ISSN 2007-9737

Towards an Automatic Mark-up of Rhetorical Structure in Student Essays

Eckhard Bick

Institute of Language and Communication, University of Southern Denmark, Denmark

eckhard.bick@mail.dk

Abstract. This paper presents and discusses a discourse relation annotation scheme for the MUCH corpus of academic writing, based on Rhetorical Structure Theory (RST). The set of proposed relational tags takes into regard both distinctiveness, pedagogical needs and implementability with automatic rules. We show how a pilot grammar with 180 rules can map discourse relations between existing syntactic nodes, exploiting lower-level grammatical/treebank markup and surface clues such as connectives (e.g., conjunctions and prepositions). In an evaluation of a live run on student essays from teacher training courses, the average false positive rate across the most frequent 21 categories was 26.7% for tags and 17.1% for relation links. Performance was best for categories with a high percentage of rules using surface connectives and, for in-sentence relations, their corresponding dependency links.

Keywords. Rhetorical structure theory, discourse annotation, student essays, MUCH corpus, constraint grammar.

1 Introduction

Over the last decade, corpus linguistics has taken an interest in the quality and pedagogical aspects of academic writing. However, most studies and corpora, e.g., the American MICUSP[1] and the British BAWE[2] corpora have focused on native (L1) speakers, single text versions and lexico-grammatical aspects only (Flowerdew 2010). The Malmö-Chalmers (MUCH) Corpus of Academic Writing as a Process (Eriksson et al. 2012, Wärnsby et al. 2016) breaks new ground by targeting Swedish students' (L2) English essays and aligning drafts, teacher/peer comments and final versions. In addition, MUCH intends to widen annotation scope beyond lexico-grammatical errors to rhetorical structure theory (RST, Mann & Thompson 1988), which not only will add linguistic value to the corpus, but also represents an important step towards a consistent semi-automatic evaluation of student essays for tasks such as grading, proofing and data-driven learning.

Finally, in a process-oriented research perspective, mark-up of rhetorical structure allows a more global interpretation of editing changes made to the texts as a result of teacher or peer intervention. One of the more ambitious goals of the MUCH project, on the corpus annotation side, is therefore the introduction of discourse relations such as reason, purpose, concession, elaboration, evaluation, contrast etc. The presence of such wide-scope mark-up will present a challenge to standard corpus interfaces[3], but it should ultimately allow RST-based searches and statistics and provide an overview of how coherently students structure their essays.

The pilot version of the corpus, collected over a 3-year period, contains about 400 essays (500,000 words), but continuous additions and a planned large-scale project, where others are invited to contribute their own texts to the MUCH infrastructure, will eventually lead to a much larger

1 Michigan Corpus of Upper-Level Student Papers [http://micusp.elicorpora.info/]

2 British Academic Written English Corpus [http://www.coventry.ac.uk/research-bank/research-archive/art-design/british-academic-written-english-corpus-bawe/]

3 For visualization of search results, we envision a relational extension to the ELAN linguistic annotator [www.mpi.nl/corpus/html/elan/]

data set. Obviously, the bigger the corpus, the more difficult it becomes to perform annotation by hand, and with ongoing additions to the corpus, any infrastructure based on manual work will eventually run out of funding. As a solution we envision an automatization of the RST mark-up process, with possible post-editing of part or all of the corpus by human annotators during the project period proper.

In principle, the same annotation tool could then also be used independently to assist teachers in their evaluation work, or permit a certain degree of self-evaluation by students. However, automatic annotation of discourse has a notoriously low accuracy with standard machine learning (ML) techniques.

Thus, Forbes-Riley et al. (2016), also working on student essays, report an F-score of 31% even when distinguishing only the 4 level-one categories of the PDTB (plus relation types). As possible issues for their target data the authors cite data noise (spelling and grammatical errors) and the importance of in-domain training data. In order to circumvent these issues, we decided to use a rule-based approach rather than ML, because the former allows transparent domain adaptation with specific rules as well as context-based recognition of grammatical errors (Bick 2015).

