
Discourse Maps — Feature Encoding for the Analysis of Verbatim Conversation Transcripts

MENNATALLAH EL-ASSADY AND ANNETTE
HAUTLI-JANISZ

4.1 Introduction

This chapter reports on work that extracts information about linguistic features from political deliberations in order to make the deliberative quality of political dialog measurable.¹ With *Discourse Maps*, a dynamic visualization that is tailored to both the requirements of the data and the theoretical framework on measuring deliberative quality as articulated within political science (Gold et al. (2016)), we showcase how Visual Analytics can combine theory-driven (top-down) analysis with a data-driven (bottom-up) view on the data. Unlike the Zhao et al. (this volume) paper, we do not work solely on the basis of discourse relations, but extract a plethora of relevant linguistic features and visualize these according to their type and contribution to the deliberative nature of the dialog.

Our system has the potential to provide political scientists, linguists, stakeholders in the debate or the general public with a visualized rep-

¹We thank our collaborators Valentin Gold, Miriam Butt, Katharina Holzinger and Daniel A. Keim for valuable discussions. The work conducted in this paper was supported by the Deutsche Forschungsgemeinschaft (FOR2111, ‘Questions at the Interfaces’) and the VolkswagenStiftung under grant 92182.

resentation of the discourse, which can be employed for the comparison of discourse patterns between speakers, speaker parties, and different sequences.

The data underlying our work are verbatim transcripts of natural language discourse in the political sphere, a type of data that has gained momentum with the increasing availability of such resources. A particular interest lies in debates, i.e., argumentative discourse that is characterized by the interaction of multiple interlocutors who try to win a discussion on a controversial topic or convince the other participants. Verbatim transcripts of such discourses capture the rapid exchange of opinions, arguments, and information between interlocutors and thus, establish a rich data source for analysis. At the same time, this type of data presents challenges to the automatic processing of language: fragmented constructions, interruptions, filled pauses (‘uhm’, ‘mh’), speech repairs, dialect, and transcription errors require a robust machinery that yields reliable results.

In order to ground our approach in theoretical work in political science, we work with the theoretical framework articulated by our political science partners, who analyze political deliberation by means of four high-level dimensions [Gold and Holzinger \(2015\)](#), namely, **(1)** ‘Argumentation & Justification’, **(2)** ‘Accommodation’, **(3)** ‘Participation’, and **(4)** ‘Atmosphere & Respect’. Using tailored *micro-linguistic discourse features* we operationalize these dimensions and make them measurable. In total, we have currently computed, together with our domain experts, a set of 53 relevant discourse features for verbatim text in two languages (English, German).

Our contribution in this area is the following: We present a robust, hybrid system that pairs shallow text mining with linguistically motivated discourse analysis in noisy data, generating a rich set of micro-linguistic features that constitutes communication in the domain. Secondly, we introduce a novel visual design that rigidly maps all relevant aspects of communication (according to the deliberation framework) onto a glyph-based representation within the Discourse Maps, making all levels of a debate (starting with a single turn, all the way to an aggregation of all turns of a speaker/ within a topic), instantly comparable with respect to the analyzed features.

This paper proceeds as follows: We first lay out the necessary background, namely relevant work in discourse processing and visualization (Section [4.2](#)). We then present the computational linguistic analysis with both shallow text mining and the deeper, more linguistically motivated annotation (Section [4.3](#)). The annotation scheme and its encoding in Discourse Maps is discussed in Section [4.4](#), followed by the discussion

of the data structure modeling and the visualization design in Section 4.5. We then present a use case, where Discourse Maps are used to shed light on a real debate scenario, namely the so-called S21 arbitration, a public arbitration process in the German city of Stuttgart in 2010 (Section 4.6). Section 4.7 provides a two-level evaluation of Discourse Maps. Section 4.8 concludes the paper.

4.2 Background

Our work is rooted in the areas of discourse processing and Visual Analytics. This section highlights the relevant related-work and literature in both areas, building the background for the design and implementation of our Discourse Maps approach.

