Reliability of Argument Mapping

SEBASTIAN CACEAN
Karlsruhe Institute of Technology
sebastian.cacean@kit.edu

This paper formulates a model to characterize the margin of interpretation in argument mapping in order to deal with hermeneutic underdetermination. Quantitative and qualitative content analysis provide their own strategy to meet the challenge of hermeneutic underdetermination, but also come with severe caveats. This paper combines the positive aspects of both strategies by introducing context dependent reliability thresholds for argument mapping. This allows generalizable results in spite of unavoidable hermeneutic underdetermination.

KEYWORDS: argument mapping, content analysis, diagramming, discourse analysis, reliability

1. OVERVIEW

The strategy followed in this paper proceeds along the following lines. I will, first, distinguish two argument mapping techniques, reconstructive argument analysis and surface analysis, and argue that both have to deal with hermeneutical underdetermination. The context provided by two related methods, Critical Discourse Analysis and Discourse Quality Index, will illustrate relevant ramifications of hermeneutical underdetermination for empirical research based on argument mapping. I will then provide a model to characterize the margin of interpretation when analysing the surface reasoning structure of a text. What will be explained is what kind of ambiguities prevail in persuasive texts and how they relate to each other. Due to combinatorial complexity it is not possible to specify the interpretational margin by an enumeration of all adequate interpretations. Instead, I will provide an alternative way of describing it that is expressive enough. Using simple graph distance metrics, it is even possible to describe the margin quantitatively in order to specify its size. Finally, a simple example will be used to illustrate how the characterization of the interpretational margin can be applied to assess such margins empirically and how context-dependent reliability measures can be introduced to enable generalizable empirical findings with argument mapping techniques.

2. ARGUMENT MAPPING

The technique of argument mapping is a method that is used to represent the reasoning structures of single arguments or whole argumentations. Argument mapping has its origins in informal logic, argumentation theory and legal reasoning, and is nowadays widely used in artificial intelligence, especially in the context of automatically extracting the reasoning structure in texts (see Reed, Walton, and Macagno 2007 for a historical overview and Lippi and Torroni (2016) for an overview of argument mining). Additionally, the abundance of argument-mapping tools (Scheuer et al. 2010) and their use in contexts of teaching critical thinking skills has recently attracted the attention of empirical researchers. Several findings suggest that the use of argument mapping improves students' critical thinking skills (e.g. Cullen et al. 2018, Eftekhari and Sotoudehnama (2018)).

But what exactly is an argument map? Though different mapping techniques differ in their detail, some features are shared by most. Reasoning structures are modelled as directed graphs. In contrast to concept maps and mind-maps, the vertices and edges of argument maps have a more precise meaning, tailored to analyse the reasoning structure of given texts (see Davies (2011)). The vertices of an argument map represent propositional-like entities such as arguments, reasons, claims and premises and the edges represent the inferential relationships between these vertices. Techniques differ, for instance, in terms of what exactly the vertices and edges may represent, whether the graph is confined to tree-like structures and how they visualize the internal structure of arguments. If an analysis is confined to single arguments or the argumentation for one major claim, tree structures will suffice despite their limitations (see Freeman 1991, p. 16). If, however, an analysis is supposed to cover multiple claims and inferential relationships between arguments, tree structures are often too constrained (see Betz and Cacean 2012, Cacean (2012) for examples of complex argument maps).

The mapping method on which this paper is based is simple, but expressive enough to represent complex argumentations and not constrained to tree structures (for a methodological background see Betz (2010) and Betz (2013)). Nodes can represent claims, i.e. single propositions, or whole arguments, which have a complex inner premise-conclusion structure. In simple cases, an argument node can be interpreted as a set of propositions or one single proposition (the premises), which are supposed to justify another proposition. There are two types of edges, visualizing support and attack relations. A support relation from a node A to another node B represents a vindicatory relation. In the case that A is an argument node and B is a claim, B can be interpreted as the conclusion of the argument represented by A. In the case that both A and B are argument nodes, A can be interpreted as justifying one of the premises of B. Attack relations represent objections, i.e. justifying the falsehood of claims or, in the case of arguments, justifying that one of the premises is false.

What can be distinguished are different depths of argumentative analysis. For the purpose of this paper, the distinction between what might be called *surface analysis* and *reconstructive analysis* of argumentation is important. A surface analysis aims at identifying the reasoning structure in a given text as it is intended by the author only (see Fisher (2004) for an overview). Text segments have to be categorized into those that the author presents as claims, conclusions, and assumptions and those that are presented as supporting reasons for or objections to claims and arguments. This identification of vindicatory relationships can be understood as a mere annotation of a text. In particular, it is not to be understood as evaluating the mentioned reasons and their relations to claims or other reasons. Whether an argument is considered to be good by certain standards is not part of this surface analysis, which is confined to the identification of intended relations only. Such a surface analysis can, however, be used as a starting point for the more exegetical technique of reconstructive analysis. Often arguments are stated incompletely: premises or even conclusions are not mentioned explicitly. A reconstructive analysis aims at making these implicit parts of an argument explicit by inferring them from the surrounding text-context with the help of hermeneutic principles, such as accuracy and charity (Brun and Hadorn 2009, chapt. 8; Betz and Brun 2016; Fisher 2004, p. 17).

It seems uncontroversial that reconstructive analysis is hermeneutically underdetermined in most cases since there are often different possibilities of adding implicit premises and conclusions. Although the mere surface analysis does not aim at identifying implicit premises, it is often hermeneutically underdetermined as well, because the understanding of the author's intended meaning will depend on the background knowledge of the person interpreting the text (Fisher 2004, p. 22). Ideally, explicit reasoning indicators point uniquely to relevant text segments and reveal their vindicatory role. However, often these linguistic cues are ambiguous. For instance, phrases using the word *'because'* might indicate a reason-relation, but also a mere explanatory relation. Often, there are not even any explicit indicators and the intended meaning has to be inferred from the text-context alone (Fisher 2004, p.

16). As a consequence, even the surface analysis is a hermeneutic process and is always tentative in its results.