The underlying morphosyntactic markup of the MUCH corpus is being carried out using an adapted version of the (rule-based) EngGram parser[4], using the Constraint Grammar (CG) formalism (Karlsson et al. 1995). The EngGram core is a modular system and has been shown to support extensions with higher-level grammars, e.g., for semantic roles and verb frames (Bick 2012).

We therefore decided to maintain methodological and annotational compatibility and extend the EngGram infrastructure to handle RST/discourse relations as well, linking the new annotation to lower level morphosyntactic and dependency mark-up, with named relations holding between clausal arguments. Such use of named relations has recently been introduced to the cg3 compiler (Bick & Didriksen, 2015), and our first experiments in 2014 indicated that the feature is up to the task and indeed can be used to map discourse relations.

However, given the task's semantic and wide-scope nature, automatic annotation at this level is extremely difficult, ambiguity across categories is likely to be high and accuracy bound to be considerably lower than in a low-level task such as part-of-speech tagging.

It is therefore paramount to identify a set of descriptive categories for the task that is large enough to allow meaningful distinctions, yet at the same time small enough to avoid excessive ambiguity which would make it impossible to formulate automatic rules and reduce inter-annotator agreement in a possible human post-editing phase.

2 Frameworks and Annotation Schemes

Common to all discourse analysis approaches is the need for segmentation in order to establish possible arguments for rhetorical/discourse relations. Though segmentation could be based on punctuation and trigger words alone, linguistic segmentation based on syntactic structures provides, if available, a more robust point of departure, because it allows a distinction between clausal and non-clausal on the form side, and verb arguments and free adjuncts in terms of function. This distinction is important, because discourse relations hold between entire predications, rather than between clause constituents. If it can be made to work with sufficient accuracy on student data, the MUCH Constraint Grammar morphosyntactic annotation will provide exactly those distinctions.

A second issue is the treatment of connectives. Though discourse does draw upon a certain number of explicit connectives (therefore, by contrast, according to .., first of all), relations may be implicit and lack surface connectives (about 50%, according to Pitler et al. 2008). A theory that limits itself to relations with a surface connective, though it may be easy to implement (English connectives are fairly predictive, Pitler et al. 2008), will therefore have limited coverage.

[4] The parser can be accessed on-line at http://visl.sdu.dk/visl/en/parsing/automatic/. Our add-on discourse module will be made available at the same site.

To avoid this problem, most theories allow abstract arguments consisting of bracketed token chains without a connective. Discourse arguments may be discontinuous, lists or coordinations, but are usually held together by syntactic coherence. In our Constraint Grammar approach, we exploit this syntactic coherence by tagging relation names onto argument heads. This way, there will always be a surface token to carry the tag, even without explicit connectives.

The third problem is how to establish a reasonable set of relational categories for discourse. In the absence of explicit and unambiguous connectives, too large a category set may lead to inter-annotator disagreement in human annotation, and to low precision in automatic annotation. Conversely, too small a category set may fail to capture important distinctions, restricting the theory's usefulness for pedagogical or linguistic research.

Furthermore, category usefulness is domain-dependent. Thus, spoken discourse exhibits mechanisms (such as repairs) that are absent from written discourse (and which we will therefore ignored for the time being), and scientific papers follow certain topic-organization rules not found in, e.g., news casts.

Two general types of discourse categories can be distinguished: On the one hand, logic-semantic categories such as CAUSE, CONDITION, ALTERNATIVE, on the other hand meta-discourse categories structuring the flow of discourse rather than relating its content: REPAIR, RESTATEMENT, ATTRIBUTION. Though a few categories in the second group are more typical (or even exclusive) of spoken discourse, and the first group is much more important for information extraction and QA, both category classes are relevant for essay evaluation, which is the target domain of the MUCH project. In the following subsections we will discuss three existing mark-up strategies and their choice of categories.