Discourse Processing Natural Language Processing (NLP) of discourse data is as varied as the type of data underlying it: An important area deals with the automatic annotation of discourse relations, i.e., relations between segments in the text. Those are annotated in different granularity and style in frameworks such as Rhetorical Structure Theory (Mann and Thompson 1988) or Segmented Discourse Representation Theory (Asher and Lascarides 2003). In English, the majority of work is based on landmark corpora such as the Penn Discourse Treebank (PDTB; Prasad et al. 2008). In German, the parsing of discourse relations has only lately received increasing attention (Versley and Gastel 2013, Stede and Neumann 2014, Bögel et al. 2014).

Another strand of research is concerned with dialogue act annotation, to which end several annotation schemes have been proposed (e.g., Bunt et al., 2010; inter alia). Those have also been applied across a range of German corpora (Jekat et al. 1995, Zarisheva and Scheffler 2015). Another area deals with the classification of speaker stance, for instance regarding personality (Mairesse et al. 2007), agreement and disagreement (Sridhar et al. 2015) or politeness (Danescu-Niculescu-Mizil et al. 2013).

With Discourse Maps, we provide the first discourse analysis pipeline which extracts a multitude of discourse features from naturally occurring dialogue data in parallel. This is done with hybrid technology: shallow text mining extracts surface-structure patterns in the discourse such as sentence complexity, interruptions and filler words (Section 4.3.1). This is complemented by a linguistically informed rule-based approach for disambiguating and annotating linguistic information such as discourse relations, speech acts, emotion, modality and rhetorical framing (Section 4.3.2).

In order to work with a fine-grained structure of the discourse, we

divide the text in smaller units of analysis, namely the *discourse unit*. While there is no consensus in the literature on what exactly these discourse units have to contain, it is generally assumed that each describes a single event (Polanyi et al. 2004). Following Marcu (2000), we term these units *elementary discourse units* (EDUs). For Discourse Maps, we aggregate the information of all EDUs on the level of the speaker turn (for more details on the aggregation see Section 4.5).

Visual Analytics Text is an inherently multimodal data source, comprised of many information channels for analysis. In particular, conversations and debates encompass a broad spectrum of information due to the diversity of their dynamics and the ambiguity of their language. Visual Analytics techniques can reveal such dynamics and enable an extensive analysis of the different aspects of discourse. One of the first examples to model the social interactions in chat systems was Chat Circles (Donath and Viégas 2002). Other approaches are GroupMeter (Leshed et al. 2009), Conversation Clusters (Bergstrom and Karahalios 2009), Trains of Thought (Shahaf et al. 2012), and MultiConVis (Hoque and Carenini 2016). The VisArgue framework (El-Assady et al. 2017a) introduced specialized visualization techniques for a faceted analysis of conversational text, most notably, the Lexical Episode Plots (Gold et al. 2015), ConToVi for mapping a conversation to a Topic Space View (El-Assady et al. 2016), NEREx for exploring named entity relationships (El-Assady et al. 2017b), the Argumentation Feature Alignment Visualization (Jentner et al. 2017), and ThreadReconstructor for untangling reply chains (El-Assady et al. 2018a). However, most of these approaches are not designed to give a full overview of discourse features and do not allow for the *finger-printing* of turns, speakers, or topics in a discourse. To achieve this, Discourse Maps utilizes the design principles and guidelines for glyph-based visualizations, as outlined by Borgo et al. (2013).

A more recent survey on glyph-based visualizations has been recently provided by Fuchs et al. (2017). They systematically reviewed the results of experimental studies on data glyphs, suggesting that the background of a glyph might not influence its readability and that aligning the glyph design to the mental models of the users enhances the understanding of its underlying data. They also express caution about encoding too many data points into a single glyph as it negatively affects search. Hence, in this work, we explore the trade-off between a stable mental model and the information density of the visualization, resulting with an interactive (turning data points on and off) representation that uses a strict visual mapping of domain knowledge.

