Guidelines for Argumentative Annotations in Randomized Controlled Trial Abstracts

1 Introduction

Argument or Argumentation Mining is a research field which focuses on the automatic detection, classification and connection of argumentative components in natural language. Due to the increasing amount of unstructured data, especially textual data, which makes it necessary to develop novel tools and methods to process this flow of information, it has become a more popular topic of research in the recent years [2]. Starting from approaches, like simple retrieval techniques, in the legal domain and on scientific essays, the ongoing work in this area ranges now far beyond that [1]. Besides the fact, that finding the structure of natural arguments in general is a complex task, another difficulty one has to face here is the limited availability of annotated resources. Especially in the medical domain, existing corpora do not comprise the annotation of argumentative components.

Therefore, one aim of this work is to build a corpus on Randomized Controlled Trial (RCT) abstracts, which annotates the different components of arguments and their relations in this type of text. The guidelines comprise a brief definition of arguments in general, an overview over RCTs and the guidelines for the annotation of the argumentative components itself.

1.1 Definition of an Argument

The theory about argumentation goes a long way back to ancient Greek rhetoricians, logicians and philosophers. The presented work here is based on the assumption that a *statement* or *claim* is an assertion that deserves attention [4] To validate if a certain *claim* holds under specific conditions, one needs *evidence* either supporting or attacking that claim. In logical terminology the *premise* (evidence) has to be true to make the *conclusion* true. In this work the preferred terms for the argument components are *claim* and *premise*.

1.2 Randomized Controlled Trials

Randomized Controlled Trials (RCT) are a common type of experimental studies in the medical domain for evidence-based decision making. In general, to test the effect of a drug or treatment, the test subjects are divided into two groups.

One receiving the hypothesized treatment(intervention arm) and the other an established treatment (control arm). The results are then compared after certain time intervals.

The documentation of the study is defined by the CONSORT[1] policies. Therefore, each reported study has a similar structure. The abstract is structured with multiple tags, either part of the *background, objective, methods, results* or *conclusion* category. The tags will be deleted before the annotation process, but the publication policies ensure a minimum consensus of provided information, which makes the studies comparable and ideal for building a corpus.

2 Annotation of Argument Components

2.1 Claim

In the context of analyzing Randomized Controlled Trial(RCT) abstracts, a *claim* is a concluding statement made by the author about the outcome of the study. It is a general hypothesis that describes the relation of a new treatment (intervention arm) to existing treatments (control arm) and should emerge from the described results.

For illustration purpose, there are some examples explained in detail below. *Claims* are written in bold and are surrounded by square brackets marked with a subscript.

Example 2.1.1 To compare viscocanalostomy, a nonpenetrating procedure for glaucoma treatment, with trabeculectomy. Randomized controlled trial. Twenty white subjects (20 eyes) with open-angle glaucoma with no history of surgery were enrolled. Ten subjects were randomly assigned to viscocanalostomy according to Stegmann's technique and 10 subjects to a modified Cairns trabeculectomy. A complete ophthalmologic examination was performed the day before surgery and postoperatively. Further visits were scheduled monthly for 6 to 8 months after surgery. Success was defined as intraocular pressure (IOP) between 7 and 20 mmHg, with no medication. After a mean follow-up of 6 months (range, 6-8 months), success was obtained in 5 of 10 cases in the trabeculectomy group and in 0 of 10 case in the viscocanalostomy group. With Kaplan-Meier's method, subjects with viscocanalostomy showed shorter postoperative IOP-reduction periods than subjects undergoing trabeculectomy. **[According to the results of this short-term study, trabeculectomy was more effective than viscocanalostomy in lowering IOP in glaucomatous eyes of white patients]$_1$**.

Example 2.1.2 To compare outcomes of selective laser trabeculoplasty (SLT) with drug therapy for glaucoma patients in a prospective ran-

[1] http://www.consort-statement.org/

domized clinical trial. Sixty-nine patients (127 eyes) with open-angle glaucoma or ocular hypertension were randomized to SLT or medical therapy. Target intraocular pressure (IOP) was determined using the Collaborative Initial Glaucoma Treatment Study formula. Patients were treated with SLT (100 applications 360 degrees) or medical therapy (prostaglandin analog). Six visits over 1 year followed initial treatment. If target IOP range was not attained with SLT, additional SLT was the next step, or in the medical arm additional medications were added. Primary outcome: IOP; secondary: number of steps. Sixty-nine patients were treated. Data collection terminated with 54 patients reaching 9 to 12-months follow-up. Twenty-nine patients were in the SLT group, 25 patients in the medical group. Baseline mean IOP for all eyes was 24.5 mm Hg in the SLT group, 24.7 mm Hg in the medical group. Mean IOP (both eyes) at last follow-up was 18.2 mm Hg (6.3 mm Hg reduction) in the SLT arm, 17.7 mm Hg (7.0 mm Hg reduction) in the medical arm. By last follow-up, 11% of eyes received additional SLT, 27% required additional medication. There was not a statistically significant difference between the SLT and medication groups. IOP reduction was similar in both arms after 9 to 12-months follow-up. More treatment steps were necessary to maintain target IOP in the medication group, although there was not a statistically significant difference between groups. [**These results support the option of SLT as a safe and effective initial therapy in open-angle glaucoma or ocular hypertension**]$_1$.

Example 2.1.1 shows a *claim* concluding the study. It is comparing the intervention with the control arm on a general level. In this case, claiming that one has a more useful property than the other. At large, the comparison could also be that there has been found no significant difference between the two arms. This must not be confused with a concrete numerical comparison of a certain measure or any reporting of the outcomes, see example 2.3.1. Those two types of comparisons differ a lot in their role as argumentative components. Whereas the reporting of experimental observations is a fact/premise, the general comparison in solitude has no factual value and is therefore a statement/claim, which needs supporting premises in order to be credible. This can also be seen in example 2.1.2, where the reporting of the IOP reduction (dashed underline) might be confused as a *claim*, but it is a reporting of the outcome and therefor a *premise*. Also notably in those two examples is, that phrases like *"According to the results", "These results support", or "This suggests"* should be included in the claim annotation.

