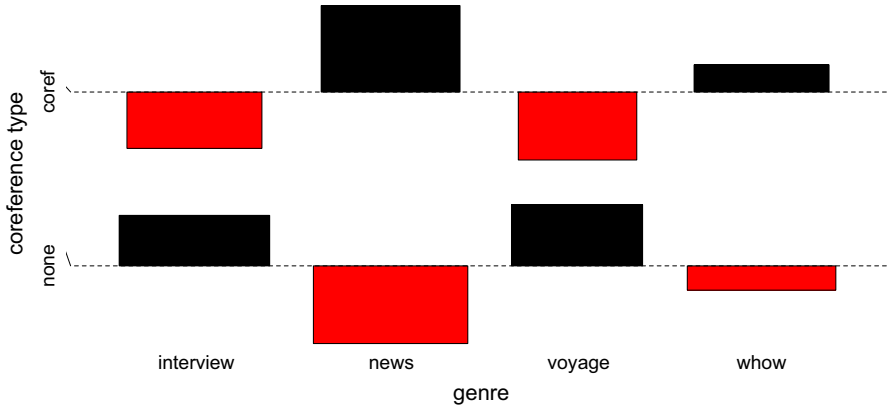

Fig. 11 Association of lexical coreference with genre

in part to different communicative functions: how-to guides are technical and contain many implied connections or part-whole relationships (ingredients, tools, etc.), whereas travel guides often supply additional information about place names using appositions.

However as the association plot in Fig. 11 shows, which only takes into account lexical coreference versus no coreference, there is more similarity between voyage and interview texts for lexical NPs, with less 'coref' than 'none' in both. The rectangles in the plot give the difference between observed and expected frequencies in each category based on an expectation of independence between the two variables (above expectation for the raised rectangles, or below for the lowered ones). From inspection of the texts the effect seems to be related to the narrower thematic focus of news and 'how-to's, both of which concentrate more on one topic (news-worthy event or topic of instruction), whereas the voyage and interview texts switch topics more frequently, leading to more, and shorter coreference chains.

It is possible to include all of the factors mentioned above in a multifactorial model predicting lexical coreference likelihood, though two caveats must be kept in mind: firstly, there is a high risk of overfitting the model to the data on account of the large amount of categories in some of the factors (grammatical functions, rhetorical relations). Secondly, model construction proceeds post hoc, after the data has been seen (and in fact inspected in great detail over the course of annotation). The first issue can be dealt with by only considering the most powerful, and intuitively plausible levels, such as not looking at all grammatical functions, but just subjecthood, which is known to be closely related to issues of salience and coreference (cf. Lee et al. 2013: 892). The second issue is more difficult: the optimism of the fitted model must be tested by cross-validation, but as far as the choice of selected factors is concerned, validation against further data completely unseen during development is desirable, and will only be possible once more data has been annotated in the same way. Nevertheless, most of the factors in question have been used before by one or more of the studies cited above, giving some independent merit to their selection.

Table 8 Mixed effects model predicting presence of an antecedent for a lexical referent

```
    AIC      BIC   logLik  deviance  Residual Degrees-of-Freedom
  4057.4   4179.7  -2008.7   4017.4     3318

Random effects:
 Groups   Name        Variance Standard Deviation
 Doc name (Intercept) 0.04444  0.2108
Number of observations: 3338, groups:  document name, 25

Fixed effects:
                              Estimate Std. Error z-value Pr(>|z|)
(Intercept)                    1.36879    0.27550   4.968 6.75e-07 ***
entity type: event            -0.64594    0.16305  -3.962 7.45e-05 ***
entity type: object           -0.68090    0.13042  -5.221 1.78e-07 ***
entity type: organization     -0.64152    0.18491  -3.469 0.000522 ***
entity type: person           -0.65285    0.13829  -4.721 2.35e-06 ***
entity type: place            -0.60493    0.14246  -4.246 2.17e-05 ***
entity type: plant            -1.63470    0.29329  -5.574 2.49e-08 ***
entity type: substance        -1.32372    0.23036  -5.746 9.12e-09 ***
entity type: time              0.72019    0.22262   3.235 0.001216 **
log(tok number in text)       -0.22414    0.03507  -6.391 1.65e-10 ***
head is a proper noun         -0.73122    0.14934  -4.896 9.76e-07 ***
head is subject               -0.62133    0.15820  -3.927 8.59e-05 ***
head is plural                 0.28769    0.09411   3.057 0.002236 **
entity is definite             0.99162    0.09674  10.251  < 2e-16 ***
rhetorically central clause -0.39553    0.13217  -2.993 0.002766 **
topical genre                 -0.31093    0.12322  -2.523 0.011623 *
log(entity length in toks)     0.61085    0.07510   8.134 4.14e-16 ***
is proper:log(ent. length)     0.46842    0.15068   3.109 0.001878 **
is subject:log(ent. length)    0.45011    0.16860   2.670 0.007593 **
---
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 8 gives a linear mixed effects model containing the factors discussed above, including significant interactions between both subjecthood and proper noun status of the referent head token with the length of the referent in words, as well as a random effect for the document in question (and by proxy, the associated annotator), generated using the lme4 package in R.

