MAD: A Corpus of Multilingual Argumentative Deliberation

Eimear Maguire, Ella Schad, Jacky Visser, Chris Reed, John Lawrence
University of Dundee, UK
emaguire001@dundee.ac.uk, e.m.schad@dundee.ac.uk, j.visser@dundee.ac.uk,
c.a.reed@dundee.ac.uk, j.lawrence@dundee.ac.uk

Abstract

We present a corpus of Multilingual Argumentative Deliberation (MAD), a manually annotated corpus of deliberative dialogues in English, German, Polish and Italian. Four groups each completed two variants of a ranking task, the NASA Survival Scenario; once in their native language and once in English. The corpus is annotated using Inference Anchoring Theory (IAT), a framework developed for analysing argument in dialogical settings, and widely used in argument mining. As an argument mining resource, MAD is distinct in offering equivalent instances of spontaneous argumentation across languages. In addition to use in argument mining, the annotation captures both argument relations and dialogue acts, enabling deeper analysis of argument and dialogue structure than typical of argument-only corpora. The design of the corpus enables studies of second-language effects in English-medium interaction, cross-linguistic argument comparisons for German, Polish and Italian, and speaker dialogue strategy consistency, amongst others. The annotated MAD corpus is freely available at https://corpora.aifdb.org/mad, while we additionally release the unannotated transcripts at https://github.com/emaguire/multiling-arg-delib to facilitate repurposing of the material for other analyses.

Keywords: argument, corpus, deliberation, dialogue, multilingual

1. Introduction

The automatic extraction of reasoning structure from natural discourse – argument mining – is an area of natural language processing that has been growing for more than a decade, spurred on by the annual Argument Mining Workshop hosted at *CL events since 2014 (Lawrence and Reed, 2020). In the past couple of years, interest has grown dramatically in how reasoning is expressed in language. Large language models compete on reasoning competence, both in an academic environment (Maslej et al., 2025) and in press releases by commercial model-builders such as OpenAI and Deepseek. Gold-standard annotated data on reasoning in language has consequently become even more valuable for evaluating not only argument mining systems, but also large language models. However, this data is slow and expensive to collect: it is not a task that lends itself well to crowdsourcing, requiring highly trained annotators. Partly as a result, almost all manually annotated gold data is available only in English. Where data is available in other languages, it is completely separate. There are no examples of data where the argumentation is thematically and structurally constrained, where participants are the same, where annotation guidelines and procedures are fixed, but multiple languages are bridged. MAD fills this gap.

This paper presents a corpus of Multilingual Argumentative Deliberation (MAD): a multilingual text corpus of eight task-focused spoken dialogues manually annotated for argument. Dialogues are produced by four groups of speakers (three pairs and one trio) in which each group completes two variants of a ranking task, the NASA Survival Scenario, through discussion. Speakers are fluent second-language users of English, and each group performs a variant of the task once in their native language (German, Italian or Polish), and once in English. Through these eight dialogues, the corpus provides argument-annotated dialogue transcripts for the same task across languages. The dialogues allow comparison across argumentation as performed in three different first languages, and consequently of three distinct first languages relative to a common second language. The corpus contains a total of 15,051 words and 1,149 argument relations.

We release both the argument annotation and the un-annotated transcript. For evaluation of argument annotation quality, we report Cohen's κ = 0.449, indicating moderate agreement (Landis and Koch, 1977). The argumentation data is available at https://corpora.aifdb.org/mad, and the transcripts are additionally made available at https://github.com/emaguire/multiling-arg-delib to facilitate reuse of the data outside its primary target audience.

2. Related Work

This corpus can be positioned relative to three areas in language resource development: resources for argument mining, multilingual corpora, and multi-party dialogue corpora.

Argument corpora are an area of resource development which has seen growth mirroring that of

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6962–6978
11–16 May 2026.

the field of argument mining, and has expanded significantly in recent years.

Due to the labour-intensive nature of argument annotation, corpora of argument are often relatively small in comparison with other NLP resources, e.g. the Microtext Corpus (Peldszus and Stede, 2015) with 576 argumentative relations, the Hotel Reviews Corpus (Liu et al., 2017) with 1,201 argument components[1], and the Persuasive Essay Corpus (Stab and Gurevych, 2017) with 3,832 annotated relations. With some rare large-scale exceptions (e.g. Visser et al., 2019; Ruiz-Dolz et al., 2021b; Hautli-Janisz et al., 2022), the number of arguments annotated typically ranges from the hundreds to low thousands, making MAD a reasonably-sized resource for its domain. It is encoded in the extended Argument Interchange Format, AIF+ (Reed et al., 2008), a standardised computer-readable format for argument data. AIF+ is the format used by AIFdb[2] (Lawrence et al., 2012), the largest available database of manually annotated argument data, allowing this corpus to be easily integrated in workflows already using resources hosted there.

The corpus is also distinguished by its multilingual nature. Given the annotator training requirements for high quality manual argument annotation, most argument corpora are monolingual (and in English – but see Rocha et al., 2022 for a monolingual Portuguese example). Peldszus and Stede (2015) provide both English and German through translation and re-annotation of their original German texts. Ruiz-Dolz et al. (2021b) develop a core annotated Catalan corpus, and additionally provide machine-translated parallel English and Spanish versions. The present work provides a multilingual corpus where all argument was originally produced in the languages as presented, across extremely similar contexts. In trading (human or machine) translation-level equivalence for contextual similarity, we produce a resource which facilitates more general cross-linguistic comparison.

As a corpus of dialogue, it can be also viewed in relation to other resources for the study of interaction. One example of partially-comparable cross-lingual pair is the pairing of *Settlers of Catan* game-playing corpora: STAC (Asher et al., 23-28, 2016-05), an English-language corpus of computer-mediated sessions annotated for discourse relations in SDRT (Asher and Lascarides, 2003), and DinG (Boritchev and Amblard, 2022), a French-language corpus of transcriptions of face-to-face sessions.

The HCRC Map Task Corpus (Anderson et al., 1991) is a corpus of task-oriented dialogues between pairs of speakers in which one aims to guide the other through a route on a map, with each speaker's copy of the map showing slightly different information. Although variants of the Map Task have been recreated across a wide range of languages,[3] we are not aware of its cross-lingual replication using the same speaker set across languages.

3. Inference Anchoring Theory

The framework used as the basis for the annotation of the corpus is Inference Anchoring Theory (Reed and Budzynska, 2011; Budzynska et al., 2014), which has been used to annotate argumentation in more than 2.5 million words across fifteen languages.