Example 2.1.3: Brimonidin tartrate is a highly selective alpha 2-agonist. This study investigates the safety and efficacy of 0.2% brimonidine administered twice daily for 1 year in patients with glaucoma or ocular hypertension. The study design was a multicenter, double-masked, randomized, parallel-group, active-controlled comparison clin-

ical trial. Subjects instilled 0.2% brimonidine or 0.5% timolol maleate twice daily for 12 months. Subjects were examined at baseline, week 1, and months 1, 2, 3, 6, 9, and 12. A subset of subjects was examined at week 2. Of 443 subjects enrolled in this study, 374 met the entry criteria; 186 received brimonidine and 188 received timolol. Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg. Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 ($P < .03$); no significant difference was observed between the groups for this variable at other visits throughout the 1-year course of the study. No evidence of tachyphylaxis was seen in either group. Allergy was seen in 9% of subjects treated with brimonidine. Dry mouth was more common in the brimonidine-treated group than in the timolol-treated group (33.0% vs 19.4%), but complaints of burning and stinging were more common in the timolol-treated group (41.9%) than in the brimonidine-treated patients (28.1%). Headache, fatigue, and drowsiness were similar in the 2 groups. **[In general, the tolerance to medication was acceptable]$_1$.[Brimonidine is safe and effective in lowering IOP in glaucomatous eyes]$_2$. [Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response]$_3$.**

Example 2.1.3 shows another type of *claim*, the assertion that one object of study, usually the intervention arm, has a specific property. One has to note here, that several stated qualities should be divided into multiple *claims* on a sentence level, see **claim$_2$** and **claim$_3$**.

Example 2.1.4 To compare two intraocular irrigating solutions, Balanced Salt Solution Plus (BSS Plus) versus Lactated Ringer's (Ringer), for the preservation of corneal integrity after phacoemulsification. 110 patients undergoing phacoemulsification were randomised to either BSS Plus (n=55) or Ringer (n=55) as the irrigating solution. Patients were examined at baseline and at 1, 8, 15, 30 and 60 days postoperatively. Evaluations included specular microscopy to evaluate endothelial cell density (ECD) and endothelial cell size variability (CV), and corneal pachymetry for central corneal thickness (CCT) measurement. Groups were well balanced regarding baseline ECD, CV and CCT ($p>0.05$). There was no statistically significant difference between ECD reduction in group BSS Plus 13.1 $\pm$ 2.0% and Ringer 9.2 $\pm$ 1.9% ($p< 0.05$) at day 60 or in any study visit. There was no statistically significant difference between CV increase in group BSS Plus 23.0 $\pm$ 3.0% and Ringer 20.2 $\pm$ 4.0% ($p< 0.05$) at day 60 or in any study visit. CCT was significantly increased ($p< 0.05$) at 1, 8, 15 and 30 days postoperatively, return-

ing to baseline at 60 days in both groups. There was no significant difference in CCT increase in both groups at any visit. Interestingly, there were statistically significant correlations between ECD loss and phacoemulsification time ($p< 0.0001$) and ECD loss and irrigation solution volume ($p< 0.0001$) in the Ringer group, but not in the BSS Plus group. **[Ringers solution was similar to BSS Plus for corneal preservation in atraumatic cataract surgery]$_1$. [However, our study demonstrates that there is a trend towards lower postoperative endothelial cell density for surgeries with longer phacoemulsification time and higher irrigation volumes if Ringer is used]$_2$.**

Example 2.1.5 Latanoprost, a new prostaglandin analogue, was compared with timolol for ocular hypotensive efficacy and side effects. In a multicenter, randomized, double-masked, parallel group study, 268 patients with ocular hypertension or early primary open-angle glaucoma received either 0.005% latanoprost once daily or 0.5% timolol twice daily for 6 months. All except ten patients from each group successfully completed the study. Intraocular pressure (IOP) was significantly ($P<0.001$) reduced and maintained by both medications without evidence of a long-term drift over 6 months. Comparing 6-month with baseline diurnal IOP values, the IOP reduction (mean +/- standard deviation) achieved with latanoprost (-6.7 +/- 3.4 mmHg) was significantly ($P<0.001$) greater than that produced with timolol (4.9 +/- 2.9 mmHg). Four patients treated with timolol and none treated with latanoprost were withdrawn from the study because of inadequate IOP control. Pulse rate was significantly reduced with timolol, but not with latanoprost. Slightly more conjunctival hyperemia appeared in latanoprost-treated compared with timolol-treated eyes. Fewer subjective side effects occurred in latanoprost-treated eyes. Both eyes of a patient with a characteristic, concentric iris heterochromia (darker centrally) at baseline showed a definite, photographically documented increase in pigmentation during latanoprost treatment, making the irides uniformly darker. Three additional patients treated with latanoprost were suspects for this color change. Otherwise, no significant difference between treatment groups occurred visual acuity, slit-lamp examination, blood pressure, and laboratory values. **[Latanoprost has the potential for becoming a new rst-line treatment for glaucoma]$_1$.**

As in example 2.1.4, discourse markers for contrastive relations (*however, but, etc.*) should also be included in the claim. Example (2.1.5) gives a concluding statement, that fewer side effects occurred (dashed underline). This is too underspecified to be a claim.