Another area of related work are dense-pixel displays. A prominent example is the work by Keim and Oelke (2007) on literature fingerprinting. In this work, pixel-based small-multiples are used for encoding measures extracted from text data. In contrast to the guidance provided for sophisticated glyph design, dense-pixel displays do not limit the number of data points encoded in one visual object. Our proposed visualization uses small, simple glyphs as pixels that are arranged using techniques comparable to the ones well known in dense-pixel displays.

4.3 Computational Linguistic Analysis

Discourse Maps are based on a hybrid set of features that are extracted via shallow text mining techniques (Section 4.3.1) or via a more in-depth, linguistically motivated annotation system (Section 4.3.2). In the following, we discuss both methods based on an sample feature set.

4.3.1 Shallow Text Mining

With shallow text mining the aim is to capture properties of the discourse that do not necessarily depend on context or a deep analysis of linguistic structure. One such property is the *average sentence complexity*, which gives us an approximation as to how complex the sentence structure of a particular speaker (or speaker position) is. To that end we count the number of EDUs in each sentence of a speaker turn and divide it by the number of sentences.

Another relevant measure is whether particular turns are *interruptions*. Given the postulate of deliberative communication to be respectful, this feature allows us to detect phases in the debate which are heated and do not adhere to deliberative standards. To determine this, we count the number of content-bearing words in a speaker turn (e.g., nouns) and check whether it exceeds a user-defined threshold, marking the turn as an interruption if it does not. In addition to some turns not significantly contributing to the conversation, we also count the *number of filler words* of each turn. With this step we do justice to the type of data we are dealing with: spontaneous, natural language speech is noisy and many turns (or parts of them) merely signal backchanneling (that the speaker is paying attention and possibly agreeing or disagreeing). These are defined using dictionaries (e.g., ‘um’, ‘hm’, ‘ah’) and regular expressions to capture variation in the transcriptions (e.g., ‘uum’, ‘hmm’). Furthermore, we also consider statistical measures and features based on the content of the text, as determined by topic modeling algorithms developed by us (El-Assady et al. 2018b). In the following, these features are described in more detail.

Statistical Measures In political science, the use of statistical measures is ubiquitous. Such measures inform models for empirical studies and are taken as essential for understanding dynamics in conversations (Gold and Holzinger 2015). In our work we implemented three measures that capture commonly studied phenomena in discourse analysis. The first two rely on a moving window approach to assess the context of a speaker turn. Hence, based on a user-defined window size (defined by the tuple (p, f) for the number of previous and the number of following turns, respectively), we regard for the neighborhood of each turn one measure, as for example, for the speaker of the turn his/her *expected probability to speak* or the *moving Gini* that determines the turn-taking distribution based on the Gini Coefficient (Ceriani and Verme 2012). The third statistical measure we included determines the eloquence of speakers, measuring the diversity of their vocabulary based on the *Maas index*, as outlined by (McCarthy and Jarvis 2007).

Topic Modeling Content analysis is one of the major tasks when dealing with discourse data. Topic modeling algorithms automatically segment the turns of a discourse into thematically coherent groups. We, thus, rely on their output to aggregate the turns into a set of topics, but also derive measures based on this segmentation. These measures determine how a particular turn is situated, given the topic distribution of the whole corpus. To define the features we extract using the topic modeling results, we select turns to consider for the similarity calculation to a turn utr_i at hand, based on three distinct factors:

speaker {self, all}: *turns that are from the same speaker as utr_i*
vs. all turns.

topics {self, all}: *turns that deal with the same topics as utr_i*
vs. all turns.

position {previous, following}: *turns that have come before utr_i*
vs. turns that have come after utr_i .

Hence, we compute a set of turns to consider based on these factors, as exemplified in the following scheme; for the similarity to all previous turns of the speaker of the selected turn utr_i , we denote: $sim_{top_{all}, spe_{self}, pos_{prev}}(utr_i)$. Note, that the similarity calculation between two turns is modular and can be defined by users — by default the cosine similarity is selected.