Inference Anchoring Theory (IAT) is a framework for modelling argument structure, the process of argument in dialogue, and how argument structure is anchored in identifiable segments of the discourse. Its three principal elements are argument structure, discourse structure, and the illocutionary connections between them. Discourse structure is made up of locutions (discourse segments connected to argument structure) and the discourse transitions between them. IAT argument structure consists of attack, support and rephrase relations among propositions. In contrast to some approaches to argument annotation, which label propositions as *claim* or *premise*, propositions which are the conclusion of some argument and those which are the premise are not treated as different classes here. The roles of premise and conclusion are instead designated entirely by position in a support relation. This neutrality enables identification of more complex structures as found in natural language argument, recognising the role of propositions which serve as the conclusion of one argument but the premise of another. Finally, illocutionary connections anchor the argument relations and propositions to specific points in the discourse. Through this anchoring process, in addition to argument data, IAT annotation produces aids in analysis of discourse behaviour (Budzynska et al., 2016) through the identification of argumentatively relevant speech acts (van Eemeren, 2018).

From the perspective of general purpose discourse annotation schemes such as SDRT (Asher and Lascarides, 2003), the IAT argument relations of conflict, inference and rephrase can be compared to opposition, justification and explanation relations. IAT argument relations may combine to form complex argument graphs, creating structurally identifiable patterns, such as linked, serial or

[1] Each argument relation in our framework requires at least two such components, making the most comparable figure 600 or less, rather than the 1,201 components themselves.

[2] aifdb.org

[3] See https://groups.inf.ed.ac.uk/maptask.

	g1-de	g1-en	g2-de	g2-en	i1-it	i1-en	p1-pl	p1-en	Total
Argument relations									
Inference	53	32	74	86	44	86	47	60	482
Conflict	13	31	14	27	28	19	18	23	173
Rephrase	60	44	73	79	63	57	61	57	494
Total	126	107	161	192	135	162	126	140	1149
Illocutionary relations									
Asserting	148	193	229	279	210	260	188	197	1704
Arguing	53	32	74	87	42	85	47	61	481
Agreeing	24	59	51	67	59	52	43	110	465
Pure Questioning	15	13	20	36	26	21	24	30	185
Assertive Questioning	19	7	10	6	16	9	10	7	84
Restating	48	37	60	63	40	49	36	33	366
Default Illocuting	10	7	11	16	23	8	24	24	123
Disagreeing	14	33	17	27	34	20	18	23	186
Rhetorical Questioning	3	1	3	2	10	1	3	5	28
Challenging	2	0	3	0	1	3	3	1	13
Total	336	382	478	583	461	508	396	491	3635
Wordcount	1256	1432	1787	2797	1751	2366	1586	2076	15051

Table 1: Argument relations, illocutionary relations and word count per dialogue. *g1-de* indicates group G1 in German, *g1-en* indicates group G1 in English, etc.

convergent arguments, as inherited from the field of argumentation theory at large (Snoeck Henkemans, 2001).

Relative to other approaches to argument annotation, IAT annotation connects its argument components to discourse structure, enabling better understanding of argument context and creation, and ultimately producing a richer and more widely applicable dataset.

IAT annotation represents the discourse and argument structure in graph form, and is encoded in a machine-readable format using the extended Argument Interchange Format (AIF$^+$) standard (Chesñevar et al., 2006; Reed et al., 2008). See section 5.3 for an example and further details.

4. Overview of the Corpus

The corpus consists of eight dialogues. Four distinct groups of speakers each produced one dialogue in their native language, and one dialogue in English. Speakers in each group retained the same anonymised alias across both dialogues. The corpus contains 1,149 identified relations across 15,051 words for an argument density of 0.76, or 0.044 if considering only inference and conflict. The non-English languages were chosen on the basis of the availability of trained annotators fluent in those languages. Each of the four participant groups performed two variants of the NASA survival scenario, switching language between scenarios.

Word counts, argument relations and illocutionary relations per dialogue are provided in Table 1. As a measure of inter-annotator agreement, we report κ = 0.449, indicating moderate agreement between annotators.

The original NASA Moon Survival Problem (Hall and Watson, 1970, from unpublished Hall, 1963), requires participants to rank the importance of items for survival in the event of a crash-landing on the moon. Participants first create their own item ranking, and then take part in a group discussion to agree on a shared ranking of the items. The task has since been expanded with an alternative scenario, placing participants in the Jamestown colony of 1607. [4] The task facilitates argument about the relative merits of a set of options, but without requiring participants to defend a particular ideological stance as in a topic- or value-based discussion, such as a mock debate or distribution of resources among disaster survivors.

This corpus uses a modified version of the task in which pre-defined roles have been removed; participants thus had equal access to task information and no artificial deference or authority was required. Participants were also actively encouraged to rank the items collaboratively, rather than first composing their own personal rankings and discussing completed rankings in a second stage.

Frequency of timestamping in the transcript differs between English and non-English transcripts as an artefact of the transcription process. All transcripts were divided into blocks of approximately 250–300 words for distribution among annotators.

[4] The Jamestown version of the task was previously available at https://www.nasa.gov/pdf/166504main_Survival.pdf, but has been removed.

English-language transcripts are fully timestamped at the beginning per turn. Non-English transcripts have added landmark timestamps at the beginning and end of each part as divided for annotation, and typically once more within the body of the part.

URLs for the subcorpora are provided in Table 2.

5. Development

5.1. Pilot

A pilot was carried out in English using an unmodified version of the modern Moonlanding survival scenario, with members of the authors' research group acting as participants. The trial was carried out by a single group of six participants, a mixture of native English speakers and fluent non-native English speakers who were living long-term in an English-language environment. Roles were assigned to participants to manipulate their specialist expertise about item utility and hierarchical power within the interpersonal context.

Participants were well-known to each other due to being members of the same research group. Pre-existing personal knowledge affected the acceptance of arguments within the group. This was especially noticeable in references to shared preconceptions about the reasoning of fellow participants, unrelated to the arguments made within the discussion. Following this, speakers for the larger corpus were recruited who did not previously know each other, limiting possible references to pre-existing external shared assumptions. The role assignments were also removed, removing the variables of enforced authority dynamics and imbalance in domain knowledge among speakers.

5.2. Data Gathering

Participants were recruited through a general call for participation which was circulated in the staff and student newsletters of the authors' university, which is in an English-speaking region. Due to the difficulty of argument annotation, recruiting and training a new set of annotators for all possible languages was impractical (see Section 5.3 for further discussion on annotation). The languages were consequently restricted to those spoken by already-trained annotators: participants were required to be fluent bilingual speakers of English and of German, Italian or Polish. There were nine participants in total, making up three pairs (both German groups, Italian) and one trio (Polish). These groups will hereafter be referred to as G1, G2, I1 and P1 respectively. The task was carried out in person (and restricted to university staff and students), ensuring the participants were living in an English-language environment. Prior task knowledge was not screened for.

The participants gave approximately an hour of their time (half an hour per language/scenario), and were compensated with a £10 Amazon voucher. Participants were unknown to each other before the task, and both variants of the task were completed in a single session. Tasks were performed by each given group in a single session per group, with one task directly after the other. These then constitute two dialogues per group: G1-DE as German Group 1 German-medium dialogue, G1-EN as German Group 1 English-medium dialogue, I1-EN as Italian Group English-medium dialogue, etc.