Example 2.1.6 To determine the analgesic effect of supplemental intracameral lidocaine 1% during phacoemulsification under topical

anaesthesia, and to assess the risk factors associated with pain. In a double-masked, randomised, clinical trial, 506 patients undergoing phacoemulsification under topical anaesthesia were randomised to receive a supplemental intracameral injection of either 0.5 cc of 1% lidocaine (277 patients, 54.7%) or balanced salt solution (BSS) (229 patients, 45.3%). Patients were interviewed by a trained interviewer using a standardised questionnaire. The main outcome measure was intraoperative pain, scored on a visual analogue scale of 0-10. Logistic regression was performed to assess ORs. 125 of 277 patients (45.1%) experienced pain in the lidocaine group, compared with 123 of 229 patients (53.7%) in the BSS group. The proportion of patients who experienced pain was significantly lower in the intracameral lidocaine group compared with the BSS group (multivariate OR 0.68, 95% CI 0.47 to 0.97; p=0.034). The median pain score (range) was 0.0 for intracameral lidocaine group compared with 1.0 for BSS group (p=0.039). Pain was more common in females (54.3% vs 43.6%; OR 1.56), non-Chinese (62.3% vs 46.9%; OR 2.13) and those who had previous cataract surgery to the fellow eye (55.3% vs 44.7%; OR 1.61). **[The use of 0.5 cc of 1% intracameral lidocaine during phacoemulsification under topical anaesthesia significantly reduces pain experienced by patients]$_1$. [Risk factors for pain include females, non-Chinese and previous cataract surgery]$_2$.**

Example 2.1.6 shows a conjunction of multiple claims. In this case it is not possible to annotated them as separate claims, contrary to example 2.1.3, because of the syntactic structure of the sentence. Such inseparable conjunctions should be annotated as one claim.

Example 2.1.7 To compare the efficacy and safety of latanoprost versus timolol in pediatric patients with glaucoma. Prospective, randomized, double-masked, 12-week, multicenter study. Individuals aged <18 years with glaucoma. Stratified by age, diagnosis, and intraocular pressure (IOP) level, subjects were randomized (1:1) to latanoprost vehicle at 8 am and latanoprost 0.005% at 8 pm or timolol 0.5% (0.25% for those aged < 3 years) twice daily (8 am, 8 pm). At baseline and weeks 1, 4, and 12, IOP and ocular safety were assessed and adverse events were recorded. Therapy was switched to open-label latanoprost pm and timolol am and pm for uncontrolled IOP. Mean IOP reduction from baseline to week 12. Latanoprost was considered noninferior to timolol if the lower limit of the 95% confidence interval (CI) of the difference was >-3 mmHg. A proportion of responders (subjects with 15% IOP reduction at weeks 4 and 12) were evaluated. Analyses were performed in diagnosis subgroups: primary congenital glaucoma (PCG) and non-PCG. In total, 137 subjects were treated (safety population; 12-18 years, n=48; 3-<

12 years, n=55; 0-< 3 years, n=34). Mean age was 8.8±5.5 years, and mean baseline IOP was 27.7±6.17 mmHg; 125 subjects completed the study, and 107 subjects were in the per protocol population. Mean IOP reductions for latanoprost and timolol at week 12 were 7.2 and 5.7 mmHg, respectively, with a difference of 1.5 mmHg (95% CI, -0.8 to 3.7; P=0.21). Responder rates were 60% for latanoprost and 52% for timolol (P=0.33). Between-treatment differences in mean IOP reduction for PCG and non-PCG subgroups were 0.6 mmHg (95% CI, -2.3 to 3.4) and 2.6 mmHg (95% CI, -0.8 to 6.1), respectively. Responder rates for latanoprost versus timolol were 50% versus 46% for the PCG group and 72% versus 57% for the non-PCG group. Both therapies were well tolerated. **[Latanoprost 0.005% is not inferior (i.e., is either more or similarly effective) to timolol and produces clinically relevant IOP reductions across pediatric patients with and without PCG]$_1$. [Both latanoprost and timolol had favorable safety profiles over the duration of this 3-month trial]$_2$.**

A special case of example 2.1.1 is example 2.1.7, the negated comparison. Like the above mentioned comparisons, this should be treated as a claim, only if it is a general statement about the experiments and not a factual report of the outcome. Furthermore, the dashed underline in example 2.1.7 is not seen as a *claim*, because it is too general, especially if you have a more concrete reformulation afterwards, see claim 2.

2.2 Major claim

Major claims as a stance of the author, as they are defined in [3], are not existent in this kind of text genre. Here *major claims* are either more a general/concluding *claim*, which is supported by other, more specific, claims or general statements about properties of treatments or diseases. In the following examples *claims* are marked as in the section above and *major claims* are marked as underlined claims.

Example 2.2.1 The purpose of this study was to establish the efficacy and safety of the Ex-PRESS (Optonol Ltd., Neve Ilan, Israel) mini glaucoma shunt in open-angle glaucoma. This was a prospective, randomized trial. Eyes from enrolled patients were randomly assigned to either Ex-PRESS implantation under a scleral flap, or trabeculectomy. The main outcome measures were: mean intraocular pressure (IOP), postoperative medication use, visual acuity, and incidence of complications. Complete success was defined as an IOP of >4 mmHg and <or=18 mmHg without the use of antiglaucoma medications. A more stringent target of IOP >4 mmHg and <or=15 mmHg was also noted. There were 78 patients (80 eyes) with primary open-angle, pseudoexfoliative, or pigmentary glaucoma enrolled in the study. A

total of 84.6% of patients receiving Ex-PRESS and 60.0% of patients receiving trabeculectomy (P=0.0230) achieved complete success. The respective proportions of patients achieving an IOP $>$ 4 mmHg and $<$ or=15 mmHg were 76.9% and 50.0% (P=0.0193). At 1-year follow-up, complete success rates were 81.8% for Ex-PRESS and 47.5% for trabeculectomy (P=0.0020), and 71.7% and 37.5% (P=0.0070), respectively, for the more stringent target. There was a similar level of postoperative interventions and complications for each group. [**In open-angle glaucoma, the Ex-PRESS mini glaucoma shunt implanted under a superficial scleral flap produces significantly higher success rates, and a similar complication rate, compared with trabeculectomy**]$_1$. [**The Ex-PRESS is a safe and effective device for treating open-angle glaucoma**]$_1$.