This method enables the segmentation of the corpus in various forms, and, in turn, allows us to define useful features based on ratios of calculated segments. In total, we define five novel features.

(1) *Topic shift* describes whether the topic of the turn advances the conversation, or whether the turn is continuing with an already estab-

lished topic. It is defined as: $Topic\ Shift_{utr_i} = \frac{sim_{top_{all},spe_{all},pos_{prev}}(utr_i)}{sim_{top_{all},spe_{all},pos_{follow}}(utr_i)}$.

(2) *Self previous recurrence* describes the relative amount of content recurrence a selected turn has to previous turns from the same speaker, considering all previous turns; i.e., how much this turn is a repetition of what this person has already said. It is defined as: $Self\ Previous\ Recurrence_{utr_i} = \frac{sim_{top_{all},spe_{self},pos_{prev}}(utr_i)}{sim_{top_{all},spe_{all},pos_{prev}}(utr_i)}$.

(3) *Self following recurrence* is the counterpart to the self previous recurrence. It describes the relative amount of content recurrence a selected turn has to the following turns from the same speaker, considering all previous turns. This can be seen as a measure of how much influence this particular speaker turn will have on the remainder of the conversation. It is defined as:

$$Self\ Following\ Recurrence_{utr_i} = \frac{sim_{top_{all},spe_{self},pos_{follow}}(utr_i)}{sim_{top_{all},spe_{all},pos_{follow}}(utr_i)}.$$

(4) *Self recurrence shift* is a measure of the relation between the self previous recurrence and the self following recurrence. Hence this is a measure of whether the recurrence of the turn is a progressive one or not, i.e., whether this particular turn is more relevant to the preceding part (for example, as a summary) or whether it will become more relevant to the following part of the conversation (for example, through setting new agenda topics). It is defined as: $Self\ Recurrence\ Shift_{utr_i} = \frac{Self\ Previous\ Recurrence_{utr_i}}{Self\ Following\ Recurrence_{utr_i}}$.

(5) *Topic persistence* describes whether the speaker of a particular turn is persistent with regard to the topic of that turn or not. This measure is defined as $Topic\ Persistence_{utr_i} = \frac{so(utr_i)/sa(utr_i)}{ao(utr_i)/aa(utr_i)}$, through the following four similarities:

$$\begin{aligned} so(utr_i) &= sim_{top_{self},spe_{self},pos_{prev+follow}}(utr_i); \\ sa(utr_i) &= sim_{top_{all},spe_{self},pos_{prev+follow}}(utr_i); \\ ao(utr_i) &= sim_{top_{self},spe_{all},pos_{prev+follow}}(utr_i); \\ aa(utr_i) &= sim_{top_{all},spe_{all},pos_{prev+follow}}(utr_i). \end{aligned}$$

In total, these measures indicate how a turn is contributing to the general content of a discourse and in which capacity a given speaker is involved. This and other shallow-linguistic features build the basis to a better understanding of the role of particular speaker turns in conversations and have proven to be insightful indicators of characteristic dynamics in political debates (Gold et al. 2016).

4.3.2 Linguistic Annotation

In contrast to the shallow text mining, the linguistic annotation pipeline extracts discourse features based on a comparatively deep linguistic analysis. The annotation system is specifically designed to deal with noisy transcribed natural speech which contains ungrammatical/fragmented constructions, backchanneling ('hm', 'ah') and interruptions. It is based on a linguistically informed, hand-crafted set of rules that deals with the disambiguation of explicit linguistic markers and the identification of their spans and relations in the text (for more details on the general structure of these rules see [Bögel et al. 2014](#)).