The task took place in person with audio and video recording. Groups performed the Jamestown scenario in a first dialogue and then Moonlanding scenario in a second dialogue directly afterwards. At the beginning of each group session, the facilitator distributed the written instructions for the first task version and a pen for participants to record their final ranking. They indicated a 30-minute time limit per task version, and stated their availability throughout the session to answer clarification questions. Clarification questions were rare but included whether the ranking was to be attempted 'in character' (i.e. without a modern understanding of medicine in the Jamestown context).

A 10-minute verbal warning was provided in cases where the task had not been finished by the 20 minute mark: all variants of all tasks were completed within the 30-minute time limit. The average length of a dialogue was 16 minutes. When the first task variant was indicated as complete by the participants, the facilitator managed switchover to the other variant in the alternative language, collecting the prior instructions and distributing the next.

Task repetition was not expected to affect the level of discussion required for ranking in the second dialogue, as the items in each task variant were distinct. Furthermore, one scenario focused on survivability in a fixed location, while the other required successful travel to a safe location, changing priorities in item ranking. Developing bias towards particular items, and so reducing necessity of discussion in the second dialogue per group, was therefore not expected. The groups G2 and P1 had approximately a minute difference in their completion of the first and second iteration of the task, while G1 and I1 spent 3-5 minutes longer on their second iteration of the task.

Audio of the sessions was sent to professional transcription services for transcription. No further anonymisation of the transcripts was performed, as no personally identifying information was found in the text. The text of these transcripts was used as the basis for annotation.

Subcorpus	URL
g1-de	https://corpora.aifdb.org/moonlandingde1
g1-en	https://corpora.aifdb.org/enjamestownde1
g2-de	https://corpora.aifdb.org/moonlandingde2
g2-en	https://corpora.aifdb.org/enjamestownde2
i1-en	https://corpora.aifdb.org/enmoonlandingit
i1-it	https://corpora.aifdb.org/jamestownit
p1-pl	https://corpora.aifdb.org/moonlandingpl
p1-en	https://corpora.aifdb.org/enjamestownpl

Table 2: URLs for the sub-corpus corresponding to each language and group. The complete corpus can be accessed at https://corpora.aifdb.org/mad.

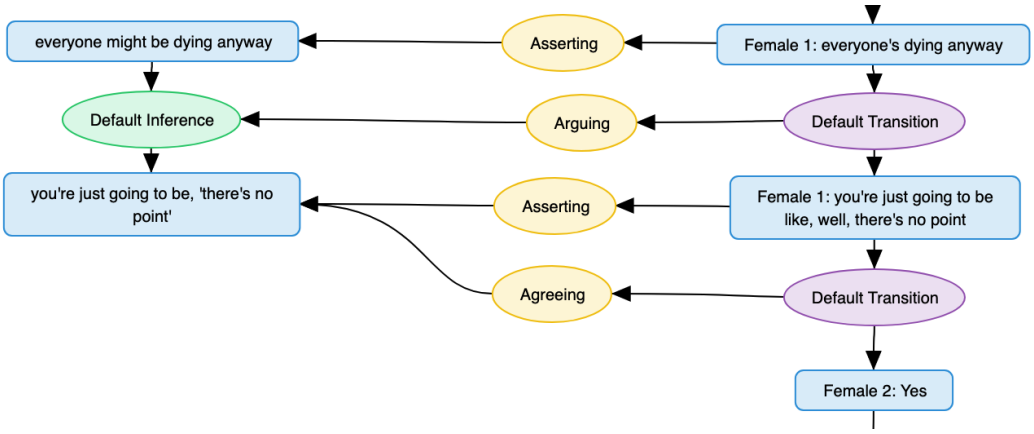

Figure 1: Example of corpus annotation from the OVA annotation tool, taken from dialogue G1 EN.

5.3. Argument Annotation

Prior to annotation, the transcript was manually divided into parts of approximately 150–300 words, with effort made to compromise between possible annotator fatigue and providing segments long enough to likely contain several instances of argument, while also avoiding splits during what appeared to be highly interconnected sequences. The non-English transcripts had not been timestamped by the transcription service, and time stamps were added to the beginning and end of the sections.

Annotation was performed by 14 members of a pre-existing annotation team. The team consisted of a number of annotators who had been previously trained in IAT annotation and who had contributed to the creation of other resources. Of the annotators who annotated the non-English portions of the corpus, the majority annotated material in their first language: 4 out of 5 annotators of German, 1 out of 2 annotators of Italian and 4 out of 4 annotators of Polish. Of the annotators who annotated the English material, 3 out of 8 were L1 English-speakers. English was the default language of the annotation team interaction and all L2 speakers had a high level of English. Annotators were paid at a rate of £11 per hour, above local minimum wage at time of annotation.

The annotation guidelines were not domain-specific, and had previously been used by this annotation team for material in other genres. They are included as an Appendix. Annotation was distributed to members of the annotation team who were fluent in the necessary languages. Annotators worked simultaneously in 3-hour online sessions, which allowed them to raise clarification questions and identify issues to an annotation session lead, who was themselves an experienced annotator contributing to the annotation. This session lead also facilitated the original recording sessions, but no active participant in the task had any interaction with the annotation process.

OVA+[5] an open-source online tool for IAT analysis, was used to carry out the annotation (Janier et al., 2014). Figure 1 shows an example of IAT annotation in OVA. The right-hand side shows the segmentation of the discourse into locutions. Each locution contains the text of an Argumentative Discourse Unit (e.g. *Yes*) and a speaker (e.g. *Female 2*): what was said and by whom. Individual locutions are connected by discourse transitions. The left-hand side shows propositions of the argument structure introduced to the context by the discourse and any inferential relations between them. In this case two propositions are identified and linked by a *Default Inference*, a support relation. This ar-

[5] ova.arg.tech

gument structure is anchored in the discourse by illocutionary force, shown between the two. In the example, the support relation is anchored by the illocutionary force of *Arguing* to the transition between two locutions: (i) the locution anchoring the premise proposition *everyone might be dying anyway*, and (ii) the subsequent locution anchoring the conclusion proposition.

Annotation quality was supported by a process of peer review within the annotation workflow itself. The review process described here was followed for this corpus, but was also established procedure for prior work by this annotation team. In the annotation workflow, every section (as divided before annotation) was annotated by an initial annotator, and then passed to a second annotator for critical review. The reviewing annotator examines the original annotation, and if they identify an error or potential error (ranging from typo to potentially incorrect argument analysis) this is discussed with the original annotator and edited according to mutual agreement. No individual section is considered annotated until it has passed this peer review. The annotation of each section is thus agreed on by an annotator-reviewer pair.