2.1 PENN Discourse Treebank

The PENN Discourse treebank (PDTB Research Group, 2008) adds discourse relations on top of the syntactic annotation, as discourse-level predicates with typically 2 arguments (clauses, vp's, np's, anaphora), just like our own discourse annotation in CG. The scheme distinguishes between explicit and implicit connections, alternative lexicalisations (AltLex) and simple entity-based coherence (EntRel). The first three are associated with discourse senses, comprising four groups of categories:

1 Temporal (asynchronous, precedence, succession),
2 Contingency (cause, condition),
3 Comparison (contrast, concession).
4 Expansion (conjunction, instantiation, restatement, alternative, exception, list),

For the categories in group 2 and 3, a distinction is made between non-pragmatic and pragmatic (e.g., pragmatic cause = justification).

2.2 Ädel's Metadiscourse Categories

Ädel's scheme uses 23 functional metadiscourse categories (Metatext categories, Ädel 2006) and is related to the MICUSP corpus of academic papers and the MICASE corpus of university lectures.

1 Metalinguistic comments (repairing, reformulating, exemplifying a.o.),
2 Discourse organization (topic handling, enumeration, asides, pre-/reviewing),
3 Speech act labels (arguing, exemplifying),
4 References to the audience (managing channel/discipline, message, response).

2.3 RST Treebanks

The Wall Street Journal-based RST Discourse treebank connects elementary discourse units (EDU), mostly clauses, including clausal adverbials (-ing, infinitive or participle clauses) and some phrases, especially PPs, but never clausal subjects or objects (with the exception of arguments of attribution verbs, i.e. cognitive predicates).

The mark-up scheme (Carlson & Marcu 2001) contains 78 relations (53 mononuclear and 25 multinuclear), belonging to 16 classes (attribution, background, cause, comparison, condition, contrast, elaboration, enablement, evaluation, explanation, joint, manner-means, topic-comment, summary, temporal, topic change). In addition, 3

structural relations are used: textual-organization, span and same-unit. For ambiguous cases, a preference order was used to decide on only one relation. Leaner versions of this scheme have been adopted for the Portuguese DiZer annotator (Pardo et al. 2004) and the Spanish (da Cunha et al. 2011) and Basque (Iruskieta et al. 2013) RST treebanks, as well as the multi-source Discourse Relations Reference Corpus (Taboada & Renkema 2008). A related scheme is used by the Potsdam Commentary Corpus (Stede 2004) for German.

2.4 Adopting a Scheme

There is a certain overlap between the RST and PDTB schemes. Both are relational, but with its focus on connectives, PDTB is more surface-oriented and "binary", while RST intends to build a tree structure for so-called EDU's (elementary discourse units). The third scheme (Ädel) is difficult to align to the other two, first because it is non-relational, and second, because it addresses meta-discourse rather than the logic of the discourse proper.

Therefore, even though some of Ädel's categories are equivalent to RST and PDTB categories, they mean something different. Rather than on the comment itself, for instance, focus is on the speech act of saying that this is a comment.

Both types of annotation, discourse and meta-discourse, appear relevant to the text types and intended uses of the MUCH corpus, but while Ädel's metadiscourse categories could be assigned fairly ambiguity-free and "mechanically" with just a large set of paraphrases for the individual category markers, it is linguistically and computationally more challenging to assign potentially ambiguous and underspecified relations between discourse elements.

Also, because of the meta-discourse surface markers, meta-discourse annotation should be more accessible to straight-forward machine learning (ML) techniques. What triggers a discourse relation, on the other hand, is less obvious.

Surface markers are often missing or ambiguous, and it is therefore likely that long distance context and deeper linguistic information will be necessary for the automatic treatment of discourse relations than for the treatment of meta-discourse. Furthermore, a structural annotation, be it binary or tree-based, should profit from structural annotation at lower levels (syntax), and could itself prepare the ground for other high level tasks, e.g., inference and summarization.