In the first example (2.2.1) the difference between the *major claim* and *claim* is only small. It is a conclusion/reformulation of the previous claims.

Example 2.2.2 To assess the efficacy and safety of fixed-combination latanoprost-timolol (FCLT) vs latanoprost or timolol monotherapy. This 12-week, randomized, double-masked, parallel-group study included patients with open-angle glaucoma or ocular hypertension treated with a beta-blocker and with baseline intraocular pressure (IOP) of 26 through 36 mm Hg. Following washout, eligible patients were randomized to once-daily FCLT in the evening, latanoprost in the evening, or timolol in the morning. Postbaseline IOP assessments at 8 am, 10 am, and 4 pm at weeks 2, 6, and 12; statistical superiority of FCLT for the 18 pairwise comparisons between FCLT and the 2 monotherapies, using analysis of variance. All therapies resulted in significant IOP reductions from baseline. Pairwise comparisons favored FCLT at all time points. When the 18 comparisons were tested simultaneously, FCLT was statistically superior to latanoprost at 7 of 9 time points and at all 9 time points when compared with timolol. In addition, FCLT was associated with greater percentage reductions in diurnal IOP levels and a greater likelihood of achieving lower mean diurnal IOP levels. Diurnal IOP reductions of 30% or more from baseline to week 12 were achieved by 73.5%, 57.5%, and 32.8% of those treated with FCLT, latanoprost, and timolol, respectively (P = .007 for FCLT vs timolol; P $<$ .001 for FCLT vs latanoprost). All therapies were well tolerated. [**Fixed-combination latanoprost-timolol therapy is as safe and effective in lowering IOP in patients with either ocular hypertension or glaucoma as monotherapy with latanoprost or timolol**]$_1$. [**Combination therapy can be used to treat patients for whom monotherapy does not provide sufficient IOP reduction**]$_1$.

Example 2.2.3 Many physicians recommend either brimonidine or latanoprost as firstline therapy for chronic open-angle glaucoma or oc-

ular hypertension. However, a search of MEDLINE indicates that there have been few head-to-head comparisons of the 2 monotherapies in a clinical setting. This study compared the clinical efficacy and tolerability of brimonidine 0.2% twice daily with those of latanoprost 0.005% once daily as monotherapy in patients with open-angle glaucoma or ocular hypertension. In this 3-month, multicenter, double-masked, parallel-group, 4-visit study, treatment-naive and previously treated patients with open-angle glaucoma or ocular hypertension and bilateral intraocular pressure (IOP) after washout of between 22 and 34 mm Hg were randomized to receive either brimonidine or latanoprost. Patients who had received previous treatment with either study drug were excluded from the study. The primary outcome measure was response rate, defined as the percentage of patients achieving > or = 20% reduction in IOP from baseline to month 3. Secondary outcome measures were mean IOP reduction from baseline to month 3 and clinical success, defined as the investigator's recommendation that the patient continue using the assigned study medication. A total of 127 patients (55 treatment naive) were enrolled, 66 in the brimonidine group and 61 in the latanoprost group. After 3 months of treatment, 80% of patients in the brimonidine group and 74% of patients in the latanoprost group had achieved > or = 20% reduction in IOP from baseline. The mean reduction in IOP from baseline at month 3 was 6.8 mm Hg with brimonidine and 6.5 mm Hg with latanoprost (27.8% vs 27.0%, respectively). Among treatment-naive patients, a significantly higher percentage of brimonidine-treated patients achieved >or = 20% decrease in IOP compared with latanoprost-treated patients (88% vs 59%, respectively; P = 0.01). In previously treated patients, a higher percentage of the latanoprost group achieved > or = 20% reduction in IOP compared with the brimonidine group (88% vs 74%, respectively); however, the difference was not statistically significant. Significantly more patients in the brimonidine group achieved clinical success at month 3 compared with patients in the latanoprost group (91% vs 74%; P = 0.01).[**At peak effect, brimonidine twice daily was as effective as latanoprost once daily in lowering IOP**]$_1$. [**In treatment-naive patients, latanoprost was associated with a significantly higher rate of nonresponse after 3 months of treatment compared with brimonidine**]$_2$. [**This suggests that brimonidine may be the more reliable choice for rst-line therapy of newly diagnosed open-angle glaucoma or ocular hypertension**]$_1$. In previously treated patients, however, [**latanoprost provided greater mean IOP reduction than did brimonidine**]$_3$. [**Significantly more patients achieved clinical success with brimonidine monotherapy than with latanoprost monotherapy**]$_4$.

Example 2.2.2 and 2.2.3 show the difference more clearly. The *claims* in both cases are specific statements about the conducted studies. The *major claims* are more general statements, for which the other *claims* serve as *premises*. The concluding statements do not have to occur at the end of the abstract. Example 2.2.4 shows that those statement may also occur at the beginning as an introductory *claim*. As previously mentioned, major claims can also be general statements about properties of a disease. Example 2.2.5 gives such an example. Those major claims are not necessarily connected to the other claims in the trial, but are still argumentative claims and should be therefore annotated as such.