Inter-annotator agreement was calculated when the complete corpus was annotated. Annotated sections corresponding to a 10% sample of the corpus (based on wordcount) were redistributed to pairs of annotators other than the sections' original annotator-reviewer pairs. The new annotator-review pair then repeated the annotation-review process. The original and new annotations of these parts were used to calculate inter-annotator agreement as κ. Sections were selected to total a word count of 1,549, a $> 10\%$ sample of the overall 15,051.

6. Discussion

This corpus exists firstly as a resource for the computational argument community. It provides multilingual argument data in a consistent dialogical context and genre. Multilingual argument corpora typically achieve cross-linguistic coverage by identifying similar sources in different languages, or through the translation (manual or automatic) of material from an original language. The resulting similarity of the material in this corpus allows argument mining approaches to be tested across highly comparable data in multiple languages. Fine-tuning, few-shot learning, and prompting with foundational LLMs have become increasingly prevalent approaches in NLP, including in argument research (Jurkschat et al., 2022; Ruiz-Dolz et al., 2021a). Within this context, the current resource provides argumentative data across four languages.

MAD also provides opportunity for small-scale case studies in language differences within a restricted context and second-language effects, and individual speaker differences.

Some immediate observations on L1-L2 differences can be made from the argument relations as reported in Table 1. Three out of four groups display higher or near-equal rates of inference over rephrase in the English-language variant of the task. This does not appear to be an artefact of first vs. second attempt at the task, as the exceptional group is G1, not I1 (I1 being the only group to use non-English for their first dialogue). In QT30, the largest IAT-annotated corpus (Hautli-Janisz et al., 2022), speakers typically display the reverse, producing relatively higher rates of rephrase. The QT30 corpus is a corpus of broadcast political debate. A difference due to genre is expected, but the L1 task variants shadow the QT30 pattern, and skew from near-equal to rephrase-dominant. This suggests a potential greater deliberateness in L2 among participants, with a reduced tendency to gradually develop a stance through elaboration, and rather to make arguments directly. Other avenues for analysis include relative complexity of argument across L1 and L2, or tendency toward particular complex argument structures such as serial or linked argument.

The primary purpose of this corpus is to provide an argumentation dataset available in the Argument Interchange Format. The additional release of the unannotated text of the dialogues allows alternative processing outside the argumentative discourse units, which exclude turns spent on communication management. While the English-language section of the corpus is all second-language, three-way comparison of L1-medium dialogues is possible between German, Polish and Italian, and leveraging not only the annotated features of argument and discourse structure, but lexical analysis. One possible path includes cross-lingual use of discourse markers in disagreement and persuasion across similar contexts (for a text-based English case, see e.g. Musi (2018), a study into lexical markers of concession in online persuasion).

7. Argumentation Across Languages

7.1. Background

There is evidence of differences in argumentation between speakers of different linguistic and cultural backgrounds, bleeding across to differences in how argumentation is constructed, used, and interpreted by those speakers when communicating in other languages (see, e.g., Rocci (2006); Fourcade et al. (2009)).

Liu and Furneaux (2014) control for effects caused merely by working in a second language

by considering first-language writing in addition to target language writing. Comparing the presence and position of thesis statements and topic sentences in essay-writing by British and Chinese students, the authors find differences between the English-language argumentation in essays by the two groups. They further find that the characteristics of English-language essays by Chinese students mirror the argumentation of a third group, Chinese students writing Chinese-language essays, indicating a transfer from argumentation as practised in L1 to L2 argumentation.

The anthropologist Hall proposed a continuum between 'high-context' and 'low-context' cultures (1976), reflected in communication by high-context cultures relying more on indirect understanding, and low-context cultures being more verbally explicit. Using Hall's framework, Copeland and Griggs (1985) rank a number of countries from high to low context cultures. Chen et al. (2022, 2025) subsequently apply argument mining to aid in cross-cultural comparison at scale, focusing on written English argumentation by second language learners. They compare results between groups on the basis of whether their first language is associated with a high-context or low-context culture. The comparison has three aspects. Firstly, they compare proficiency through overall quality of essay organisation and argument strength. Secondly, small-scale structure is examined by identifying recurring 'flow' patterns of *premises* and *claims*, individual components which are automatically identified and classified within the text. Finally, they draw on the work of (Suzuki, 2010) for higher-level structural comparisons, adapting measures of climactic vs. anti-climactic and horizontal vs. vertical structures. Whether an essay is climactic or anti-climactic is determined based on the position of the automatically-identified *major claim* within the essay, and is classified as horizontal or vertical on the basis of whether more than half of paragraphs contained claims which were supported by premises.

7.2. Approach

Within Copeland and Griggs' ranking, Italy and Poland are identified as high-context cultures, and Germany is identified as low. We use the current corpus as the basis for a sample analysis of differences in argumentation between speakers of different language backgrounds in their own language. In contrast to the work discussed above, we examine spontaneous dialogical material rather than written work.

Suzuki (2010) examines five features: two features at the macro-structure level, a general measure of argument length, and two features at the micro-structure level. The granularity of annotation in the current corpus allows for the three features below macro-structure level to be encoded and measured. These are features which cannot be analysed on the basis of identification of Argumenative Discourse Units alone, as they require fine-grained identification of the structure. An argument as discussed here is not only an individual pair of propositions connected by an inference (two I-nodes connected through edges to/from a single RA-node), the minimal possible argument, but a connected argument structure, which may include multiple inferences (such as multiple reasons in support of some proposition, or further support for a supporting reason).

A comparison of climactic vs. anti-climactic macro-structure is not applicable to this dataset for multiple reasons. The task itself is an ongoing deliberation which leads to a final outcome. This necessitates the conclusion being available only at the end of the discourse, making a comparison of placement trivial. We also do not recognise *major claim* as a distinct proposition role within the data: see Lawrence and Reed (2020) for a discussion on the flaws in the notion of a major claim. We instead report on order of presentation at the micro-structure level, identifying the prevalence of arguments in which the conclusion is presented before the supporting premises.

The measurement of horizontal vs. vertical macro-structure does not directly transfer to this dataset. We do not consider unsupported statements to be claims in an argumentative sense: unsupported 'claims' are not annotated as not distinct from other non-argumentative text in this dataset. However, Chen et al. (2025) consider an essay to be vertical if the majority of its paragraphs include claims supported by premises. We mirror this, adapting the notion to dialogue by identifying verticality with the percentage of turns which contain both premise and conclusion of the same inferential support relation.

Argument length is measured by the number of 'argument units' within a single argument. We operationalise this as the number of propositions within a discrete argument structure. We will refer to this as *argument size* to avoid confusion with the depth/length of chains of serial argument.