We therefore decided to address the more challenging discourse relation mark-up in the MUCH corpus with a Constraint Grammar approach, leaving meta-discourse annotation to a possible later ML stage. Because Constraint Grammar is a token-based approach, we suggest to link the necessary relational tags to the heads of existing syntactic constituents (first of all, clauses). Such a head with all its dependents ("descendents") will then constitute what RST calls elementary discourse units (EDU's), which makes RST a more natural framework than PDTB with its need for implicit (i.e. token-less) connectives[5].

3 Choosing a Category Set

We implemented a pilot discourse grammar in the CG framework, using example sentences from the RST corpus annotation manual for development and formulating relational CG rules for individual RST categories. Based on these experiments, we selected those categories that could be operationalized in terms of text-based linguistic clues (lemma, syntax, semantic roles, verb frames etc.)[6], ending up with a reduced CG set of 33-37[7] RST categories, for each of with we introduced a (mostly 4-letter) abbreviation tag. 11 of these are directly equivalent to adverbial semantic roles, making it possible to directly "translate" the corresponding EngGram tags (e.g., cause, condition, consequence/effect, blue in table 1). Our tag set has a substantial overlap with those cited in (Pardo et al. 2004) and (Da Cunha et al. 2011) who also use a streamlined tag set smaller than the

5 Of course, even with a PDTB category set, connectives could simply be used as names of relations, while still attaching tags to clause heads rather than the connectives themselves, avoiding the problem of missing surface tokens.

6 The presence of an overt surface connector was not a condition, all linguistic hints were considered

7 A few difficult categories are included in the grammar, but filtered back into a hypernym category in actual corpus annotation (* in the table 1).

Table 1. Category tag set

Relational tag	Category name	Relational tag	Category name
BACK*	background	**MEANS**	means
CAUS	cause	**OTHR***	otherwise
CIRC	circumstance	**PREF***	preference
COCL	conclusion	**PSOL**	problem solution
COMP	comparison	**PURP**	purpose
COMT	comment	**QA**	question answering
CONC	concession	**QUOTE**	quote/attribution
COND	condition	**REAS**	reason
CONS	consequence	**RESU**	result
COTR	contrast	**RETQ**	rhetor. question
ELAB	elaboration	**RSTA**	restatement
ENAB*	enablement	**SEQU**	sequence
EVAL	evaluation	**STAR**	statem.-response
EVID	evidence	**SUMA**	summary
EXAM*	example	**TEMP-AFT**	temporal:after
EXPL	explanation	**TEMP-BEF**	temporal:before
ITPR	interpretation	**TEMP-SAM**	temporal:same
LIST	list	**TXTO**	text organisation
MANR	manner		

English original, and differs from the former mainly by including a more fine-grained set of "adverbial" and "illocutionary" RST categories (e.g., temporal, manner and comment, statement-response). Because of this superset-subset correspondence (rather than a many-to-many correspondence), it is possible to automatically convert our CG annotation into the categories used for Spanish and Portuguese.

Another reason for not adopting all categories from the RST scheme was that many are not sufficiently disjunct for our purposes, and difficult to reliably distinguish for both human annotators and CG context rules:

- **Background** is very close to **Circumstance.** Though the latter should contain a temporal element, this needn't be visible, and background information may include time markers, too (tense, adverbs), so it would be easiest to fuse these categories (CIRC).
- **Analogy** should be subsumed under **Comparison** (COMP) because its defining criterion (correspondence in more than one respect) is difficult to operationalize.
- **Antithesis** should be fused with **Contrast** (COTR). The RST manual itself suggests to use nuclearity for the distinction (mononuclear

for Antithesis, multinuclear for Contrast), but for automatic annotation we deem nuclearity too soft a distinction.