Example 2.2.4 [**Oral carbonic anhydrase inhibitors used to treat glaucoma have significant systemic side effects**]$_1$. Brinzolamide 1.0%, a new topical ocular carbonic anhydrase inhibitor, is effective apparently without significant systemic side effects. This study was performed to establish the long-term safety and efficacy of brinzolamide 1.0% two and three times daily for primary open-angle glaucoma and ocular hypertension. An 18-month, multicenter, double-masked, parallel, controlled study was conducted. Patients were randomized to brinzolamide two or three times daily or timolol 0.5% twice daily in a 2:2:1 ratio (n = 150, 153, and 75, respectively). Intraocular pressure was measured at 8:00 AM at eligibility and months 1, 3, 6, 9, 12, 15, and 18. Efficacy was based on intraocular pressure reduction from baseline. Safety was also evaluated. All regimens produced clinically relevant and statistically significant (P<.05) intraocular pressure reductions from baseline. Mean changes in intraocular pressure trough measurements ranged from -2.7 to -3.9 mm Hg with brinzolamide twice-daily dosing and -2.8 to -3.8 mm Hg three times daily dosing compared with -4.7 to -5.6 mm Hg with timolol. The intraocular pressure reductions with brinzolamide two and three times daily were clinically and statistically equivalent. One hundred forty-four patients were discontinued from the study after randomization with the most common reasons being the occurrence of an adverse event (46), inadequate intraocular pressure control (23), patient decision unrelated to study medication (11), lost to follow-up (16), and noncompliance (9). Adverse events were nonserious and resolved without sequelae. There were no clinically relevant changes in safety parameters. [**Brinzolamide produced less ocular discomfort (burning/stinging) than timolol, and total carbonic anhydrase inhibition levels remained below that known to cause systemic side effects**]$_1$. [**Brinzolamide produced significant and equivalent reductions in intraocular pressure when dosed two and three times daily for 18 months**]$_2$. [**Brinzolamide was safe and well tolerated by patients, with minimal ocular discomfort**]$_3$.

Example 2.2.5 [**Dermatitis is a frequent adverse effect of adjuvant breast radiotherapy**]$_1$. [**It is more likely in full-breasted women and when the radiation is distributed nonhomogeneously in the breast**]$_2$. Breast intensity-modulated radiation therapy (IMRT) is a technique that ensures a more homogeneous dose distribution. A multicenter, double-blind, randomized clinical trial was performed to test if breast IMRT would reduce the rate of acute skin reaction (notably moist desquamation), decrease pain, and improve quality of life compared with standard radiotherapy using wedges. Patients were assessed each week during and up to 6 weeks after radiotherapy. A total of 358 patients were randomly assigned between July 2003 and March 2005 in two Canadian centers, and 331 were included in the analysis. Breast IMRT significantly improved the dose distribution compared with standard radiation. This translated into a lower proportion of patients experiencing moist desquamation during or up to 6 weeks after their radiation treatment; 31.2% with IMRT compared with 47.8% with standard treatment (P = .002). A multivariate analysis found the use of breast IMRT (P = .003) and smaller breast size (P < .001) were significantly associated with a decreased risk of moist desquamation. The use of IMRT did not correlate with pain and quality of life, but the presence of moist desquamation did significantly correlate with pain (P = .002) and a reduced quality of life (P = .003). [**Breast IMRT significantly reduced the occurrence of moist desquamation compared with a standard wedged technique**]$_3$. [**Moist desquamation was correlated with increased pain and reduction in the quality of life**]$_4$.

2.3 Premise

A *premise* in a RCT context is an observation or measurement in the study (ground truth), which supports or attacks another argument component, usually a *claim*. Those observations comprise side effects and the measured outcome of the intervention and control arm. It is important that they are observed facts, and therefore credible without further evidence, since this is the ground truth the argumentation is based on.

Illustrating examples are listed below. *Claims* are marked in bold as in the previous sections, whereas *premises* are underlined and written in italic and surrounded by square brackets with subscripts.

Example 2.3.1: To compare the intraocular pressure-lowering effect of latanoprost with that of dorzolamide when added to timolol. This randomized, open-label study with two parallel groups was conducted in five centers in Greece. The study enrolled 148 patients with inadequately controlled open-angle or pseudoexfoliation glaucoma (intraocular pressure of at least 22 mm Hg) or ocular hypertension

(intraocular pressure of at least 27 mm Hg) who were receiving monotherapy with a beta-blocker or dual therapy in which one of the agents was a beta-blocker. The patients were switched to timolol 0.5% twice daily for 2 to 4 weeks (run-in period) before the start of the study (baseline). At baseline, the patients were randomized to receive latanoprost 0.005% once daily or dorzolamide 2% twice daily as add-on therapy to timolol. The intraocular pressure was recorded at 9:30 AM, 12:30 PM, and 3:30 PM at baseline and at 3 months. Safety was followed throughout the study. [*The diurnal intraocular pressure reduction was significant in both groups ($P < 0.001$)*]$_1$. [*The mean intraocular pressure reduction from baseline was 32% for the latanoprost plus timolol group and 20% for the dorzolamide plus timolol group*]$_2$. [*The least square estimate of the mean diurnal intraocular pressure reduction after 3 months was -7.06 mm Hg in the latanoprost plus timolol group and -4.44 mm Hg in the dorzolamide plus timolol group ($P < 0.001$)*]$_3$. Drugs administered in both treatment groups were well tolerated. This study clearly showed that [**the additive diurnal intraocular pressure-lowering effect of latanoprost is superior to that of dorzolamide in patients treated with timolol**]$_1$.

The first example (2.3.1) shows different reports of the experimental outcomes as *premises*. Those can be results without concrete measurement values, see premise 1, or exact measured values, see premises 2 and 3. Like for *claims*, it is important to annotate different measures as multiple *premises*. The argument component boundaries might hereby differ from the by punctuation defined sentential structure, see premise 5 and 6 in example 2.3.2. As for claims, discourse markers for contrastive relations should be included in the annotation, see premise 6. In cases of comparison, it is sometimes necessary that the *premise* comprises more than one sentence, see example 2.3.3 premise 1 and 2. Here it needs the first two sentences, respectively, to complete the comparative structure.