We measure the use of serial and compound micro-reasoning structures on the basis of the number of inferences which are part of these structure types. We operationalise the measurement of serial reasoning as the number of inferences which form a part of a chain of serial argument. 'Compound' reasoning is defined in Suzuki (2010) as a structure in which "a statement is directly supported by two or more reasons", a structure which is usually described in the field of argumentation as *convergent* argument. We operationalise the measurement of convergent argument as the number

	g1-de	g1-en	g2-de	g2-en	i1-it	i1-en	p1-pl	p1-en
context	l	l	l	l	h	h	h	h
lang.	1	2	1	2	1	2	1	2
% concl. first	0.547	0.562	0.486	0.453	0.636	0.442	0.468	0.500
verticality	0.140	0.092	0.160	0.197	0.085	0.195	0.119	0.110
avg. size	3.385	2.222	3.214	2.415	2.314	2.406	2.531	2.444
% serial	0.566	0.188	0.459	0.326	0.205	0.302	0.298	0.417
avg. depth	1.475	1.103	1.355	1.203	1.122	1.208	1.220	1.280
% conv.	0.226	0.125	0.189	0.116	0.159	0.070	0.128	0.033
avg. breadth	1.340	1.100	1.388	1.148	1.175	1.084	1.136	1.119

Table 3: Results for each dialogue, marking whether L1 is classed as being spoken in a high-context or low-context culture, and whether the dialogue took place in L1 or L2: (i) Percentage of inferences in which the conclusion is presented first; (ii) Percentage of turns which contain an inferentially connected premise and conclusion; (iii) Average number of propositions connected through inference per argument; (iv) Number and percentage of inferences which form part of a serial argument, average depth of argument; (v) Number and percentage of inferences which form part of a convergent argument, average breadth of argument.

of inferences which support a claim which has at least one other supporting reason.

7.3. Results

Findings are presented in Table 3. Each dialogue is additionally marked as being associated with a low- or high-context group (based on L1), and taking place in L1 or L2.

Preference in order of presentation is expressed as the prevalence of inferences where the conclusion is presented in the discourse before any of its supporting premises, operationalised as the percentage of RA-nodes in which the conclusion proposition is anchored at an earlier point in the discourse than any premise proposition. Order of presentation is mixed for all groups, with no clear pattern based on context group or use of L1 or L2 in the dialogue.

Verticality, the percentage of turns which contain both premise and conclusion of an argument, is likewise mixed, without a clear pattern. This measure may be strongly affected by the change in genre to spontaneous spoken interaction. A high number of short discourse-management, primarily phatic, or confirmatory turns can be produced in a relatively brief exchange without this reflecting a high volume of unjustified claims, in contrast to a high number of paragraphs in an argumentative essay that lack a complete internal argument.

Average argument size is the average number of propositions per argument: note that as each argument must contain at least two propositions, two is the minimum size. Argument size is expected to be generally higher in low-context cultures, where steps in the inferential process are expected to be made explicit. Conversely, it is expected to be smaller in high-context cultures, where fewer or no reasons may be explicitly given. The low-context groups G1 and G2 produce larger arguments in their L1 German (3.385 and 3.214) than the high-context groups in their respective L1s of Italian and Polish (2.314 and 2.531). Argument size for G1 and G2 reduces to similar levels as the high-context groups in English-medium dialogues, despite English also being considered low-context.

Identification of serial and convergent argument provides a more granular look at differences in argument size. We report the percentage of inferences which form part of a serial argument chain or which contribute directly to a convergent argument, along with the average depth and breadth for arguments in the dialogue. Low-context cultures are expected to use more serial and convergent arguments, providing deeper and broader explicit justifications. The low-context groups G1 and G2 when speaking in their L1 show the highest concentration of both serial argument and convergent argument of any groups (0.566 and 0.459), also reflected in these groups having the longest average argument depth (1.475 and 1.355) and widest average argument breadth (1.340 and 1.388). The greater argument size found in these dialogues is driven by a higher prevalence of both serial and convergent argument structures.

Overall, the low-context groups G1 and G2 display the expected features in their argument construction when speaking in their L1. Perhaps surprisingly however, they are not distinct from the high-context groups once speaking in English, despite English also being considered low context.

8. Conclusion

This paper presented MAD, a corpus of argumentatively annotated deliberation, and to our knowledge unique among argumentation corpora in its multi-

lingual nature. As a resource, MAD provides argument data in dialogue across multiple instances of a single context and activity, and provides a novel multilingual resource offering comparable argument annotation across languages.

As an illustration, we have provided a brief analysis of argument structure features based on Suzuki (2010), comparing argument size and aspects of argument structure across languages in the corpus. The paired corpus design and the detailed annotation framework also encourage its usage beyond the argument mining community for the study of argument, the study of inter-speaker differences, and the study of cross-lingual effects in second language speakers.

9. Limitations

The corpus is modest in size, reflecting the high cost and expertise required for manual IAT annotation. The language selection would ideally be more diverse and include non-European languages. By its nature it is also unbalanced, with more material in the shared second language than in the three first languages. More comprehensive relevant participant information could have been collected at the time of data collection: only gender was recorded. Although more information on language background (other languages spoken, history with learning or speaking English, length of time spent living in an English-language environment) was requested post hoc, not all participants provided this information. A formal record of participant age range should also have been made. Deliberation about these NASA tasks may not be representative of argumentative and dialogical behaviour in other contexts.

10. Acknowledgements

This work has been supported in part by: Volkswagen Stiftung under grant 98 543, Deliberation Laboratory; the 'AI for Citizen Intelligence Coaching against Disinformation (TITAN)' project, funded by the EU Horizon 2020 research and innovation programme under grant agreement 101070658, and by UK Research and innovation under the UK government's Horizon funding guarantee grant numbers 10040483 and 10055990; the 'CLARUS' project, funded by the EU Horizon Europe Framework Programme (HORIZON) under grant agreement 101121182; and the 'AI4Deliberation' project, funded by the European Union under grant agreement 101178806. The authors would like to acknowledge the contribution of all consortium partners to formulate the project. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

11. Bibliographical References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 23-28, 2016-05. Discourse structure and dialogue acts in multiparty dialogue: The STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge university press, Cambridge.

Maria Boritchev and Maxime Amblard. 2022. A multi-party dialogue ressource in French. In *Proceedings of the Language Resources and Evaluation Conference*, pages 814–823, Marseille, France. European Language Resources Association.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument*, pages 185–196.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint-Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

Mei-Hua Chen, Wei-Fan Chen, Garima Mudgal, and Henning Wachsmuth. 2025. Cross-Cultural Comparison of Argument Structures Among English Learners: Argument Proficiency, Patterns, and Communication Styles. *Argumentation*.

Wei-Fan Chen, Mei-Hua Chen, Garima Mudgal, and Henning Wachsmuth. 2022. Analyzing culture-specific argument structures in learner essays. In *Proceedings of the 9th Workshop on Argument Mining*, pages 51–61, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(4):293–316.

Lennie Copeland and Lewis Griggs. 1985. *Going International: How to Make Friends and Deal Effectively in the Global Marketplace*, 1st ed edition. Random House, New York.

Sayaka Fourcade, Michael Hazen, and Narahiko Inoue. 2009. Cross-cultural responses to arguments in high and low context messages. *Gengo Kagaku / Linguistic Science*, 44:33–43.