3. POLITICAL DISCOURSE ANALYSIS & DISCOURSE QUALITY INDEX

Hermeneutical underdetermination is not necessarily an obstacle and is dealt with constantly within empirical research designs. It is instructive to have a look at *Political Discourse Analysis (PDA)* as representative of a qualitative paradigm and the *Discourse Quality Index (DQI)* as representative of a quantitative paradigm in order to see how they deal with hermeneutic underdetermination.

The argumentative turn in policy analysis, a term coined by Fischer and Forester (1993), introduced the use of methods from philosophy and argument analysis to understand political discourse first and foremost as practical reasoning (see Hansson and Hirsch Hadorn (2016) for an overview). According to this account, political decision-making is primarily a deliberation over different possibilities for political action. Non-argumentative elements such as narratives and explanations can be understood inasmuch as they are embedded within practical reasoning as premises of practical arguments (I. Fairclough and Fairclough 2012, p. 13). The approach of Political Discourse Analysis of (I. Fairclough and Fairclough 2012) provides an account of the structure of practical argumentation and demonstrates the feasibility of that method by analysing the political discourse surrounding the financial and economic crisis that began in 2007 (I. Fairclough and Fairclough 2012, pp. 1–2). PDA is not limited to a mere descriptive analysis of argumentation but strives to enable a critical evaluation of practical argumentation (I. Fairclough and Fairclough 2012, p. 11). By using argumentation theoretic methods, which are heavily influenced by Walton Schemes (e.g. as described in Walton (1996) and D. N. Walton (2006)), PDA proceeds along the following lines: First, the premises and conclusions of arguments have to be identified in a text or have to be construed from the text-context. Having made the reasoning structure of the practical argument explicit, arguments can then be critically examined, by either questioning the acceptability of premises and conclusions or by questioning the vindicatory relation between the premises and their conclusions (I. Fairclough and Fairclough 2012, p. 12). PDA expands and refines the approach of Critical Discourse Analysis and, in consequence, shares its main features (I. Fairclough and Fairclough 2012, p. 10). Discourses are understood as historical and embedded in cultural contexts. Consequently, discourse analysis must consider this context-dependence and is always an open-ended hermeneutic process of interpretation (Titscher et al. 2000, p. 146 and p. 167). This is mirrored in the quality criteria for critical discourse analysis. Since the results of such analysis remain relative to a specific interpretation, the strong quality criteria of quantitative research methods such as reproducibility or validity do not play an important role in PDA. Rather, the results of discourse analysis should be transparent and recognizable and the interpretations must be intelligible (Titscher et al. 2000, p. 164).

Similar to PDA the Discourse Quality Index is used to analyse political discourse. The DQI is methodologically based on content analysis (see Krippendorff (2012) and Neuendorf (2002) for an overview) and intended as a quantitative measure to assess the quality of discourse. The DQI uses seven different coding categories to classify text segments, which are based on Habermas' discourse ethics (Steenbergen et al. 2003, p. 21). In a first step, relevant text segments, so-called coding units, have to be identified according to some relevance criteria. In analysing parliamentary debates, coding units of DQI are speech acts containing *"proposal[s] on what decisions should or should not be made"* (Steenbergen et al. 2003, p. 27). In a second step, these relevant text segments have to be categorized. For instance, the subcategory 'level of justification' has four values: 'no justification', 'inferior justification', 'qualified justification' and 'sophisticated justification' and is used to identify formulated reasons and to assess their quality (Steenbergen et al. 2003, p. 28). The DQI enables empirical research of discourse for a diverse spectrum of questions. For instance, Baccaro, Bächtiger, and Deville (2016) investigated how different procedural ways of structuring deliberation relate to the discourse quality, Caluwaerts and Deschouwer (2014) investigated in a deliberative experiment how group-compositions and the applied decision-making rule are related to the discourse quality and Caluwaerts Didier and Min (2014) scrutinized the effects of discourse quality on attitude change. In other words, the DQI-approach helps to understand under which conditions deliberation works and which ends can be achieved by deliberation. Such insights can be considered to restructure deliberative politics and to estimate the limitations of deliberation. Admittedly, this is only possible if the results of such research are in some sense general statements about the relationship between the discourse quality and other factors. To that end, the quality criteria for DQI are much stronger than those for PDA. The results of applying DQI must at least be reliable. That is, the measurements must be reproducible and lead to sufficiently similar results if repeated under the same conditions. Reliability amounts to an agreement in the categorization of text segments. That is, if one coding unit is categorized independently by different coders who were instructed in the same way, the coding should yield the same results. There are different quantitative measures for the assessment of this intercoder reliability, which usually take the possibility of agreement by mere chance into account (see Artstein and Poesio (2008) for an overview).

I introduced DPA and DQI as different approaches analysing argumentation to exemplify two different ways of handling interpretational leeway. DPA is already a reconstructive analysis of argumentation and just accepts that it is a hermeneutical approach with a non-diminishing margin of interpretation. Different analysts with different background knowledge and possibly different opinions as to what context should be considered may come two different results. Given the focus on specific case studies, this is not necessarily a drawback. DQI, on the other hand, attempts to provide general empirical insights about what drives different forms of deliberation. Reliability, in this context, is necessary and can be achieved if the coding instructions are precise enough. The strategy of reliability driven content analysis is to diminish the margin of interpretation in the application of the categories by providing precise coding instructions.

4. HERMENEUTICAL UNDERDETERMINATION IN ARGUMENT MAPPING

The question addressed in this paper is whether an argument mapping in the form of the described surface analysis can be applied as a social-empirical research method to generate interesting results. There are numerous interesting research questions. For instance, Betz (2013) simulates complex multi-agent debates to investigate the effects of different argumentation strategies on reaching a consensus. The question is whether the findings are mirrored in reality and under which conditions. A corresponding argument mapping of real debates could answer these kinds of questions. However, such empirical research designs face a hermeneutical challenge: In order to provide generalizable results, the corresponding argument mappings of texts must be reliable. That is, the argument mappings of one text by different coders should result in sufficiently similar argument maps. As hinted at above, the described argument mapping approach often allows for different interpretations. As a consequence, coders can choose between different interpretations and we should not expect high reliabilities even if the coding results are adequate.