Example 2.3.2 Brimonidin tartrate is a highly selective alpha 2-agonist. This study investigates the safety and efficacy of 0.2% brimonidine administered twice daily for 1 year in patients with glaucoma or ocular hypertension. The study design was a multicenter, double-masked, randomized, parallel-group, active-controlled comparison clinical trial. Subjects instilled 0.2% brimonidine or 0.5% timolol maleate twice daily for 12 months. Subjects were examined at baseline, week 1, and months 1, 2, 3, 6, 9, and 12. A subset of subjects was examined at week 2. Of 443 subjects enrolled in this study, 374 met the entry criteria; 186 received brimonidine and 188 received timolol. [*Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg*]$_1$. [*Brimonidine lowered*

mean peak IOP significantly more than timolol at week 2 and month 3 ($P < .03$)]$_2$; no significant difference was observed between the groups for this variable at other visits throughout the 1-year course of the study. [*No evidence of tachyphylaxis was seen in either group*]$_3$. [*Allergy was seen in 9% of subjects treated with brimonidine*]$_4$. [*Dry mouth was more common in the brimonidine-treated group than in the timolol-treated group (33.0% vs 19.4%)*]$_5$, [*but complaints of burning and stinging were more common in the timolol-treated group (41.9%) than in the brimonidine-treated patients (28.1%)*]$_6$. [*Headache, fatigue, and drowsiness were similar in the 2 groups*]$_7$. In general, the tolerance to medication was acceptable. [**Brimonidine is safe and effective in lowering IOP in glaucomatous eyes**]$_1$. [**Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response**]$_2$.

Example 2.3.3 To report a randomized clinical trial of postoperative subconjunctival injections of low-dose 5-fluorouracil in patients undergoing primary trabeculectomy. In a prospective, randomized clinical trial, 40 eyes of 40 patients were randomized to the low-dose 5-fluorouracil group and received three subconjunctival injections of 5 mg each over 11 postoperative days, and 40 eyes of 40 patients were randomized to trabeculectomy without 5-fluorouracil. [*Mean (+/-SD) preoperative and 1-year postoperative intraocular pressures in the 5-fluorouracil group were 26.9 (+/-9.5) and 15.3 (+/-5.8) mm Hg, respectively. In the control group these were 25.9 (+/-8.1) mm Hg, and 15.8 (+/-5.1) mm Hg, respectively*]$_1$. [*The patients who received 5-fluorouracil had a mean reduction in intraocular pressure of 11.5 (+/-9.1) mm Hg at a median follow-up of 52.3 weeks. The control group had a mean reduction in intraocular pressure of 10.2 (+/-8.7) mm Hg at a median follow-up of 52.6 weeks. These differences were not statistically significant.*]$_2$ [**Three postoperative subconjunctival 5-fluorouracil injections of 5 mg each after trabeculectomy in eyes at low risk for failure had no statistically or clinically significant effect on reduction of intraocular pressure with 1-year follow-up**]$_1$. Enhancement of success in this group of patients may require a larger total dose of 5-fluorouracil.

As mentioned in the introductory paragraph, the reporting of side effects are also considered as *premises*, see premises 5-7 in example 2.3.2. Furthermore, negative observations should be annotated as *premises*, see example 2.3.2 premise 3. General formulated phrases, like example 2.3.4 dashed underline, should not be annotated as *premises*.

Example 2.3.4 Latanoprost, a new prostaglandin analogue, was compared with timolol for ocular hypotensive efficacy and side effects. In

a multicenter, randomized, double-masked, parallel group study, 268 patients with ocular hypertension or early primary open-angle glaucoma received either 0.005% latanoprost once daily or 0.5% timolol twice daily for 6 months. All except ten patients from each group successfully completed the study. [*Intraocular pressure (IOP) was significantly (P<0.001) reduced and maintained by both medications without evidence of a long-term drift over 6 months*]$_1$. [*Comparing 6-month with baseline diurnal IOP values, the IOP reduction (mean +/- standard deviation) achieved with latanoprost (-6.7 +/- 3.4 mmHg) was significantly (P<0.001) greater than that produced with timolol (4.9 +/- 2.9 mmHg)*]$_2$. Four patients treated with timolol and none treated with latanoprost were withdrawn from the study because of inadequate IOP control. [*Pulse rate was significantly reduced with timolol, but not with latanoprost*]$_3$. [*Slightly more conjunctival hyperemia appeared in latanoprost-treated compared with timolol-treated eyes*]$_4$. Fewer subjective side effects occurred in latanoprost-treated eyes. [*Both eyes of a patient with a characteristic, concentric iris heterochromia (darker centrally) at baseline showed a definite, photographically documented increase in pigmentation during latanoprost treatment, making the irides uniformly darker*]$_5$. [*Three additional patients treated with latanoprost were suspects for this color change*]$_6$. Otherwise, [*no significant difference between treatment groups occurred visual acuity, slit-lamp examination, blood pressure, and laboratory values*]$_7$. **[Latanoprost has the potential for becoming a new rst-line treatment for glaucoma]$_1$.**

Important to note is also, that weaker outcomes, e.g. 'statistically not significant', are split into two annotations, if the syntactic structure allows it, see dashed premise 4/5 in example 2.3.5. Definition of terminology, see the definition of 'success' in example 2.3.6 (underlined) should not be annotated.

Example 2.3.5 Many physicians recommend either brimonidine or latanoprost as firstline therapy for chronic open-angle glaucoma or ocular hypertension. However, a search of MEDLINE indicates that there have been few head-to-head comparisons of the 2 monotherapies in a clinical setting. This study compared the clinical efficacy and tolerability of brimonidine 0.2% twice daily with those of latanoprost 0.005% once daily as monotherapy in patients with open-angle glaucoma or ocular hypertension. In this 3-month, multicenter, double-masked, parallel-group, 4-visit study, treatment-naive and previously treated patients with open-angle glaucoma or ocular hypertension and bilateral intraocular pressure (IOP) after washout of between 22 and 34 mm Hg were randomized to receive either brimonidine or latanoprost. Patients who had received previous treatment with either study drug were excluded from the study. The primary outcome measure was response rate, defined as the percentage of patients