The random effect of document/annotator is responsible for a small amount of variance compared with the fixed effects' influence, suggesting that annotator bias given a fixed effect involving the genre is not problematic.

In terms of predictive power, the model correctly classifies 68.06 % of the data, a substantial improvement over a majority baseline of 55.22 % (assuming referents never have an antecedent, the more common option), or 57.8 % if we predict that all and only definite NPs are coreferential. Although this classification accuracy is not nearly sufficient for independent application, it should be kept in mind that it is rather unrealistic to expect an accurate prediction of the presence of an antecedent

based only on the properties of the referent itself (i.e. without actually looking for an antecedent in previous text), and the evaluation is quite stringent in not considering pronouns at all, which are 'the easy cases' for this task.[21] Despite the small size of the corpus, it seems that the model, which is based on 3337 individual referents out of the 5113 annotated in the corpus (including anaphors etc.), is quite stable and not substantially over-fitted: a tenfold cross validation evaluating models based on non-overlapping randomly divided 90 % slices of the data achieves a mean accuracy of 66.83 % on held out data, with a standard deviation of 3.52 %. All of the effects including the interactions remain significant in all slices. This suggests that the factors used here are truly relevant for the phenomenon in question, across genres, and that adding training data may improve the model further.

From a descriptive perspective, this model is a starting point for considering the properties of a typical coreferring, non-pronominal NP, which is more likely than not (cf. the negative coefficients in the table, which predict coreference):

- A person or concrete object, a substance or plant[22] but not a time, and less often a place.
- Quite likely headed by a proper noun.
- Often the subject of the sentence.
- Usually not plural.
- Not marked indefinite.
- Not part of a rhetorically circumstantial, conditional, concessive or antithetical clause.
- Short.
- Later in the text.
- If it is somewhat long, being subject or proper noun is less important.
- The prior likelihood of finding such referents is higher in 'narrow topic' genres such as how-to and news (less in interviews and travel guides).

These features coincide in part with the features used deterministically by Lee et al. (2013), and also with those used to predict given status by Ritz (2010) (specifically grammatical function, proper noun inclusion and definiteness), but some of them are novel (rhetorical relations, interactions of length with other factors, position in the text and genre), as is the stochastic approach relating them together in a mixed effects model classifying coreferentiality.

[21] Adding pronouns to the evaluation results in accuracy of 74.14 %, and makes the interaction between referent length and proper noun status insignificant (possible interference from the inherent shortness of pronouns). This brings results in line with the precision score (72.2) reported in Recasens et al. (2013), however in my opinion mixing evaluation on pronouns and nominals obscures the theoretical issue somewhat, except in the context of in situ evaluation within a coreference resolution system, which is the goal of Recasens et al.'s paper.

[22] The latter is admittedly due almost entirely to one document dealing with the cultivation of Basil, though the latest iteration of GUM is set to introduce a rather similar document on growing cactuses.