The basic idea to meet this challenge is that an unavoidable hermeneutic underdetermination in the case of argument mapping does not necessarily lead to a purely qualitative analysis. An important point is that even in the face of underdetermination one can often distinguish adequate from inadequate interpretations. This allows for capturing the margin of interpretation in a specific context. Instead of dispensing with reliability measures altogether, I suggest using context-dependent reliability constraints. The margin of interpretation can vary even from one text to another. Having captured the margin quantitatively for, say, a specific text type, one can define reliability thresholds that an adequate argument mapping has to fulfil.

5. CAPTURING INTERPRETATIONAL MARGINS

An apparently simple way to describe the margin of interpretation of a given text is an explicit listing of all of the valid argument maps. Often, however, the number of valid argument maps is vast due to combinatorial complexity.[1] Therefore, such an extensional way of describing the margin of interpretation is at the very least impractical and often even impossible since computational and human capacities are limited. An alternative way of describing the margin of interpretation uses the concept of a maximum argument map, which in some way contains all valid argument maps as subgraphs. This basic idea will be elaborated by explicating the concept of a maximum argument map and by formulating validity criteria without referring to sets of all valid argument maps.

This approach, in some way, provides an explication of the concept of valid argument maps, which can, however, be easily misunderstood. Namely, I do not intend to elaborate criteria to evaluate the validity of an argument map with respect to a given text. Such an account would represent a systematic theory of text-interpretation, which would rely on hermeneutic principles since validity cannot be evaluated by formal text-characteristics alone. What I intend is much more modest and assumes that there are such criteria in a systematic or informal way. What I address is merely the described combinatorial challenge by providing an alternative representation of the margin of interpretation. This alternative representation can then be used to check a given argument map for its validity.

Figure 1 provides an illustration: This paper presupposes that there are criteria that can be used to assess whether an argument map A constitutes a valid interpretation of the reasoning structure of a given text T ($Val_1(T,A)$). I do not provide an explication of Val_1 but of another concept Val_2. In order to enable validity checks, the margin of interpretation has to be described by a maximum argument map A_{max}. Then the validity of a given map A can be evaluated with the help of a validity concept Val_2, which relies on the concept of a maximum argument map alone ($Val_2(A,A_{max})$). A_{max} can be understood as the result of a mapping $MoI(T)$ from the given text T. Needless to say, the

[1] This is, of course, an empirical claim, which cannot be justified here rigorously and depends heavily on the given text.

description of the margin of interpretation via A_{max} has to be obtained with the help of a content analysis itself.

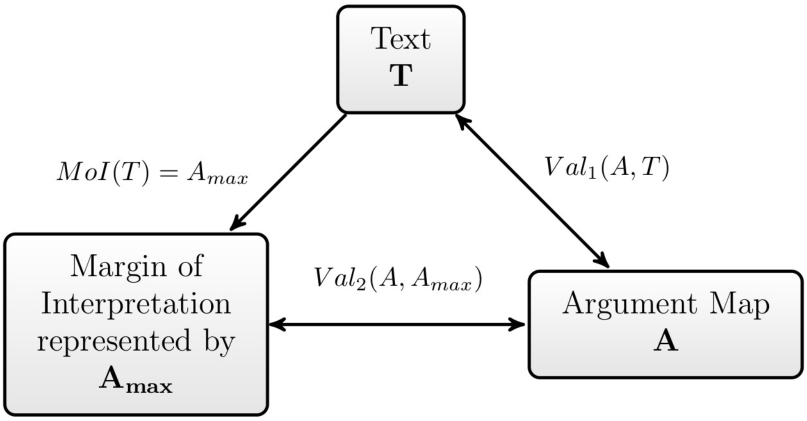


Figure 1 – Relationship of both validity concepts

6. MINIMUM & MAXIMUM ARGUMENT MAPS

The basic idea of characterising the interpretational margin is to restrict it from two sides by specifying those elements that have to be represented in every valid map and those that should not be represented. Most persuasive texts have a non-vanishing interpretational margin, which can be explained by the existence of text segments for which it is unclear whether any vindicatory function is intended. Sometimes, for instance, it is unclear whether a text segment is formulated as an additional argument or, say, a mere explanation or illustration of a point already made. The elements of an argument map can be divided into those that represent text segments that unambiguously have a vindicatory function and those that can be interpreted as having one but also allow for other interpretations. Minimum argument maps are maps that are not further reducible: Any further removal of an element results in an invalid map. Maximum argument maps contain all these unambiguous elements but additionally all the ambiguous ones. If a text segment might be interpreted as having a vindicatory function it should be represented in a maximum argument map. Maximum argument maps cannot be enlarged without jeopardizing their validity. In sum, the minimum and maximum argument maps represent the lower and upper bounds for the valid argument maps. Any valid argument map should contain all elements of a minimum argument map, but should not contain more elements than a maximum argument map. Often, the interpretational margin of a given text must be represented by more than one minimum argument map. Fortunately, often one maximum map will do to represent the interpretational margin. For simplicity and brevity, I will assume in the remaining description that there is exactly one maximum argument map. In this case, the set of all valid argument maps has a tree-structure as illustrated in figure 2. The edges represent a subgraph-relation: The argument maps $A_1 - A_9$ have fewer elements than the maximum and more elements than the minimum argument maps $A_{10} - A_{15}$.