- The RST scheme lists some **-[A-Z]** subcategories, for instance negated attribution, **Attribution-N** (e.g., *yesterday's statememt didn't say whether ...*), but negation is a semantic operator not specific to discourse relations, and might better be kept separate. Another case is **Consequence-N** and **Consequence-S,** indicating whether it is the nucleus or the satellite that is the consequence, in analogy to the Cause-Result distinction. Since our own scheme does not distinguish between nucleus and satellite, we will simply use uppercase 'CONS' for the consequence and lower case 'cons' for the underlying situation statement. Similarly, we do not distinguish between **-N** and **-S** forms for RST's categories of **Evaluation**, **Interpretation,** **Problem-Solution-N** and **Summary.**
- A category **Comment-Topic** or **Topic-Comment** is stipulated in the RST scheme, and difficult to distinguish from ordinary (subjective) **Comment** as non-subjective, but examples are close to **Explanation** or **Elaboration** (incl. Definition), so it might be an idea to drop this category.
- A distinction between **Sequence** and **Inverted-Sequence** according to chronological order is not strictly necessary for discourse annotation, and could be left to a TIME-relation parsing stage.
- **Definition** is a separate category in the RST manual, but unless there's actually a verb like "define", definitions read like elaborations, and will be treated as such (ELAB) in our CG scheme. **Example** works a bit like Definition, and could be classified as ELAB, but has so far been kept as an independent category.
- Similarly, the six RST subcategories of Elaboration, **Elaboration-Additional, Elaboration-General-Specific, Elaboration-Object-Attribute, Elaboration-Part-Whole, Elaboration-Process-Step and Elaboration-Set-Member** are just tagged as ELAB. Making these distinctions in an automatic fashion would be challenging, and is left to future research. Elaboration-Process-Step has the added problem of ARG2 being a multi-part list of satellites. In CG, this will either be seen as a coordination (and tagged as a whole), or as multiple parallel arguments.
- The **Hypothetical** seems problematic as a relation and independent category, and is logically subsumed as the parent end of COND (condition) or RESU (result).
- The RST scheme introduces a "symmetric" category for cause/result, **Cause-Result,** which we avoid as superfluous, if true ambiguity/symmetry should occur, double tagging with CAUS and RESU could be used as a fail-safe.
- In the RST scheme, the **Condition** category has a competitor, **Contingency,** for habitual/recurrent conditions or time/place contingencies *(whenever, wherever).* In practice, however, ordinary *where* or *when* can fulfill these functions, too, and the distinction is even more difficult without a connective. We therefore use **Condition** or **Temporal:same** in these cases.
- Finally, the category of **Same-Unit** is not necessary in our scheme, because CG dependency trees do not share the discontinuity problem a constituent grammar would suffer from.

It might be useful to add PDTB categories without a direct match in the RST scheme, in particular **Exception,** which often has clear surface connectives. Furthermore, PDTB categories could be used where a subdivision of RST **Textual-Organization** is desired (e.g., introducing topic, previewing, endophoric marking).

Furthermore, there is the issue of PDTB *pragmatic* versions of certain categories: Pragmatic concession, Pragmatic contrast, Justification (pragmatic reason), Relevance, Implied Assertion. Both RST and PDTB mark topic change, with Topic-Drift / Topic-Shift and Adding-Topic, respectively. However, it seems near impossible to identify surface-oriented or structural clues for these categories in automatic annotation, and bag-of-word comparisons, that would work between texts, are of less use on small chunks such as sentences or paragraphs.

4 Writing a Discourse Grammar

In our CG annotation RST tags appear in upper case for the ARG2 discourse unit, and in lower case for the ARG1 discourse unit, linked by token IDs, e.g., <REF:CONC:+10> and <REF:conc:-10>. For an RST nucleus-satellite relation, ARG2 is the satellite and ARG1 the nucleus. However, our CG annotation does not make the distinction between mono- and multi-nuclear relations. Rather, it will follow the syntactic annotation and call @ADVL constituents for ARG2 satellites[8]. With 2 main clauses, the second will be ARG2, the first ARG1.

The following CG rule, for instance, will tag a concession relation (CONC) between two main verbs (@MV) and their EDU clauses.