5 Conclusion

This paper has explored the construction, evaluation and exploitation of a multilayer corpus collected in the context of classroom teaching. Although the corpus collected in this project is still relatively small, we have seen that its quality is quite high compared to automatically annotated material, and that the depth of annotations it contains allows the study of hitherto unexplored interactions between linguistic levels of description across genres. In terms of best practices for classroom corpus collection, some lessons from this project include:

1. The positive motivational effect of allowing students to select their own text topics from within a restricted, extensible pool.
2. The utility of collaborative online interfaces such as Arborator or WebAnno which save time and avoid technical problems by letting students annotate using a Web browser. This has motivated the creation of rstWeb, a web interface for RST annotation, for the second iteration of the course.
3. The synergy between NLP and human annotation in adjudicating errors after limited training.
4. The importance of expert review by an instructor or teaching assistant of as much material as possible, in view of the remaining errors.
5. The documentation of evolving guidelines as the project progresses to secure consistency across iterations (and accordingly, the need to revise resources if guidelines are altered).

Although the overall experience of compiling the corpus has been received very positively by participants, some desiderata and negative lessons should also be pointed out. A major concern in the development of any annotation project is the refinement of annotation guidelines on test sets before the final corpus data is approached. In the limited time context of a single semester course this was not feasible, which had consequences for guideline revisions. For example, decisions about whether or not to annotate possessive determiners for referentiality and coreference, or how to handle different types of non-canonical sentences in syntax annotation were revisited, forcing a review step that is both time consuming and potentially detrimental to annotation consistency. Lessons from such a review cannot be learned from experience if new students work on the project in a subsequent semester. For future iterations the problem may therefore be alleviated only if guidelines from previous semesters are documented and updated during the course. A further very helpful resource in dealing with unanticipated complex constructions was to give students access to a search engine with manually prepared treebanks (such as constituent and dependency versions of the WSJ and Switchboard corpora), or OntoNotes for coreference. In retrospect, these resources should have been introduced earlier in the course, but were only made available towards the end of syntax annotation. Another useful technology not utilized for the first round of this project but which is being constructed in the second round is dynamic documentation of the work in an online Wiki across semesters. This should ease the documentation and discussion of problematic cases, serve as a gateway for inclusion in new versions of the guidelines and teach students

how to work with a Wiki and document their work in a sustainable way. An evaluation of the effectiveness of using Wiki-based guidelines in the classroom context remains outstanding.

The work on multi-layer corpus analysis presented in the last section is set to continue and should benefit from the addition of new data as the corpus grows, as well as new annotations. Currently a conversion adding Universal Stanford Dependencies (de Marneffe et al. 2014) to the Typed Dependencies is planned, as is the addition of a sentence mood/rough speech act layer. Students also have access to the data and are encouraged to write final papers using it both in the course in which GUM is constructed and in other courses, so that more work on and with the data is expected. The second cycle of data collection is ongoing, currently using the same genres in order to increase the variety within the existing subcorpora, but future extensions to other genres are conceivable as well. All materials have been, and will be made available online under a Creative Commons license in the hopes of promoting more work on interactions between linguistic levels, variation across genres and the development of tools for multilayer corpora.

Acknowledgments I would like to thank the participants, past, present and future, of the course LING-367 'Computational Corpus Linguistics', for their contributions to the corpus described in this paper. Special thanks are due to Dan Simonson for his help in preparing the data. For a current list of contributors and a link to the course syllabus, please see http://corpling.uis.georgetown.edu/gum. I am also grateful for very helpful suggestions from Aurelie Herbelot, Anke Lüdeling, Mark Sicoli, Manfred Stede, the editors, and three anonymous reviewers; the usual disclaimers apply.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and Speech, 34*, 351–366.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257.

Blackwell, C., & Martin, T. R. (2009). Technology, collaboration, and undergraduate research. *Digital Humanities Quarterly, 3*(1). http://digitalhumanities.org/dhq/vol/003/1/000024/000024.html.

Burnard, L., & Bauman, S. (2008). *TEI P5: Guidelines for electronic text encoding and interchange*. Technical report. http://www.tei-c.org/Guidelines/P5/.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. (2010). The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation, 44*(4), 387–419.