Edward T. Hall. 1976. *Beyond Culture*. Anchor/Doubleday.

Jay Hall. 1963. *The Rejection of Deviates as a Function of Threat*. Ph.D. thesis, University of Texas.

Jay Hall and W. H. Watson. 1970. The Effects of a Normative Intervention on Group Decision-Making Performance. *Human Relations*, 23(4):299–317.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A Corpus of Argument and Conflict in Broadcast Debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. OVA+: An argument analysis interface. In *Computational Models of Argument*, Frontiers in Artificial Intelligence and Applications, pages 463–464, Netherlands. IOS Press.

Lena Jurkschat, Gregor Wiedemann, Maximilian Heinrich, Mattes Ruckdeschel, and Sunna Torge. 2022. Few-shot learning for argument aspects of the nuclear energy debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 663–672, Marseille, France. European Language Resources Association.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. 2012. AIFdb: Infrastructure for the Argument Web. In *Frontiers in Artificial Intelligence and Applications*. IOS Press.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using Argument-based Features to Predict and Analyse Review Helpfulness. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1363, Copenhagen, Denmark. Association for Computational Linguistics.

Xinghua Liu and Clare Furneaux. 2014. A multidimensional comparison of discourse organization in English and Chinese university students' argumentative writing. *International Journal of Applied Linguistics*, 24(1):74–96.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. The AI Index 2025 Annual Report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

Elena Musi. 2018. How did you change my view? a corpus-based study of concessions' argumentative role. *Discourse Studies*, 20(2):270–288.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action. Proceedings of the 1st European Conference on Argumentation*, volume II, pages 801–816, Lisbon. College Publications.

Chris Reed and Katarzyna Budzynska. 2011. How dialogues create arguments. In *Proceedings of the 7th Conference of the International Society for the Study of Argumentation*, Amsterdam. SicSat.

Chris Reed, Simon Wells, Joseph Devereux, and Glenn Rowe. 2008. AIF: Dialogue in the argument interchange format. In *Computational Models of Argument*, number 172 in Frontiers in Artificial Intelligence and Applications, pages 311–323, Netherlands. IOS Press.

Andrea Rocci. 2006. Pragmatic inference and argumentation in intercultural communication. *Intercultural Pragmatics*, 3(4):409–442.

Gil Rocha, Luís Trigo, Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, Bruno Martins, and Miguel Won. 2022. Annotating arguments in a corpus of opinion articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1890–1899, Marseille, France. European Language Resources Association.

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barberá, and Ana García-Fornes. 2021a. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Ramon Ruiz-Dolz, Montserrat Nofre, Mariona Taulé, Stella Heras, and Ana García-Fornes. 2021b. VivesDebate: A New Annotated Multilingual Corpus of Argumentation in a Debate Tournament. *Applied Sciences*, 11(15):7160.

A. Francisca Snoeck Henkemans. 2001. Argumentation structures. In Frans Hendrik van Eemeren, editor, *Crucial concepts in argumentation theory*, pages 101–134. Amsterdam University Press, Amsterdam.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Shinobu Suzuki. 2010. Forms of written arguments: A comparison between Japan and the United States. *International Journal of Intercultural Relations*, 34(6):651–660.

Frans H van Eemeren. 2018. *Argumentation Theory: A Pragma-Dialectical Perspective*. Number 33 in Argumentation Library. Springer Berlin Heidelberg, New York, NY.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2019. Argumentation in the 2016 US presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

A Quick Start Guide to Inference Anchoring Theory (IAT)

Centre for Argument Technology
www.arg.tech

May 2, 2022

Segmentation

Segmenting text (or transcribed speech) into units called *argumentative discourse units* (ADUs) is described in two parts. First, we describe analysing propositions: the contents of individual utterances. After describing relations between such propositions, we then return to the issue of segmentation to complete the description of how to map from original text to analysis, by looking at *locutions*.

- An *argumentative discourse unit* (ADU) is any text span which (a) has a **propositional content** anchored in either the locution (ADU) itself or a transition targeting this locution, regardless of whether or not that content is atomic; and (b) has **discrete argumentative function**, in that the propositional content stands in relation to one or more other propositions via one or more instances of inference, conflict or rephrase (described in Section 2).

1 Propositions

1.1 **Basics** Punctuation, delimitation, discourse indicators and other extraneous material that occurs at the boundaries of ADUs are always excluded from the ADU proper

1.2 **Reconstruction** Anaphoric references are typically reconstructed in the text associated with a proposition (i.e. the original text is edited to resolve, e.g., pronouns). The reconstructed ADU should have a form of full grammatical sentence (with subject, predicate, etc.) and should be understandable without the context of what previously has been said, but you should stay as close as possible to what originally has been said, i.e. include as little implicit material as possible.

1.3 **Reconstruction** Propositional content has to be always a sentence – it has have a subject, a verb, a predicate, etc. In other words, you have to reconstruct the missing, implicit material like with anaphoric references. For instance, for the utterance "I didn't say that", you have to reconstruct what she didn't say using the material before this utterance, e.g. if Trump said to Clinton "You said before that taxes should be increased" and she responds "I didn't say that" the propositional content of Clinton's EDU should be: "Clinton didn't say that taxes should be increased". **Exception: the pronoun "we" is not resolved.**

1.4 **Reconstruction** Do as much reconstruction as possible so that you end up with a full sentence which will be understandable without any context, i.e. without knowing what has been said before. At the same time, do as little reconstruction of an implicit material as possible so that you stay close to the original text which you annotate. Think about someone else looking at your analysis without the context of when and who said that. Apply test: "*Will this person be able to understand this sentence?*", i.e. whether this person will understand the argument when looking at the annotation which shows "I didn't say that" as an element of it (who didn't say and what wasn't said?). One the other hand you don't want to over-interpret - this way you might associate a speaker with a standpoint to which they didn't really committed themselves.

1.5 **Splitting** Utterances such as "Yes, but A" (or "No, A") have two segments: "Yes" and "A" ("No" and "A" resp.)

1.6 **Splitting** In many cases, a span will combine clauses that could be identified as separate ADUs. Examples include conjunctions ("A and B"), conditional clauses ("If A then B"), epistemic modalities ("I think that A") and ted speech ("Bob said that A"). In every case, each span with discrete argumentative function should be analysed separately. That said, each category is also associated with a most typical segmentation:

(a) Conjunctions are typically analysed as the two constituent conjunct ADUs because the conjunction itself rarely has discrete argumentative function (except in cases such as 'and'-introduction)

(b) The analysis of conditionals varies: any combination of the three spans ("If A then B", "A", and "B") might have discrete argumentative function

(c) Epistemically qualified statements are typically analysed as a single segment that drops the epistemic modality – e.g. "I think that A" is typically analysed as just, "A"

(d) Reported speech is almost always analysed into two ADUs the first corresponding to the complete span and the second to what was reported to have been said

Interruptions In OVA+, interposed text can present a problem, e.g., "the liquid, because it is so dangerous, is not allowed in the building". In such cases, the text that is interrupted should be identified as one segment, and the interposed text as the other – i.e., in this example, there are two segments: (a) "the liquid, because it is so dangerous, is not allowed in the building"; and (b) it is so dangerous. The first of these two can then be edited to remove the interposed text.