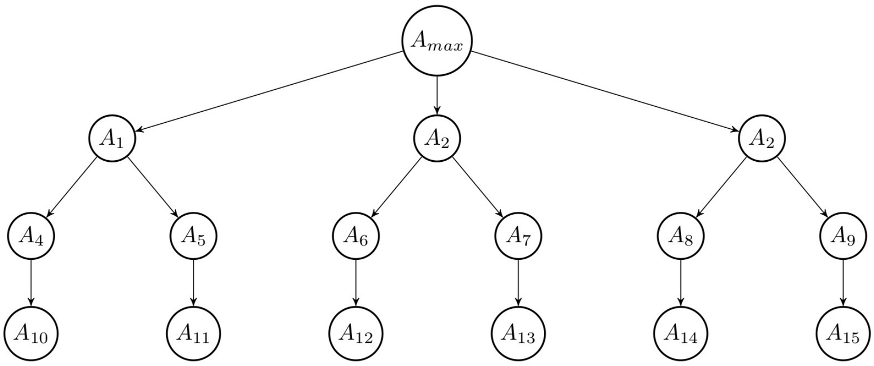


Figure 2 – Tree structure of valid argument maps.

An argument map contains nodes, which represent propositions or sets thereof and directed edges, representing attack and support relations. To elaborate on the described idea of minimum and maximum maps it is crucial to understand in what way ambiguities of the given text relate to elements in the argument map. There are two principal different kinds of ambiguities when it comes to the reasoning structure of a given text. The first kind concerns the question of whether a specific text segment has an intended vindicatory function. That translates into the question of whether this text segment should be represented in the argument map. In the case that a text segment is interpreted as having an intended vindicatory function another kind of ambiguity can occur. It might be ambiguous which vindicatory function is intended by the author. For instance, it might be unclear what exactly is intended to be justified by the text segment.

Given the structure of argument maps the following ambiguities can be distinguished:

- *Node-ambiguity*: Whether a given text segment has some intended vindicatory function can be ambiguous. Even in the case of explicit indicator words, such ambiguities might prevail. For instance, phrases like *"... because"* can indicate causal and vindicatory relationships. Sometimes the whole text context does not determine uniquely which one is intended.

- *Relation-ambiguity*: Similarly as with nodes, relations between nodes are often hinted at by linguistic cues in the text. A relation is called unique if the text context and/or explicit indicators unambiguously point to a relation between nodes.

The latter two ambiguities concern the question of whether a text segment has a vindicatory function. The following concern the exact type of the vindicatory function:

- *Ambiguity of edge-type*: It might be unclear whether an identified relation is supposed to be intended as a support or an attack.
- *Ambiguity of direction*: The edges of argument maps are directed. That is, they have a source and a target. Similarly as with the edge type, the direction of an edge might be formulated in the text ambiguously.

Both, the type and the direction of an edge are usually stated in an unambiguous way in texts. What occurs more often is that either the source or the target is ambiguous, for which the following technical termini are introduced:

- *Source-unique relations*: If the source of a relation is stated unambiguously in a given text, the relation is called source-unique. That is, the text states clearly from which node the relation comes.
- *Sink-unique relations*: If the target of a relation is stated unambiguously in a given text, the relation is called sink-unique. That is, the text states clearly to which node the relation aims.

The question of whether a relation is source- or sink-unique is independent of the relation-ambiguity. Relation-ambiguity concerns the existence of a vindicatory relation. Source- and sink-uniqueness, on the other hand, are related to ambiguities with respect to the source and target of a relation. That is, there might be ambiguous source-/sink-unique relations and there might be unique source-/sink-ambiguous relations. With respect to sink- and source-uniqueness there are four combinatorial possibilities.

Having introduced the relevant types of ambiguities, the construction of a maximum argument map can be outlined. A maximum argument map is an argument map complemented with additional information about the ambiguities found in the text.

1. *Nodes*: Text segments that have a vindicatory function, which are either being used to justify something or are being justified,

should be represented in the maximum argument map by an ambiguous or unambiguous node.

As illustrated above, there might be unambiguous relations for which there are different valid interpretations of their target. The question is how to represent the different interpretations in the maximum argument map. An unambiguous relation, which lacks sink-uniqueness, has to be represented by not only one edge but different edges, each representing one possible interpretation. The suggestion is then to represent a relation by an equivalence class of edges, which are called representatives of the relation. The construction of edges in the maximum argument map is in consequence as follows:

2. *Relations*: Text segments indicating the existence of a vindicatory relation in an ambiguous or unambiguous way have to be represented by equivalence classes of edges. Relations that are not sink-unique and/or not source-unique or exhibit some other type of ambiguity have a corresponding edge for each valid interpretation in the equivalence class.

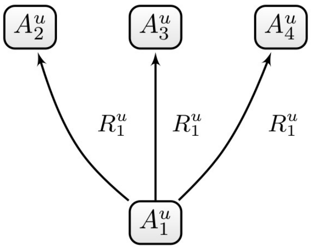


Figure 3 – Example of one unique relation with three representatives

Figure 3 provides an illustration. The given maximum argument map can be read as follows: There are four text segments that are identified as arguments in a unique way ($A_1 - A_4$) and one unique relation, represented by the equivalence class R_1.[2] That is, the text uniquely indicates that the text segment represented by A_1 is being used to justify something. However, it is not clear what exactly the author intends to justify with it: A_2, A_3 or A_4 or all of it. This abstract example already hints at why the margin of interpretation might be combinatorically complex. If, as the example is constructed, every combination of the

[2] The illustrations use u-indices as superscripts to indicate uniqueness and '$\sim u$' to indicate ambiguity respectively.

representations of that relation represents a valid interpretation, there are already *3!*, i.e. *6* valid argument maps. This illustrates that the amount of valid argument maps is roughly in the magnitude of *n!* with *n* being the number of ambiguities.

The properties of relations being sink-unique or source-unique do not have to be encoded or visualized separately since they can be defined as follows:

- A vindicatory relation is *source-unique* if and only if all representatives of that relation have the same source.
- A vindicatory relation is *sink-unique* if and only if all representatives of that relation have the same target.

As a consequence, relations that are unique in every way have exactly one edge as representative in their equivalence class.

There are cases of simple maximum argument maps that allow for exactly one minimum argument by a stepwise reduction of the maximum map. Consider the example of figure 4: There are two unambiguous main claims (C_1 and C_2), one ambiguous node (A_1) and two ambiguous relations (R_1 and R_2), each of which have one edge as representative. If you remove all ambiguous elements, the resulting minimum argument includes only the main claims.