- ADDRELATIONS (CONC) (conc) TARGET @MV
- (*-1 ("although" KS) OR ("even=if")
- OR ("though") BARRIER @MV)
- TO (1 (*) LINK *-1 @FS-ADVL
- BARRIER NON-ICL/ADV LINK p @MV) ;

The rule's conditions are basically that the first (TARGET) main verb (@MV) should have a concessive conjunction to the left (*-1) without other verbs in between (BARRIER), that its clause function should be that of adverbial subclause (@FS-ADVL), and that the other (TO) main verb should be the dependency parent (p) of this subclause. An example annotation (word tokens with annotation tags) is shown below for the following sentence:

"Although Scotland has chosen to stick with the union, Cameron will still face political fallout over the vote."

Although [although] <clb> KS @SUB #1->4
Scotland [Scotland] <Proper> <Lcountry> N S @SUBJ> #2->4
has [have] <aux> V PR 3S @FS-ADVL> #3->14
chosen [choose] **<REF:CONC:+10>** <mv> V PCP2 AKT
@ICL-AUX< #4->3 **ID:4**
to [to] INFM @INFM #5->6
stick [stick] <mv> V INF @ICL-<ACC #6->4
with [with] PRP @<PIV #7->6
the [the] <def> ART S/P @>N #8->9
union [union] <HHorg> <def> N S @P< #9->7
, [,] PU @PU #10->0
Cameron [Cameron] <*> <Proper> <hum> N S @SUBJ>
#11->14
will [will] <aux> V PR @FS-STA #12->0
still [still] <atemp> ADV @<ADVL #13->14
face [face] **<REF:conc:-10>** <mv> V INF @ICL-AUX<
#14->12 **ID:14**
political [political] ADJ POS @>N #15->16
fallout [fallout] <event><idf> N S @<ACC #16->14
over [over] PRP @<ADVL #17->14
the [the] <def> ART S/P @>N #18->19
vote [vote] <act-s> <occ> <def> N S @P< #19->17
. [.] PU @PU #20->0

Note that the discourse-level annotation (in red) is fully integrated into the rest of the corpus mark-up. For each token ("word") there are well-defined tag fields, e.g., lemma [....], part-of-speech and morphology (upper case letters), syntactic function (@tags), dependency links (#n->m) and secondary tags such as semantic class (<...>).

Most discourse relations hold between clauses and are therefore tagged on clause heads, i.e. main verbs, but sometimes a discourse function will hold between a prepositional phrase and a main verb. In these cases, we map the relational tag on the semantic head of the pp, i.e. the argument of the prepositions, as in the QUOTE-relation below:

*[No fossils had been **found**], [according_to a NASA **representative**].*

5 Evaluation

Though the focus of this paper is on annotation design decisions such as category set and rule formalism, we have done a small pilot evaluation of the current performance of the parser, using a section of the MUCH corpus containing essays from teacher training courses (85,000 tokens). For the time being, we are interested in methodologically important performance

8 Quotes are an exception to this, with the quoting main clause constituting an ARG2.

Table 2. Category frequency and surface trigger percentage

Relation	n	surf %	Relation	n	surf %
ELAB	3178	0.7	**CIRC**	159	83.0
BACK	832	0	**QA**	115	0
COMT	646	0	**CONC**	113	52.2
QUOTE	561	82.5	**RETQ**	89	0
COORD	376	100	**MEANS**	70	100
COTR	369	96.7	**EVID**	62	100
PURP	361	(infm)	**CONS**	43	72.1
COND	221	100	**RESU**	30	46.7
REAS	214	100	**COMP**	30	100
LIST	206	(adv)	**TEMP-AFT**	21	100
TEMP-SAM	201	100			

differences across categories, rather than in absolute performance as such.

A very important methodological distinction holds between cases, where a discourse relation can be built upon overt surface markers, and where it cannot, assuming the further to be more reliable than the latter. Thus, of 7496 binary relations added in all, 40.3 % were based on rules involving conjunctions and prepositions, or 54.7 % when sentence-internal ELAB (such as relative clauses) was ignored, and even higher when including rules with adverbial lexeme triggers, e.g., LIST.