Interruptions In the case of interposed word order that interacts with reported speech such as "Bob, who is an expert, said A", the same rules are applied, allowing three segments to be identified: "Bob, who is an expert", "A", "Bob said A", (the last one is obtained by selecting the whole span of text and then by editing the node to delete "who is an expert")

2 Propositional relations

Relations between propositional contents are about a speaker's (intended) use of linguistic material. It is important that as analysts, we allow arguers to express not just good arguments, but poor, weak, incoherent and fallacious arguments. We are asking ourselves if a speaker intended the content of her utterance to be understood to be related to previous material in a given way.

2.1 **Inference (Support, Default Inference, RA)** Holds between two propositions when one proposition is used in order to provide a reason to accept another proposition. Support may be of a specific kind, depending on the theoretical context an analyst is working in – Modus Ponens, Argument from Expert Opinion, and (the prima facie reasoning from) Perception are all examples of such kinds. If a support relation is not associated with a specific kind, it defaults to 'Default Inference.' Any given support is thus an Application of some Rule of inference, hence RA.
There can be several inference structures identified, with examples provided in the accompanying document:

(a) **Serial arguments** Serial arguments occur when there is an inference relation from a first proposition to a second proposition, and another inference relation from the second proposition to a third proposition.

(b) **Convergent arguments** Convergent arguments occur if there is an inference relation from a first proposition to a second proposition, and an independent inference relation from a third proposition again to the same second proposition.

(c) **Linked arguments** Linked arguments occur if there is an inference relation from a first proposition together with a second proposition to a third proposition.

(d) **Divergent arguments** Divergent arguments occur if there is an inference relation from a first proposition to a second proposition, and an independent inference relation from the first proposition to a third proposition.

2.2 **Conflict (Attack, Default Conflict, CA)** Holds between two propositions when one proposition is used in order to provide an incompatible alternative to another proposition. Conflict may also be of a given kind (e.g., Conflict from Bias, Conflict from Propositional Negation) and defaults to 'Default Conflict.' Note that conflict need not be symmetric. Some kinds (such as Conflict from Propositional Negation) typically are symmetric, which must be captured with two distinct Conflict relations, one in each direction. In contrast to inference, conflict is always structurally the same – it

has only one incoming and one outgoing edge.
There can be a few conflict structures identified, with examples provided in the accompanying document:

(a) **Rebutting Conflict** If a conflict relation targets a proposition (indicating the latter is not acceptable), then the conflict is rebutting.

(b) **Undermining Conflict** If a conflict relation targets the premise of another argument, then the conflict is undermining.

(c) **Undercutting Conflict** If a conflict relation targets the inference relation between two propositions, then the conflict is undercutting.

2.3 **Rephrase (Default Rephrase, MA)** Holds between two propositions when one proposition is used to rephrase, restate or reformulate another proposition. Rephrasing is not repeating: repetition involves multiple utterances with the *same* (i.e. just a single) propositional content. Rephrase involves different propositions connected through a variety of different relations, such as Specialisation, Generalisation, Instantiation, etc. Question answering often involves rephrasing because the propositional content of a question is stereotypically instantiated, resolved, or refined by its answer. In contrast to inference, rephrase is always structurally the same – it has only one incoming and one outgoing edge.

In contrast, conflict (CA) and rephrase (MA) structures are always the same – they have only one incoming and one outgoing edge.

3 Locutions

ADUs are typically directly analysed as *locutions* with an eye on both their propositional content and their discrete argumentative function. One locution typically has one propositional content (to which it is connected by illocutionary connections – see section 5).

Locutions typically have speakers (a term we use to encompass utterers in any medium) and may also have timestamps.

3.1 **Basics** ADUs expressing propositions may overlap and need not be minimal

3.2 **Reconstruction** Ellipsis, pronominalisation, etc., should not be reconstructed in the ADU (which should be simply a span of the original discourse material as uttered); however they *should* be reconstructed in the expression of the propositional content

3.3 **Speakers** Particularly in analysis of dialogues, speakers of particular utterances are identified as part of text of locutions by a convention: "SPEAKER: *ADU*". You should try to go back in the transcript backwards to try to identify the speaker, if it is not immediately clear. In case it is impossible, include as a name "Unknown" as the first name and "Speaker" as the last name.

In OVA, you can associate a locution with either a pre-existing speaker or a new speaker – in the latter case you can type in the forename and surname (with usual capitalisation) which OVA will then prepend with a colon to the locution content. Once you've added a new speaker they'll be available for subsequent selection.

Locutions should be created in order as they temporally appear in the transcript, and kept in that order on the right hand side.

The maps should be clearly laid out, without too many overlapping nodes and edges

4 Transitions

Transitions (TA) connect locutions.

4.1 **Basics** A transition captures a response or a reply and embodies a functional relationship between predecessor locution and successor locution. TAs capture temporal relation as well, but only to a very limited extent - in the sense that the response will always happen later than the locution to which it responds. But the response can refer to something that happened much earlier in the dialogue.

4.2 **Types of TAs** Transitions are of many types, though there are not many good names for these types – our example is *substantiating* used in responding to a challenge. These types of transitions available in a given dialogue type (or context, or activity type) is governed by the protocol in use in that context. A protocol (or dialogue game) is a high level specification of the set of transition types that are available. Usually in practical analysis, TAs are left untyped, and default, therefore, to "Default Transition". Each instance is an Application of a scheme of Transition, hence TA node

4.3 **Adjacency** Transitions often hold between adjacent segments, but not always – a significant minority capture long-distance relationships in cases where, for example, a claim is returned to and given additional support, or an earlier question is refined, and so on. On the other hand, because transitions capture a functional response relation, they never hold in opposition to temporal flow. That is, the directionality of transitions is the same as the temporal ordering. Because the transition structure is branching, however, it is not necessarily possible to reconstruct an absolute ordering over all locutions from TA connections alone

4.4 **Mirroring propositional relations** If you find a long-distance relation on the side of propositional relations (RAs, CAs and MAs between a proposition and another proposition that was said some time long before), then you need to also annotate the connection on the dialogical side which links through TA the locutions for these propositional contents. In this way, you show that the later locution is a response to something which happened some time earlier in the dialogue.

5 Illocutionary connections

Illocutionary connections (IC) link locutions with propositions & propositional relations (i.e. right hand side of the IAT diagram with its left hand side). They are based on the concept of illocutionary force introduced by speech act theory (see e.g. Searle 1969, Searle and Vanderveken 1985). The act $F(p)$ is a communicative act which ties together the locution the propositional content p through the illocutionary force F of, e.g., asserting p, asking about p, requesting p, ordering p, promising p and so on.