achieving > or = 20% reduction in IOP from baseline to month 3. Secondary outcome measures were mean IOP reduction from baseline to month 3 and clinical success, defined as the investigator's recommendation that the patient continue using the assigned study medication. A total of 127 patients (55 treatment naive) were enrolled, 66 in the brimonidine group and 61 in the latanoprost group. [*After 3 months of treatment, 80% of patients in the brimonidine group and 74% of patients in the latanoprost group had achieved > or = 20% reduction in IOP from baseline*]$_1$. [*The mean reduction in IOP from baseline at month 3 was 6.8 mm Hg with brimonidine and 6.5 mm Hg with latanoprost (27.8% vs 27.0%, respectively)*]$_2$. [*Among treatment-naive patients, a significantly higher percentage of brimonidine-treated patients achieved >or = 20% decrease in IOP compared with latanoprost-treated patients (88% vs 59%, respectively; P = 0.01)*]$_3$. [*In previously treated patients, a higher percentage of the latanoprost group achieved > or = 20% reduction in IOP compared with the brimonidine group (88% vs 74%, respectively)*]$_4$; [*however, the difference was not statistically significant*]$_5$. [*Significantly more patients in the brimonidine group achieved clinical success at month 3 compared with patients in the latanoprost group (91% vs 74%; P = 0.01)*]$_6$.[**At peak effect, brimonidine twice daily was as effective as latanoprost once daily in lowering IOP**]$_1$. [**In treatment-naive patients, latanoprost was associated with a significantly higher rate of nonresponse after 3 months of treatment compared with brimonidine**]$_2$. This suggests that [**brimonidine may be the more reliable choice for rst-line therapy of newly diagnosed open-angle glaucoma or ocular hypertension**]$_1$. [*In previously treated patients, however, latanoprost provided greater mean IOP reduction than did brimonidine*]$_7$. [**Significantly more patients achieved clinical success with brimonidine monotherapy than with latanoprost monotherapy**]$_3$.

Example 2.3.6 To report the longer term results of a randomized, clinical trial comparing the 350-mm2 and the 500-mm2 Baerveldt glaucoma implants. Extended follow-up on a randomized, controlled trial. Between March 1991 and April 1993, 107 patients with uncontrolled intraocular pressure (IOP) due to non-neovascular glaucoma associated with aphakia, pseudophakia, or failed filters were randomly assigned for surgical placement of either the 350-mm2 or the 500-mm2 Baerveldt implant at the Doheny Eye Institute. A random-numbers table was used to assign each patient to one of the two groups. Preoperative IOPs and visual acuities were recorded. Clinical records were reviewed to ascertain postoperative IOPs, visual acuities, number of medications used, and implant-related complications that occurred throughout the follow-up period. Success was

defined as IOP of 6 mmHg or greater and of 21 mmHg or less in two or more consecutive follow-up visits without further glaucoma surgery or loss of light perception attributable to glaucoma. [*The overall success rates were 87% for the 350-mm2 group and 70% for the 500-mm2 group (P = 0.05)*]$_1$. Average follow-up was 37 months (range, 1-76 months) for the 350-mm2 group and 34 months (range, 5-77 months) for the 500-mm2 group. [*The life-table success rates declined over time for both implant groups, from a high of 98% for the 350-mm2 group and 92% for the 500-mm2 group at 1 year to a cumulative success rate of 79% for the 350-mm2 group and 66% for the 500-mm2 group at 5 years*]$_2$. [*Visual acuities were better or remained the same in 50% of the patients in the 350-mm2 group and 46% of those in the 500-mm2 group*]$_3$. Complications during the 5-year follow-up were also statistically similar. **[The longer term results show that the 350-mm2 Baerveldt implant is more successful than the 500-mm2 implant for overall IOP control]$_1$. [Interval comparisons indicate a higher rate of success for the 350-mm2 implant in the first, second, third, fourth, and fifth years of implantation]$_2$.** Visual acuities, implant-related complications, and average IOPs were statistically indistinguishable between the two groups.

3 Annotation of Argumentative Relations

As a next step towards modeling the arguments in the data, it is crucial to annotate the argumentative relations. Those relations are connecting argumentative components to form the graph like structure of an argument. The relation is a directed link from an outgoing node, *source*, to a *target* node. The nature of the relation can be *supporting* or *attacking*, meaning that the *source* component is justifying or undermining the *target* component. Links can occur only between certain combinations of components: *Premises* can be connected to either a *claim* or another *premise*, whereas *claims* can only point to other *claims* (including *majorclaims*). The polarity of the relation (*supporting* or *attacking*) does not limit the possibility to what type of component a component can be connected. Theoretically, all types of relations are possible between the allowed combination pairs. Practically, some relations occur rather seldom compared to the frequency of others. The number of outgoing links from a component may exceed one. Furthermore, in rare cases components can not be connected at all. This can happen for *majorclaims* in the beginning of an abstract, where their function is to point out a related problem, but which is unconnected to the outcome of the study itself.

As in the previous section, *claims* are written in **bold** and *premises* are underlined and written in *italic*. Components are enumerated with subscripts.

3.1 Attack and Partial-Attack

For *attacking* relations one has to consider the different degrees of the attack. There are *partial-attacks*, which are weakening the *target* by limiting its statement, but not completely contradicting it, and *attacks*, which are in a strong contrastive relation to the target, contradicting the *target* statement. Furthermore, a strong assumption is made w.r.t. to the significance of a measurement: we consider all measurements as statistically significant, so that statements that a result is not statistically significant can attack the observation without contradicting its existence.

3.1.1 Attack

A component is *attacking* another one, if it is (i) contradicting the proposition of the *target* component or (ii) undercutting its implicit assumption of significance, i.e. stating that the observed effects are not statistically significant.

The latter case is shown in example 3.1.1. Here, *premise 1* is attacked by *premise 2*, challenging the generalizability of the prior observation.