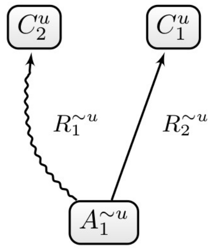
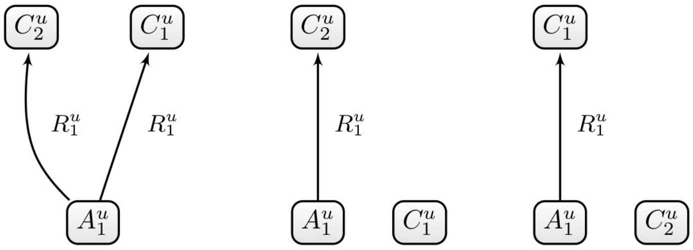


Figure 4 – Maximum map

However, often there is more than just one minimum argument map as exemplified by the following example.

In figure 5 every node is unique. Additionally, there is a unique relation (R_1), which has two representatives. The relation is source unique, but lacks sink uniqueness. The question is now, what elements can be removed from the maximum argument map without jeopardizing validity. Given the terminology introduced, it is required that there is at least one edge representing a relation in the case that the relation is unambiguous. That is, a valid argument map should provide at least one interpretation of a unique relation, even if the relation is not sink unique or not source unique. Consequently, both edges cannot be removed simultaneously in figure 5. However, each one can be removed separately, resulting in two different argument maps (figure 6 and 7) which cannot be reduced any further.



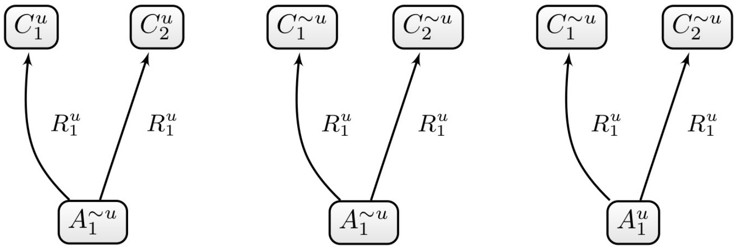
Figures 5-7

7. THE VALIDITY CONCEPT

The concepts elaborated so far can now be used to formulate the conditions that a valid argument map has to fulfil. Although the concept of a minimum argument map is helpful to understand the basic idea, it is possible to formulate the validity criteria without reference to minimum argument maps. The following intuitions seem to be a good starting point:

a. A valid argument map should contain all unique nodes.
b. A valid argument map does not have to contain ambiguous nodes.
c. For every unique relation there has to be at least one representative in a valid argument map.
d. A valid argument map should not contain elements that are not in the maximum argument map.
e. Every edge should have a source and a target.

Condition *(e)* merely ensures that the argument map is really a directed graph, that is, that the edges do not point into the void or come from the void. The intuitions *(b)* and *(c)* are, however, in some tension with each other, because the existence of representatives of a relation depends on the existence of corresponding source and target nodes. Consider the cases of figures 8-10.



Figures 8-10

According to the intuition that ambiguous nodes do not have to be in a valid argument map *(b)* the argument node A_1 could be removed in figure 8. That would, in turn, lead to the removal of all edges according to condition *(e)*. As a consequence, the relation R_1 would not have any representatives, which violates the intuition *(c)* that all unique relations should have a least one representative. A similar consideration applies to figure 9 with respect to all nodes and to figure 10 with respect to the removal of C_1 and C_2. There are three possibilities to resolve the tension between the intuitions *(b)* and *(c)*:

1. A *precedence over unique relations* suggests demanding that every unique relation must have at least one representative. All of the cases discussed would be handled in the same way with the consequence that sometimes ambiguous nodes could not be removed.
2. A *precedence over the removal of ambiguous nodes* suggest that ambiguous nodes can be removed, even if that implies the removal of the last representatives of unique relations. According to this approach, the discussed maps could be reduced until no representative of R_1 is in the map. The map in figure 9 could even be reduced to an empty map.
3. A *mixed approach* suggests handling the cases discussed differently. To take precedence over unique relations in some cases and precedence over the possibility of removing ambiguous nodes on other.

The second option is advantageous if one prioritizes a principle of charity with regard to the analysis of the reasoning structure. It simply allows more valid interpretations of the given maximum maps. Whereas according to the first option no node in figure 8 and only one node in figures 9 and 10 could be removed, the second option allows the removal of more nodes (one in figure 8, two in figure 10 and even three in figure 9).

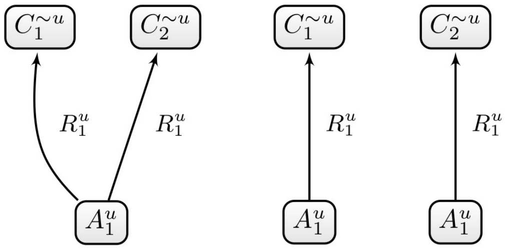
Although the principle of charity might favour the second option, the precedence over unique relations has a pressing appeal especially with respect to figure 10). Let's consider the interpretational situation of this case in an abstract way. Suppose there is an explicit argument indicator that unambiguously points to a text segment that is used to justify something. That motivates the representation of that text segment by node A_1 and the unique support relation R_1. However, the linguistic cue does not single out the target of that node, i.e. what is supposed to be justified. Two other text segments might be interpreted as main claims, which are represented by C_1 and C_2 respectively. The fact that A_1 and R_1 are uniquely pointed at by a linguistic cue demands some interpretation of what is supposed to be justified by A_1. In other words: The given text provides a text segment as a justification. However, it is unclear what exactly is being justified - either C_1 or C_2. A valid interpretation has to provide at least one answer to the question of what is being justified.

Instead of choosing either option one or two, I opt for using a mixed approach that handles the cases of figures 8-10 differently. In particular, I suggest that the ambiguous nodes of figures 8 and 9 can be removed, even if by doing so some unique relations remain without representatives. The case of figure 10 should, however, be treated differently. The fact that both the source of the source-unique relation and the relation itself are unique prohibits the removal of all ambiguous nodes.