Since some categories are much more reliant on surface triggers than others (and hence safer), it is possible to use these counts to assign automatic confidence measures or to support informed decisions about selective annotation.

Table 2, containing all categories with n>10, shows, that of the larger categories, QUOTE (quote), COTR (contrast), COND (condition), REAS (reason), MEANS, EVID (evidence) and the temporal categories are the most surface-anchored. PURP (purpose) and LIST could be added, since both have fairly safe constructions, with infinitive markers and certain adverbs as surface markers, respectively.

With a rule-based approach, where part of the research goal is identifying the most operationalizable categories, it is not easy to find or create a manual gold corpus, but we still wanted to know how the individual categories perform in a live parse.

The easiest accessible measure for inspection in this setting is precision, i.e. the percentage of false positive tags and relations (tag % and rel % in table 3). For our experiment, we ran a live parse from raw text, including pos, syntax, frames, roles and, finally, discourse relations, then selecting the first 10 tagged instances of each discourse relation category.

As expected, the "surface-heavy" categories (# in table 3) had a good relation attachment (7.3 % errors compared to 28 % for other categories), because the parser could simply follow the syntactic dependency link based on conjunctions or prepositions, and some of the errors were in fact caused by syntactic parse errors. For the category tags (average 19 % vs. 35 %), the effect was less pronounced, mainly due to ambiguity issues with words such as "as" and "since".

6 Conclusion

We have presented an RST-based discourse annotation scheme for the MUCH corpus, arguing that the category set.

– should have sufficient distinctive power to be useful for linguistic and pedagogical purposes.

Table 3: Precision errors (false positives)

Relation	% cat error	% rel error	Relation	% cat error	% rel error
ELAB	10	40	**CIRC #**	40	10
BACK	40	30	**QA**	40	30
COMT	60	60	**CONC**	10	10
QUOTE #	40	0	**RETQ**	50	40
COORD #	30	0	**MEANS #**	0	30
COTR #	10	0	**EVID #**	0	0
PURP	50	40	**CONS**	40	30
COND #	20	10	**RESU**	0	0
REAS #	10	10	**COMP #**	20	20
LIST	50	0	**TEMP-AFT#**	40	0
TEMP-SAM#	0 (ambi.)	0	**average**	26.7	17.1

- should be implementable as an automatic system, without too fuzzy/many categories.
- should be compatible with, and integratable to, the Constraint Grammar approach used for lower level annotation of the corpus.
- We suggest to largely ignore meta-discourse annotation at the present stage and to focus on discourse relations between existing syntactic nodes. Relation classes should be independent of nucleus-satellite distinctions.

We have implemented and tested a first set of discourse annotation rules to run on top of the EngGram CG parser, prioritizing rules based on surface clues (connector particles such as conjunctions) and confirming our expectation that such rules have a higher precision, for both categories and relation target links, than rules trying to link predications without such clues.

Acknowledgments

The work reported here has been supported by Craafordska Stiftelsen through a pilot grant for the MUCH project. Thanks are also due to the Swedish MUCH team: Anna Wärnsby, Asko Kauppinen, Maria Wiktorsson (Malmö University) and Andreas Eriksson (Chalmers University of Technology), for their work on corpus compilation and valuable discussion feedback.

References

1 **Ädel, A. (2006).** Metadiscourse in L1 and L2. Studies in Corpus Linguistics, John Benjamins Publishing, Vol. 24.

2 **Bick, E. (2012).** Towards a semantic annotation of English television news - building and evaluating a constraint grammar FrameNet. 26th Pacific Asia Conference on Language, Information and Computation, pp. 60–69.

3 **Bick, E., Didriksen, T. (2015).** CG-3 - Beyond classical constraint grammar. 20th Nordic Conference of Computational Linguistics (NODALIA), pp. 31–39.