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of 2nd SIGDIAL workshop on discourse and dialogue, Eurospeech 2001* (pp. 1–10). Aalborg, Denmark.

Cer, D., de Marneffe, M.-C., Jurafsky, D., & Manning, C. D. (2010). Parsing to stanford dependencies: Trade-offs between speed and accuracy. In *7th International conference on language resources and evaluation (LREC 2010)* (pp. 1628–1632). Valletta, Malta.

Chambers, N., & Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP of the AFNLP* (pp. 602–610). Suntec, Singapore.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 740–750). Doha, Qatar.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing, 8*, 259–265.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of 9th international*

conference on language resources and evaluation (LREC 2014) (pp. 4585–4592). Reykjavík, Iceland.

de Marneffe, M.-C., & Manning, C. D. (2013). *Stanford typed dependencies manual*. Stanford University, Technical Report.

Dipper, S., Götze, M., & Skopeteas, S. (Eds.) (2007). Information structure in cross-linguistic corpora: annotation guidelines for phonology, morphology, syntax, semantics, and information structure. *Interdisciplinary Studies on Information Structure,* Working papers of the SFB 632, *7*.

Durrett, G., & Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2013)*. Seattle: ACL.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121). London: Longman.

Gerdes, K. (2013). Collaborative dependency annotation. In *Proceedings of the second international conference on dependency linguistics (DepLing 2013)* (pp. 88–97). Prague.

Givón, T. (Ed.). (1983). *Topic continuity in discourse. A quantitative cross-language study (Typlological Studies in Language 3)*. Amsterdam: John Benjamins.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92* (pp. 517–520). San Francisco, CA.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203–225.

Haug, D. T., Eckhoff, H. M., Majer, M., & Welo, E. (2009). Breaking down and putting back together: Analysis and synthesis of New Testament Greek. *Journal of Greek Linguistics, 9*(1), 56–92.

Hirschman, L., Robinson, P., Burger, J. D., & Vilain, M. B. (1998). *Automating coreference: The role of annotated training data*. AAAI, Technical Report SS-98-01. http://www.aaai.org/Papers/Symposia/Spring/1998/SS-98-01/SS98-01-018.pdf.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90 % solution. In *Proceedings of the human language technology conference of the NAACL, companion volume: Short Papers* (pp. 57–60). New York: Association for Computational Linguistics.

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT workshop on active learning for natural language processing* (pp. 27–35). Boulder, CO.

Hunston, S. (2008). Collection strategies and design decisions. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 154–168). Berlin: Mouton de Gruyter.

Ide, N., Baker, C., Fellbaum, C., & Passonneau, R. (2010). The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 68–73). Uppsala, Sweden.

Jha, M., Andreas, J., Thadani, K., Rosenthal, S., & McKeown, K. (2010). Corpus creation for new genres: A crowdsourced approach to PP attachment. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk* (pp. 13–20). Los Angeles, CA.

Jiang, L., Wang, Y., Hoffart, J., & Weikum, G. (2013). Crowdsourced entity markup. In *Proceedings of the 1st international workshop on crowdsourcing the semantic web* (pp. 59–68). Sydney.

Krause, T., Lüdeling, A., Odebrecht, C., & Zeldes, A. (2012). Multiple tokenizations in a Diachronic Corpus. In *Exploring ancient languages through Corpora*. Oslo.

Krause, T., & Zeldes, A. (2014). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*. http://dsh.oxfordjournals.org/content/digitalsh/early/2014/12/02/llc.fqu057.full.pdf.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day English*. Providence: Brown University Press.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics, 39*(4), 885–916.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., & Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache, 2*, 67–73.

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web (Language and computers—studies in practical linguistics 59)* (pp. 7–24). Amsterdam: Rodopi.

Lüdeling, A., Ritz, J., Stede, M., & Zeldes, A. (2016). Corpus linguistics and information structure research. In Féry, C., & Ichihara, S. (Eds.), *The Oxford handbook of information structure*. Oxford: Oxford University Press.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60). Baltimore, MD.