This mixed approach can be captured by the following conditions: An argument map AM is valid with respect to a text T only if

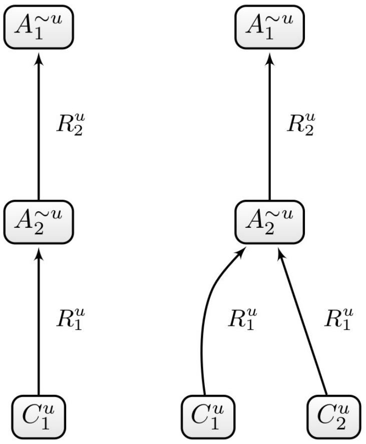
i. every edge in AM has a source and a target,
ii. every node and edge of AM is also an element in the maximum argument map of T,
iii. every unique node is a node of AM,
iv. for all unique relations which are source unique, exists at least one representative in AM, in the case that the source node is in AM, and
v. for all unique relations which are sink unique, exists at least one representative in AM, in the case that the target node is in AM.

In order to illustrate the consequences of these conditions, let's consider, first, the case of figure 11, which I have already used to motivate the conditions (see figure 10):



Figures 11-13

In comparison to figure 11, the argument maps in figures 12 and 13 lack one of the nodes of the maximum argument map of figure 11. However, they do not violate any of the conditions (i-v). The question is whether these submaps can be further reduced. Relation R_1 is unique and source-unique. The corresponding source node is unique. Hence, according to (iii) and (iv) neither the source nor the last representative of R_1 can be removed and consequently the corresponding target nodes have to remain. Hence, the maps in figures 12 and 13 are minimum argument maps.



Figures 14 and 15

The case illustrated in figure 14 shows that there are constellations in which last representatives cannot be removed even if the source and target node are ambiguous: The source node C_1 cannot be removed since it is unique. There is only one representative of R_1, which renders R_1 source (and sink) unique. Hence, the ambiguous node A_2 cannot be removed according to condition (iv). The relation R_2 is also source unique and since A_2 is in every valid map, so is the only representative of R_2 and its target A_1. In sum, the maximum map is also a minimum map, since no element can be removed without violating the validity conditions. The map in figure 15 is different. If nodes A_1 and A_2 and with them all edges are removed, the resulting map is still valid. The relation R_1 is sink-unique but not source-unique. As a consequence, condition (iv) does not apply. Since A_2 is not part of the reduced map, condition (v) does not apply either.

8. CONTEXT DEPENDENT RELIABILITY THRESHOLDS

Reliability requires that different coders agree in their codings of the same text (inter-coder reliability) and that a repeated coding of the same text by one coder yields the same results (intra-coder reliability). Whereas reliability quantifies the agreement among repeated codings of the same text, validity is a concept concerned with truth. Only if a measurement instrument measures what it is supposed to measure can the results be called valid. Though height reliabilities do not guarantee validity, they are at least necessary for validity (Krippendorff 2012, p. 213). If coders are not consistent with each other some of them must be wrong or the categories of the coding scheme are not appropriately precise (Artstein and Poesio 2008, p. 557). In agreement with this positivistic view of coding, the thresholds for adequate levels of reliability are not context-dependent. Although the numerical specifications of these thresholds might depend on whom you ask and on the particular reliability measure being used they do not depend on the interpretational margin of the text. If the category-system allows a margin of interpretation, it is simply not suited for reproducible measurements of text characteristics and as a consequence does not allow generalizable results.

What I suggest is taking a stance between the sketched positivistic picture of coding as a measurement process and giving up on striving for reproducibility entirely in the case of hermeneutical underdetermination. It is true that argument mapping is, like many methods of semantic text analysis, an interaction between the text and the reader of the text and often allows for different interpretations. Nevertheless, it can be susceptible to reliability constraints. The simple idea is to specify reliability thresholds relative to the margins of interpretation.

Before providing an outline of this approach, I want to address an important worry, which could be formulated as follows: If we are provided with a coding of the interpretational margin of the reasoning structure of a particular text, we do not need any further valid codings in the form of argument maps of that text. We already have them in an encoded form via the maximum argument map. Hence, there is also no need for additional argument maps to be checked for validity and reliability. If the specification of the reliability threshold for a particular text relies on coding the interpretational margin, we do not need the reliability threshold for the same reason. Its only purpose is to assess whether the given argument maps satisfy reliability constraints. This concern shows that the idea of context-dependent reliability thresholds is only fruitful if margins of interpretation can be estimated without coding them for every text explicitly. What is called for are text features that can serve as proxies for the interpretational margin and that are easier to detect than the interpretational margin itself. Whether such proxies exist is an open empirical question. Perhaps margins differ so severely that there is no other viable way to estimate interpretational margins than to code them explicitly via maximum argument maps. The small contribution of this paper is to enable such empirical research.

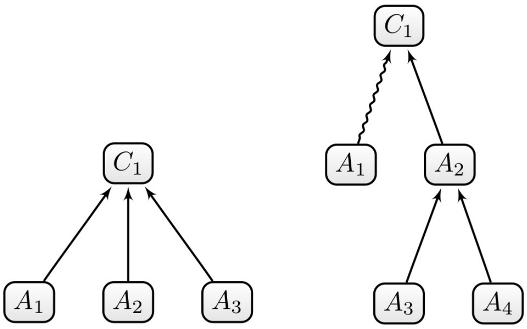


Figure 16 – Illustration of the Hamming distance

In addition to that, the account of context-dependent reliabilities hinges on a crucial assumption. It should be possible to capture the margin of interpretation in some quantified way. Fortunately, this concern is easily dealt with. Both the margin of interpretation and the agreement among different codings can be assessed quantitatively by using graph distance measures. A very simplistic one is a non-normalized Hamming distance, which simply counts the number of elements that are not shared by two argument maps. Figure 16 illustrates what is meant by this: There is only one node that is not an element in both argument maps (A_4). There are two edges in the first argument map that are not present in the second one (the supporting edges from A_1 to C_1 and from A_3 to C_1) and three edges in the second map that are not present in the first one (the attacking edged from A_1 to C_1 and the supporting ones from A_3 and A_4 to A_2). In sum, the Hamming distance is six.