4 **Bick, E. (2015).** DanProof: Pedagogical spell and grammar checking for Danish. International Conference Recent Advances in Natural Language Processing (RANLP), pp. 55–62.

5 **Carlson, L., Marcu, D. (2001).** Discourse tagging reference manual. ISI Technical Report ISI-TR-545, Information Science Institute, Vol. 54.

6 **Da Cunha, I., Torres-Moreno, J. M., Sierra, G. (2011).** On the development of the RST Spanish treebank. 5th Linguistic Annotation Workshop, pp. 1–10.

7 **Eriksson, A., Finnegan, D., Kauppinen, A., Wiktorsson, M., Wärnsby, A., Withers, P. (2012).** MUCH: The Malmö University-Chalmers Corpus of academic writing as a process. 10th Teaching and Language Corpora Conference (TALC10).

8 **Flowerdew, L. (2010).** Using corpora for writing instruction. In: **O'Keeffe, A., McCarthy, M., eds.,** The Routledge Handbook of Corpus Linguistics, pp. 444–457.

9 **Forbes-Riley, K., Zhang, F., Litman, D. (2016).** Extracting PDTB discourse relations from student essays. 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 117–127. DOI: 10.18653/v1/W16-3615.

10 **Iruskieta, M. Aranzabe, M. J., de Ilarraza, A. D., Gonzalez-Dios, I., Lersundi, M., de Lacalle, O. L. (2013).** The RST Basque treebank: an online search interface to check rhetorical relations. 4th Workshop RST and Discourse Studies, pp 40–49.

11 **Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. (1995).** Constraint grammar: A language-independent system for parsing unrestricted text. Mouton de Gruyer, pp. 1–88.

12 **Mann, W. C., Thompson, S. A. (1988).** Rhetorical structure theory: Toward a functional theory of text organization. TEXT – Interdisciplinary Journal for the Study of Discourse, Vol. 8, No. 3, pp. 243–281.

13 **Marcu, D., Amorrortu, E., Romera, M. (1999).** Experiments in constructing a corpus of discourse trees. ACL Workshop on Standards and Tools for Discourse Tagging, pp. 48–57.

14 **Pardo, T. A. S., Nunes, G., V., Rino, L. H. M. (2004).** DiZer: An automatic discourse analyzer for Brazilian Portuguese. Advances in artificial Intelligence–SBIA, Lecture Notes in Computer Science, Vol. 3171, pp. 224–234. DOI: 10.1007/978-3-540-28645-5_23.

15 **The PDTB Research Group (2008).** The Penn discourse treebank 2.0 annotation manual. Technical Report IRCS-08-01. Institute for Research in Cognitive Science, University of Pennsylvania.

16 **Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., Joshi, A. K. (2008).** Easily identifiable discourse relations. Technical Reports (CIS), Report 884, Institute for Research in Cognitive Science, University of Pennsylvania.

17 **Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., Webber, B. (2008).** The penn discourse treebank 2.0. 6th International Conference on Language Resources and Evaluation (LREC), pp. 2961–2968.

18 **Stede, M. (2004).** The Potsdam commentary corpus. Workshop on Discourse Annotation, Association for Computational Linguistics, pp. 96–102.

19 **Maite, T., Renkema, M. (2008).** Discourse relations reference Corpus [Corpus]. Simon Fraser University and Tilburg University. Available from: http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

20 **Wärnsby, A., Kauppinen, A., Eriksson, A., Wiktorsson, M., Bick, E., Olsson, L. -J. (2016**). Building interdisciplinary bridges - MUCH: The Malmö University-Chalmers Corpus of academic writing as a process. In: **Olga, T., Gardner, A. C., Honkapohja, A., Chevalier, S., eds**., New Approaches to English Linguistics: Building Bridges, John Benjamins Publishing, pp. 197–211.

Article received on 18/02/2018; accepted on 11/01/2021. Corresponding author is Eckhard Bick.