The system analyzes several layers of information. With respect to *discourse relations*, we annotate spans as to whether they represent: reasons, conclusions, contrasts, concessions, conditions or consequences ([Bögel et al. 2014](#)). For German, we rely on the connectors in the Potsdam Commentary Corpus ([Stede and Neumann 2014](#)), for English we use the PDTB-style parser by [Lin et al. \(2014\)](#). In order to identify relevant *speech acts*, we annotate speech act verbs signaling agreement, disagreement, arguing, bargaining and information giving/seeking/refusing. In order to gauge *emotion*, we use EmoLex, a crowdsourced emotion lexicon ([Mohammad and Turney 2010](#)) available for a number of languages, plus our own curated lexicon of politeness markers. With respect to *event modality*, we take into account all modal verbs and adverbs signaling obligation, permission, volition, reluctance or alternative. Concerning *epistemic modality* and speaker stance we use modal expressions conveying certainty, probability, possibility and impossibility. Finally, we add a category called *rhetorical framing* ([Hautli-Janisz and Butt 2016](#)), which accounts for the illocutionary contribution of German discourse particles. Here we look at different ways of invoking Common Ground, hedging and signaling accommodation in argumentation, for example.

Preprocessing We first divide up all turns into EDUs. For German, we approximate the assumption made by [Polanyi et al. \(2004\)](#) by inserting a boundary at every punctuation mark and every clausal connector (conjunctions, complementizers). For English we rely on clause-level splitting of the Stanford PCFG parser ([Klein and Manning 2003](#)) and create EDUs at the SBAR, SBARQ, SINV and SQ clause levels. The annotation is performed on the level of these EDUs, therefore relations that span multiple units are marked individually at each unit.

We were not able to use an off-the-shelf parser for German. For instance, an initial experiment using the German Stanford Dependency parser ([Rafferty and Manning 2008](#)) showed that 60% of parses are

incorrect due to interruptions, speech repairs and multiple embeddings. We therefore hand-crafted our own rules on the basis of morphological and POS information from DMOR (Schiller 1994). For English, we used the POS tags from the Stanford parser.

Disambiguation Many of the crucial linguistic markers are ambiguous. We developed hand-crafted rules that take into account the surrounding context to achieve disambiguation. Important features include position in the EDU (for instance for lexemes which can be discourse connectors at the beginning of an EDU but not at the end, and vice versa) or the POS of other lexical items in the context. Overall, the German system features 20 disambiguation rules, the English one has 12.

Relation Identification After disambiguation is complete, a second set of rules annotates the spans and the relations that the lexical items trigger. In this module, we again take into account the context of the lexical item. An important factor is negation, which in some cases reverses the contribution of the lexical item, e.g., in the case of ‘possible’ to ‘not possible’.

With respect to discourse connectors, for instance the German causal markers *da*, *denn*, *darum* and *daher* ‘because/thus’, we only analyze relations within a single speaker turn, i.e., relations that are expressed in a sequence of clauses which a speaker utters without interference from another speaker. As a consequence, the annotation system does not take into account relations that are split up between turns of one speaker or turns of different speakers. For causal relations (reason and conclusion spans), the system performs with an F-score of 0.95 (Bögel et al. 2014).

Taken together, shallow text mining and linguistic processing yields a set of 53 features that encode various properties of the debate. All of them serve as operationalizing features for analyzing communicative strategies in deliberative communication. For Discourse Maps, we combine all features into an annotation scheme with dimensions that are well motivated from the viewpoint of political science. This annotation scheme serves as the mental model for Discourse Maps and is discussed in the following.

4.4 Annotation Scheme

The framework used to analyze deliberative communication comes from political science and comprises four larger dimensions, namely Argumentation & Justification, Accommodation, Atmosphere & Respect

and Participation (Gold and Holzinger 2015). This model serves as the backbone for the annotation scheme and is populated with the features from the shallow text mining and the linguistic annotation pipeline. It also defines the design structure of the Discourse Maps visualization in that the map is divided into four quadrants — illustrated by the template in Figure 1. The annotation scheme and its relation to Discourse Maps is discussed in more detail in the following.