Example 3.1.1 [*True acupuncture was associated with 0.8 fewer hot flashes per day than sham at 6 weeks,*]$_1$ [*but the difference did not reach statistical significance (95% CI, -0.7 to 2.4; P = .3).*]$_2$

Example 3.1.2 [*At 24-month follow-up, patients in the hysterectomy group were significantly more satisfied than those in the UAE group (P = .02).*]$_1$ **[No differences were observed between groups regarding HRQOL at 24-month follow-up.]$_1$**

Example 3.1.3 [*Grade 3 or 4 myelosuppressions were more frequent with GC[...]*]$_1$ **GC is as effective as PE in terms of overall survival and progression-free survival.]$_1$**

Example 3.1.2 and 3.1.3 show the other case of *attack* relation, when an assertion is contradicted by another. Here, **claim 1** states something which is opposed by what *premise 1* states. Notably is, contrary to the previous example, that the relation is not marked with an explicit contrastive discourse marker (*but*).

3.1.2 Partial-Attack

Contrary to *attack*, the *partial-attack* is used when the *source* component is not in full contradiction, but weakening the *target* component by constraining or restricting its proposition. Those can be implicit statements about the significance of the outcome of the study, which usually occur between two claims, see example 3.1.4 and 3.1.5

Example 3.1.4 [SLN biopsy is an effective and well-tolerated procedure.]$_1$ [However, its safety should be confirmed by the results of larger randomized trials and meta-analyses.]$_2$

Example 3.1.5 **[Preoperative PFME may improve early continence and QoL outcomes after RP.]$_1$ [Further studies are needed to corroborate our results.]$_2$**

Furthermore, this can be exceptions for stated regularities, see example 3.1.6, or lowering the strength of an observed effects, as in example 3.1.7 or pointing out a trade-off, as in example 3.1.8.

Example 3.1.6 [*There were no statistically significant differences between treatments in hematological and nonhematological toxic effects,*]$_1$ [*except more patients in the thalidomide group had rash, constipation, or neuropathy.*]$_2$

Example 3.1.7 [*Swallowing and coughing problems decreased more in the TPF arm than in the PF arm at the end of cycle 2,*]$_1$ [*but to a limited extent.*]$_2$

Example 3.1.8 [*Fatigue, nausea, and vomiting worsened,*]$_1$ [*but there was a reduction in pain and no change in global QOL.*]$_2$

Similar to the examples 3.1.4 and 3.1.5, statements about the prematurely end of the study are considered as a *partial-attack* on every claim in the study, since this weakens the overall significance of the study and therefore every assertion made in this context. Example 3.1.9 is such a case. Here, *premise 1* partially attacks the **claim 1** and **claim 2**.

Example 3.1.9 [*The trial had to be stopped prematurely due to insufficient patient recruitment.*]$_1$ [...] **[The LAP approach is feasible for restorative proctocolectomy.]$_1$ [IPAA seems at least as safe as CON surgery.]$_2$**

3.2 Support

Contrary to the *attack* relations, the *supportive* relation is not further subdivided. It would have been not always unambiguous to distinguish between partially and fully supportive relations, especially w.r.t. to the impact of observed adverse drug effects. Thus, for simplicity reasons, all statements or observations justifying the proposition of the *target* component are considered as *supporting* the *target* without any further distinction in degree. Even if they are justifying only parts of the *target* component, as in example 3.2.1. Here, *premise 1* supports the first half of *claim 1*, while *premise 2* serves as evidence for the second half.

Example 3.2.1 [*Among patients with limited-stage disease, there was no evidence of a survival difference (HR for death = 0.91, 95% CI = 0.73 to 1.15).*]$_1$ [*Thalidomide was associated with an increased risk of having a thrombotic event, mainly pulmonary embolus and deep vein thrombosis (19% thalidomide vs 10% placebo; HR = 2.13, 95% CI = 1.41 to 3.20; $P < .001$).*]$_2$ **[Thalidomide in combination with chemotherapy did not improve survival of patients with SCLC but was associated with an increased risk of thrombotic events.]$_1$**

Usually, *support* relations occur mainly between claims, see example 3.2.2, and premises and claims, see example 3.2.3 and 3.2.4. Theoretically they are also possible between premises, as long as the *source* component somehow justifies the proposition of the *target* component, but this is rarely the case.

Example 3.2.2 **[The findings of this study suggest that stellate-ganglion block can provide survivors of breast cancer with relief from hot flushes and sleep dysfunction with few or no side-effects.]$_1$ [Long-term relief of symptoms has the potential to improve overall quality of life and increase compliance with anti-oestrogen medications for breast cancer.]$_2$**

Example 3.2.3 [*The SF-36 MCS and PCS, Health Utilities Index Mark 3, EuroQol 5D, and UDI scores were improved significantly in both groups at 6 months and afterward (P < .05).*]$_1$ **[Both UAE and hysterectomy improved HRQOL.]$_1$**

Example 3.2.4 [*No serious adverse events were reported.*]$_1$ **[In this small study, CDB-2914 was well-tolerated without serious adverse events.]$_1$**

Furthermore, we consider objective reports of side-effects, as *premise 1* in example 3.2.5, as supporting the evaluation of side-effects made by the author of the study. Since we do not have the medical expertise to judge the impact and gravity of such effects. On the example, we believe the author that the reported erectile dysfunction is within an acceptable range and mark therefore the relation between *premise 1* and **claim 1** as support. In cases with clear polarity markers(*'worsened'*), as in *premise 1* in example 3.1.8, the polarity of the relation should be switched to attack.

Example 3.2.5 [*281 (85%) of 332 in the combined-treatment group and 227 (72%) of 313 in the endocrine-only group had erectile dysfunction (p=0.0002).*]$_1$ **[The increase of symptoms seems acceptable and has little extra effect on quality of life.]$_1$**

References

[1] Lippi, M., and Torroni, P. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn. 16*, 2 (2016), 10:1–10:25.

[2] Peldszus, A., and Stede, M. From argument diagrams to argumentation mining in texts: A survey. *IJCINI 7*, 1 (2013), 1–31.

[3] Stab, C., and Gurevych, I. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)* (Aug. 2014), Dublin City University and Association for Computational Linguistics, pp. 1501–1510.

[4] Toulmin, S. E. *The Uses of Argument*. Cambridge University Press, 2003.