Marcu, D., Amorrortu, E., & Romera, M. (1999). Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL workshop towards standards and tools for discourse tagging* (pp. 48–57). College Park, MD.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics, 19*(2), 313–330.

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., Meyers, A., Brunstein, A., Ferro, L., & Sundheim, B. (2003). *ACE-2 Version 1.0*. Linguistic Data Consortium, Technical Report LDC2003T11, Philadelphia.

Nissim, M. (2006). Learning information status of discourse entities. In *Proceedings of the 2006 conference on empirical methods in natural language processing (EMNLP 2006)* (pp. 94–102). Sydney, Australia.

O'Donnell, M. (2000). RSTTool 2.4—A markup tool for rhetorical structure theory. In *Proceedings of the international natural language generation conference (INLG'2000)* (pp. 253–256). Mitzpe Ramon, Israel.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on speech and natural language, HLT '91* (pp. 357–362). Stroudsburg, PA: ACL.

Ragheb, M., & Dickinson, M. (2013). Inter-annotator Agreement for Dependency Annotation of Learner Language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 169–179). Atlanta, GA.

Recasens, M., de Marneffe, M.-C., & Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of NAACL 2013* (pp. 627–633). Atlanta, GA.

Redeker, G., Berzlánovich, I., van der Vliet, N., Bouma, G., & Egg, M. (2012). Multi-layer discourse annotation of a Dutch text corpus. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eight international conference on language resources and evaluation (LREC'12)* (pp. 2820–2825). Istanbul: ELRA.

Reppen, R. (2010). Building a corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 31–38). London: Routledge.

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., & Andreas, T. (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen*. Humboldt-Universität zu Berlin, Technical Report Version 2.01, Berlin.

Riester, A., Killmann, L., Lorenz, D., & Portz, M. (2007). *Richtlinien zur Annotation von Gegebenheit und Kontrast in Projekt A1. Draft version, November 2007*. SFB 732, University of Stuttgart, Technical Report, Stuttgart.

Ritz, J. (2010). Using tf-idf-related measures for determining the anaphoricity of noun phrases. In *Proceedings of KONVENS 2010* (pp. 85–92). Saarbrücken.

Ritz, J., Dipper, S., & Götze, M. (2008). Annotation of information structure: An evaluation across different types of texts. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the 6th international conference on language resources and evaluation (LREC-2008)* (pp. 2137–2142). Marrakech.

Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. Reykjavik: ELRA.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project* (3rd Revision). University of Pennsylvania, Technical Report.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the conference on new methods in language processing* (pp. 44–49). Manchester.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S. R., Connor, M., Bauery, J., & Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the ninth international conference on language resources and evaluation (LREC-2014)* (pp. 2897–2904). Reykjavik, Iceland.

Sinclair, J. (2004). *Trust the text*. London: Routledge.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing (EMNLP 2008)* (pp. 254–263). Honolulu, HI.

Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 455–465). Sofia, Bulgaria.

Stede, M. (2004). The Potsdam commentary corpus. In Webber, B., & Byron, D. K. (Eds.), *Proceeding of the ACL-04 workshop on discourse annotation* (pp. 96–102). Barcelona, Spain.

Stede, M. (2008). Disambiguating rhetorical structure. *Research on Language and Computation, 6*(3), 311–332.

Stede, M., & Neumann, A. (2014). Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the language resources and evaluation conference (LREC '14)* (pp. 925–929). Reykjavik.

Taboada, M., & Mann, W. C. (2006). Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies, 8*, 423–459.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., & Beck, K. (2012). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Technical Report.

Weischedel, R., Pradhan, S., Ramshaw, L., Kaufman, J., Franchini, M., El-Bachouti, M., Xue, N., Palmer, M., Hwang, J. D., Bonial, C., Choi, J., Mansouri, A., Foster, M., Hawwary, A.-A., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., & Houston, A. (2012). *OntoNotes Release 5.0*. Linguistic Data Consortium, Philadelphia, Technical Report.

Yimam, S. M., Gurevych, I., Castilho, R. Eckart de, & Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 1–6). Sofia, Bulgaria.