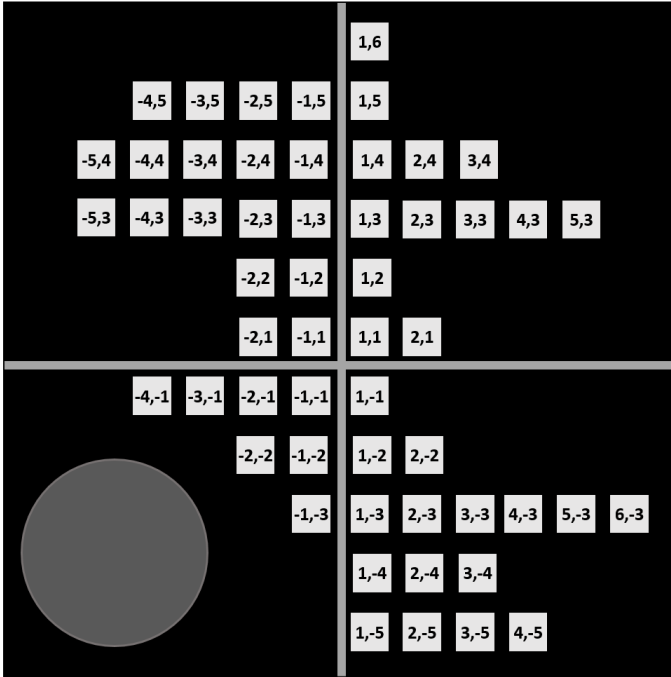


FIGURE 1 Index map, with each index position indicating the position of a glyph in the Discourse Map. Overall, the map shows the four quadrants, depicting the four deliberation dimensions. In addition, all quadrants combine 19 subdimensions, that, in turn, span 53 measures. The circular shape on the bottom left is scaled to the length of the underlying turns, indicating the relative size of the underlying text.

4.4.1 Atmosphere & Respect

The first subdimension, Atmosphere & Respect, is encoded in the upper right (Northeast; NE) quadrant of the Discourse Map. This dimension encompasses features that represent central standards in deliberative

communication, namely respect, conscientiousness and civility (Gerhards 1997, Fishkin and Luskin 2005; inter alia). For a good deliberative process, Landwehr and Holzinger (2010) also require a conversational back-and-forth in the debate instead of a passive listening to monologues by the interlocutors.

TABLE 1 Dimension “Atmosphere & Respect”

Subdimension	Feature/Measure	Index	Type
Emotion	Emotion Count	1, 1	NUM CONT
	Emotion Relation	2, 1	NUM BIPOL
Interruptions	Interruption	1, 2	BINARY
Responsiveness	Topic Shift	1, 3	NUM BIPOL
	Self Previous Recurrence	2, 3	NUM CONT
	Self Following Recurrence	3, 3	NUM CONT
	Self Recurrence Shift	4, 3	NUM BIPOL
Conventional Politeness	Topic Persistence	5, 3	NUM BIPOL
	Politeness	1, 4	NUM CONT
	Impatience	2, 4	BINARY
Face Issues	Unobtrusiveness	3, 4	BINARY
	Resignation Acceptance	1, 5	NUM CONT
Sentiments	Sentiment	1, 6	NUM BIPOL

In order to group these diverse aspects in a meaningful way, we introduce subdimensions, namely “Emotion”, “Interruptions”, “Responsiveness”, “Conventional Politeness”, “Face Issues” and “Sentiment” (see Table 1). Each of these subdimensions is represented as an individual row, with each row consisting of square glyphs that represent the individual features (see second column in Table 1). The exact position of these glyphs in the quadrant is described via the index in the third column in Table 1. The underlying feature can be revealed by mousing over the glyphs in the visual interface, for instance the amount of positive or negative emotion is represented by the glyph in position (1,1) in the Discourse Map, the amount of interruptions is captured in position (1,2). The fourth column in Table 1 encodes how the different discourse features are measured. This defines the type of color-coding, a property of Discourse Maps that is discussed in detail in Section 4.5.