5.1 **Basics** Each locution will typically anchor a single illocutionary connection, but may anchor more than one

5.2 **Types of ICs** There is no prescribed set of illocutionary schemes, i.e. types of illocutionary forces (any more than there is are prescribed sets of inference, conflict or transition schemes). Illocutionary schemes suitable for negotiation (that might involve, e.g., *offer* and *reject*) might be different from those involved in a court (where *testify* and *object* might be more appropriate).
That said, many domains share a number of schemes. These commonly applicable schemes are described below:

(a) **Asserting (A)** The speaker S is *asserting* p to communicate his opinion on p. It does not imply that S really believes p: it is rather a public declaration to which the speaker can be held.

(b) **Questioning (Q)** S is *questioning whether* p when S formulates p as interrogative sentence of the form using a Yes/No question or a Wh-question. In both cases, the propositional content is treated as underspecified – as a disjunction for a yes/no question (so, "Is it the case that p?" has the content, "It is or is not the case that p") or as a lambda sentence for a Wh-question: (so, "What time is it?" has content, "The time is x"). We distinguish three categories of questioning: ***Pure Questioning (PQ)***, ***Assertive Questioning (AQ)***, and ***Rhetorical Questioning (RQ)***. In the case of PQ, S is asking for the hearer H's opinion on p: whether H believes p, or not, or has no opinion on p. AQ and RQ, in contrast, carry some degree of assertive force. For AQ, S not only seeks H's opinion on p, but also indirectly publicly declares his own opinion on p. This IC is typically linguistically strongly signalled by cues such as "Isn't it the case that...", "Can we agree that...", "Doesn't...". Finally for RQ, S is grammatically stating a question, but in fact is only conveying that he does (or does not) believe p. A good

test for deciding between RQ and AQ is to check whether it is discursively possible for H to reply to a given question, e.g., whether his response "Yes" to the question "Does the pope wear a funny hat?" would be treated as irrational (or humorous or naive) discursive behaviour.

(c) **Challenging (Ch)** When S is *challenging p*, S declares that he is seeking (asking about) the grounds for H's opinion on *p*. Challenges are a dialogical mechanism for triggering argumentation.

(d) **Agreeing (Agr)** Agreeing is used for expressing a positive reaction, i.e. when the speaker S declares that they share the opinion of the interlocutor. This can take the basic form of signalling such as "Yes", "Indeed", "Most definitely", "Sure", but may as well be a complete sentence. Note that it is not "Yes" on its own that is a bearer of agreement: this is "Yes" as a *reaction to (in relation to)* e.g. an assertive question, that is conveying agreement. Thus this type of IC is anchored in the transition between, in our example above, locution which anchors AQ and the locution "Yes". Agreeing takes as a content a proposition earlier uttered with which the agreement has been expressed (in example – with the propositional content of AQ).

(e) **Disagreeing (Disagr)** Disagreeing is used for expressing a negative reaction, i.e. when S declares not to share the interlocutor's opinion. This can take the form of utterances which have similar meaning to "No" (e.g. "I'm not saying that", "Actually, that's not correct", "Definitely not", "No, it's not") or it can be an utterance with a complete propositional content. In the same way as agreeing, the force of disagreeing being anchored in a transition captures the idea that the full reconstruction of this IC structure (and then its automatic recognition) requires knowing not only that the disagreement has been expressed, but also *at what* the disagreement was targeted. This IC takes as its content a relation of conflict.

(f) **Restating (Rest)** *Restating* is used for expressing the relation of rephrase between propositional contents, i.e. it anchors MA between two propositions in TA between two locutions which takes these propositions as their contents). This can take the form of an utterance that slightly modifies the original content of the locution being restated. When Clinton says: "I want to invest in you. I want to invest in your family.", the latter sentence is a rephrase of a former one. The rephrase relation differs from repeating: "I want to invest in your family" is not a pure repetition of "I want to invest in you", as the propositional content is a specification of what sort of investing is meant by the speaker.

(g) **Arguing (Arg)** S is *arguing* when he defends a standpoint. This IC is signalled by linguistic cues such as "therefore" and "because", however, these indicators occur infrequently in spoken natural language. Arguing takes as a content a relation of inference (i.e. an RA). In other words, the *inference relation* between two propositional contents – a premise and a conclusion (the left hand side of the IAT diagram) is anchored in the *transition relation* between two matching locutions (the right hand side of the diagram) by means of an illocutionary force called *Arguing*.

(h) **Default Illocuting** If an illocutionary connection does not match the guidelines for any IC described above, then it can be labelled *Default Illocuting*. This illocutionary connection is also currently used to connect an MA to its TA anchor when the MA is being used to answer a question.

5.3 **Reported speech** This is not a type of IC, but it accompanies the occurrences of IC in the discourse, when one speaker reports what another speaker said. In case of reported speech, the original locution contains the text as it was said, e.g. TAPPER: Senator Rubio, last October, you said that you're, quote, "generally very much in favor of free trade". This locution is linked with the locution being reported (RUBIO: I'm generally very much in favor of free trade) using *Asserting*.

5.4 **Repetitions** This again is not a type of IC. If the second locution repeats the content of the first locution, then the content of the second locution is the same as the content of the first one. The content of the second locution is anchored in this locution via an appropriate IC annotated according to the guidelines defined above. In implementation, identity conditions for propositions are currently effectively string matching (hence the need for anaphoric and deictic reconstruction).

5.5 **Connections with propositional relations** All RAs, CAs and MAs must be anchored through ICs in TAs.

6 Further reading

- K. Budzynska, M. Janier, C. Reed, P. Saint-Dizier (2016) Theoretical Foundations for Illocutionary Structure Parsing, *Argument and Computation*, IOS Press, vol. 7, no. 1, pp. 91-108

- M. Janier, M. Snaith, K. Budzynska, J. Lawrence, C. Reed (2016) A System for Dispute Mediation: The Mediation Dialogue Game, *Frontiers in Artificial Intelligence and Applications. Proc. of 6th International Conference on Computational Models of Argument (COMMA 2016)*, Pietro Baroni, Thomas F. Gordon, Tatjana Scheffler, Manfred Stede (Eds.), vol. 287, IOS Press, pp. 351-358.

7 Annotation software

The OVA+ argument diagramming tool

OVA+ is an online tool for argument analysis facilitating the representation of the structure of argumentative discourse. You can start using OVA+ freely at the website `ova.arg.tech`. A manual for using OVA+ is available at `arg.tech/index.php/projects/ova-2/`.

The AIFdb and AIFdb Corpora repositories

Analyses produced with OVA+ can be saved as 'argument maps' in AIFdb, an online searchable repository of analysed arguments freely available at `aifdb.org`. The argument maps can be collected in corpora at `corpora.aifdb.org`.