The Hamming distance can be used to introduce different measures to quantify the reliability of argument mapping and the interpretational margin. For simplicity, let's use the mean distance between argument maps as a numerical value for their reliability. The leading idea of the context-dependent reliability measure can now be described as follows: If there is some margin of interpretation, a coder is allowed to pick any interpretation that is valid. Hence, we should expect that different valid argument mappings might result in different argument maps. As a consequence, low reliability does not imply invalidity. That is, high reliability is not necessary for validity as in the positivistic picture described above. However, that does not mean that anything goes. If, for instance, the distance between two argument maps exceeds the maximum distance within the set of all valid argument maps, one of them must be invalid. This can be generalized: low reliability is still a probabilistic indicator of invalidity. But how much disagreement among different mappings is tolerable? A strong requirement of discrimination might demand that the disagreement between coded argument maps should not exceed the mean distance of all valid argument maps. A more apt specification of reliability thresholds should take into account the distribution of argument maps within the interpretational margin. However, without further findings about these margins, no particular constraints can be formulated.

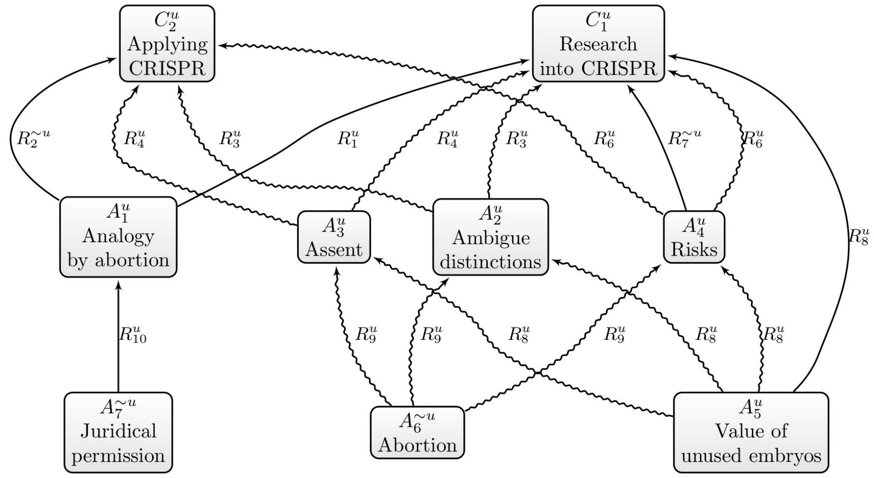


Figure 17 – Maximum argument map

Let us, however, consider an example to illustrate the approach described. The maximum argument map of figure 17 represents the reasoning structure of answers to a questionnaire with open questions. The respondents were asked to formulate their opinions about human germline editing and to provide reasons for it. Additionally, they were asked to deal with objections they know of. Both, the particular response to that questionnaire visualized in figure 17 and the different codings of it were generated by students during a research seminar, which was part of a larger participatory research project at the Karlsruhe Institute of Technology (KIT).[3] The research seminar served as a test vehicle to assess whether the evaluation of questionnaires with argument mapping techniques is feasible. Figure 18 visualizes the distances of different argument maps. Each point represents an argument map and the length of a line linking two points represents the Hamming distance between the corresponding argument maps. There are seven different argument maps as coding results by students and fifty randomly generated valid argument maps based on the given maximum argument map. The mean distance of the coded argument maps is 12.9 ± 3.6 and 7.3 ± 2.0 of the randomly generated valid maps. All of the coded argument maps are, however, invalid, which might be explained by insufficient coding instructions. The students had no argumentation theoretic background and had only a short-term introduction to argument mapping (roughly one and a half hours) before creating the argument maps.

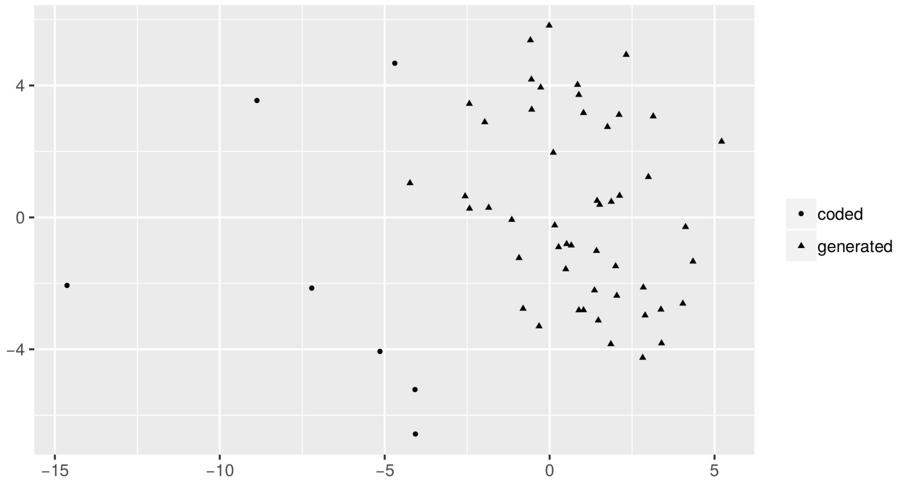


Figure 18 – Visualized distance in two dimensions

The given numbers illustrate how interpretational margins can be captured quantitatively. However, they describe only one particular case. Further empirical research has to show, which reliability thresholds would be appropriate and whether there are properties of texts that allow inferences to their interpretational margin.

[3] See http://www.buedeka.de/ for more information about the project "Citizen-Delphi", which was funded by the German Federal Ministry of Education and Research.

Let me finally provide an outline of how context-dependent reliability thresholds meet the challenge of hermeneutical underdetermination. The question is, how the described technique of argument mapping can be applied as an empirical research method to allow generalizable results. The problem of generalizing the results of a content analysis with non-optimal reliabilities can be described roughly with the following picture: Empirical research strives for justified general statements about the causal or at least correlational relationships between observable properties of phenomena. Put simply, the researcher asks whether a difference in some independent variable makes a difference in some other dependent variable. In order to answer this question, the corresponding differences have to be measured. Since the researcher is interested in differences in the phenomena, the measured differences must be an indicator of differences in the phenomena and not a mere artefact of the measurement process. Applied in the context of argument mapping: If two argument maps are different, these differences should be the result of differences in the coded reasoning structure and not the result of interpretational differences only. The latter case would say more about the analyst than about the coded text. But how can we exclude that, if there is a non-vanishing interpretational margin? Capturing interpretational margins quantitatively might solve these problems since it allows for estimating whether a difference in argument maps could be explained by interpretational differences alone, given that the argument mapping is valid. That is to say, capturing the margin of interpretation quantitatively in argument mapping is similar to specifying confidence intervals or error bars in measuring other quantitative data. It allows for evaluating whether the observed differences are significant enough to infer differences in the phenomena observed.

ACKNOWLEDGEMENTS: This work has been funded by the Federal Ministry of Education and Research (BMBF, Germany) in the project "Citizen-Delphi" (grant number 01GP1776).

REFERENCES

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, 34(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Baccaro, L., Bächtiger, A., & Deville, M. (2016). Small Differences that Matter: The Impact of Discussion Modalities on Deliberative Outcomes. British Journal of Political Science, 46(3), 551–566. https://doi.org/10.1017/S0007123414000167

Betz, G. (2010). Theorie dialektischer Strukturen. Frankfurt am Main: Klostermann.

Betz, G. (2013). Debate Dynamics: How Controversy Improves Our Beliefs.

Betz, G., & Brun, G. (2016). Analysing Practical Argumentation. In S. O. Hansson & G. Hirsch Hadorn (Eds.), The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty. (pp. 39–77). Cham: Springer.

Betz, G., & Cacean, S. (2012). Ethical Aspects of Climate Engineering. Retrieved from http://digbib.ubka.uni-karlsruhe.de/volltexte/1000028245

Brun, G., & Hadorn, G. H. (2009). Textanalyse in den Wissenschaften: Inhalte und Argumente analysieren und verstehen (1. Aufl.). Stuttgart: UTB.

Cacean, S. (2012). Ethische Aspekte von Cognitive Enhancement. In G. Spitzer & E. Franke (Eds.), Sport, Doping und Enhancement Ergebnisse und Denkanstöße (pp. 151–220). Köln: Sportverl. Strauß.

Caluwaerts, D., & Deschouwer, K. (2014). Building bridges across political divides: Experiments on deliberative democracy in deeply divided Belgium. European Political Science Review, 6(3), 427–450. https://doi.org/10.1017/S1755773913000179

Caluwaerts D., & Reuchamps M. (2014). Does Inter-group Deliberation Foster Inter-group Appreciation? Evidence from Two Experiments in Belgium. Politics, 34(2), 101–115. https://doi.org/10.1111/1467-9256.12043

Cullen, S., Elga, A., Fan, J., & Brugge, E. van der. (2018). Improving Analytical Reasoning and Argument Understanding: A Quasi-Experimental Field Study of Argument Visualization with First-Year Undergraduates. Npj Science of Learning, 3.

Davies, M. (2011). Concept mapping, mind mapping and argument mapping: What are the differences and do they matter? Higher Education, 62(3), 279–301. https://doi.org/10.1007/s10734-010-9387-6

Eftekhari, M., & Sotoudehnama, E. (2018). Effectiveness of computer-assisted argument mapping for comprehension, recall, and retention. ReCALL, 30(3), 337–354. https://doi.org/10.1017/S0958344017000337

Fairclough, I., & Fairclough, N. (2012). Political discourse analysis: A method for advanced students. London: Routledge.

Fischer, F., & Forester, J. (Eds.). (1993). The Argumentative Turn in Policy Analysis and Planning. Durham, N.C: Duke University Press Books.

Fisher, A. (2004). The Logic of Real Arguments (2nd ed.). New York: Cambridge University Press.

Freeman, J. B. (1991). Dialectics and the macrostructure of arguments: A theory of argument structure.

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika, 53(3/4), 325–338. https://doi.org/10.2307/2333639

Hansson, S. O., & Hirsch Hadorn, G. (Eds.). (2016). The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty. Cham: Springer.

Harrell, Mara. (n.d.). Using Argument Diagramming Software to Teach Critical Thinking Skills. Retrieved from https://www.academia.edu/772355/Using_Argument_Diagramming_Software_to_Teach_Critical_Thinking_Skills

Harrell, Maralee. (n.d.). Argument diagramming and critical thinking in introductory philosophy. Higher Education Research &amp; Development, 30(3), 371–385.

Krippendorff, K. H. (2012). Content Analysis: An Introduction to Its Methodology. SAGE Publications Inc.

Lippi, M., & Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. ACM Trans. Internet Technol., 16(2), 10:1–10:25. https://doi.org/10.1145/2850417

Neuendorf, K. A. (2002). The Content Analysis Guidebook. SAGE Publications Inc.

Reed, C., Walton, D., & Macagno, F. (2007). Argument diagramming in logic, law and artificial intelligence. The Knowledge Engineering Review, 22(01), 87.

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. International Journal of Computer-Supported Collaborative Learning, 5(1), 43–102. https://doi.org/10.1007/s11412-009-9080-x

Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring Political Deliberation: A Discourse Quality Index. Comparative European Politics, 1(1), 21–48. https://doi.org/10.1057/palgrave.cep.6110002

Titscher, S., Meyer, M., Wodak, R., & Vetter, E. (2000). Methods of Text and Discourse Analysis: In Search of Meaning (First edition). London ; Thousand Oaks Calif.: SAGE Publications Ltd.

Walton, D. (1996). Argumentation Schemes for Presumptive Reasoning (1st ed.). Routledge.

Walton, D. N. (2006). Fundamentals of critical argumentation. Cambridge: Cambridge University Press.