4.4.2 Argumentation & Justification

The Argumentation & Justification dimension is situated in the lower right (Southeast; SE) quadrant of the Discourse Map. It contains the subdimensions listed in Table 2: “Information Certainty”, “Reason-giving”, “Event Modality”, “Common Ground” and “Information Exchange”. “Information Exchange” is relevant in the Dimension “Argumentation & Justification” because participants in a deliberative

process should argue and justify their positions, consequently we expect structures where information is provided, sought or refused. The certainty with which this information is provided is subsumed under the subdimension “Information Certainty”: Here we use the scale of Lassiter (2010) which maps expressions of epistemic modality on a scale from 0 (impossible) to 1 (certain).

We further expect argumentative structures, in particular causal argumentation with premises and/or conclusions (subsumed in the “Reason-giving” subdimension). Deontic modals, i.e., those modals that denote how the world should be according to norms or speaker desires, e.g. ‘have to’ and ‘should’, are encoded in the subdimension “Event Modality”. The “Common Ground” originates in a linguistic concept, whereby interlocutors share an abstract knowledge space (Stalnaker 2002). In German, the Common Ground is frequently referred to via particles, for instance *ja* ‘lit. yes’, a linguistic category that is highly frequent in spontaneous speech — speakers use these relate themselves or their contributions to the shared knowledge of the discussion partners (Zimmermann 2011).

TABLE 2 Dimension “Argumentation & Justification”

Subdimension	Feature/Measure	Index	Type
Information Certainty	Epistemic Value	1, -1	NUM BIPOL
Reason-giving	Reason	1, -2	NUM CONT
	Conclusion	2, -2	NUM CONT
Event Modality	Obligation	1, -3	BINARY
	Volition	2, -3	BINARY
	External Constraint	3, -3	BINARY
	Permission	4, -3	BINARY
	Alternative	5, -3	BINARY
	Reluctance	6, -3	BINARY
Common Ground	Common Ground (CG)	1, -4	BINARY
	Reject CG	2, -4	BINARY
	Activate CG	3, -4	BINARY
Information Exchange	Information Giving	1, -5	BINARY
	Elucidation	2, -5	BINARY
	Information Seeking	3, -5	BINARY
	Information Refusing	4, -5	BINARY

As in the Atmosphere & Respect dimension above, the position of the glyphs that represent those features in the Discourse Map are given in the third column of Table 2. For instance, premise units are represented by the glyph in position (1,-2), conclusions are encoded with the glyph in position (2,-2).

4.4.3 Participation

The lower left (Southwest; SW) quadrant of the Discourse Map represents the Participation dimension, a dimension that measures the involvement of individual speakers in the discourse. This is operationalized by looking at the “Equality of Speaker Capabilities” (measured by features that indicate the eloquence of speakers), the “Equality of Speaker Participation” (measured by comparing the number of contributions of one speaker to those of the other interlocutors) and “Topic Comprehensiveness” (measured by the network density of all thematic relations of a speaker turn) (see Table 3). As in the dimensions above, each subdimension is encoded as one row and each feature is represented by one glyph. Again, the position of the glyphs in the Discourse Map is shown in the index column in Table 3.

TABLE 3 Dimension “Participation”

Subdimension	Feature/Measure	Index	Type
Equality of Speaker Capabilities	Sentence Complexity	-1, -1	NUM BIPOL
	Maas Index	-2, -1	NUM BIPOL
	Filler Words	-3, -1	NUM CONT
	Stalling	-4, -1	BINARY
Equality of Speaker Participation	Exp Prob to Speak	-1, -2	NUM BIPOL
	Moving Gini Index	-2, -2	NUM BIPOL
Topic Comprehensiveness	Network Density	-1, -3	NUM BIPOL

4.4.4 Accommodation

Another dimension with a large array of subdimensions is Accommodation, situated in the upper left (Northwest; NW) quadrant in the Discourse Map and detailed in Table 4. In this dimension we capture all linguistic structures that are relevant in negotiation situations, such as instances that signal agreement or disagreement, hint at conditions that need to be fulfilled in order to come to an agreement and are used to achieve some kind of consensus. In total we have five subdimensions: “Condition”, “Agreement vs. Disagreement”, “Agreement”, “Disagreement” and “Arguing vs. Bargaining”. In “Condition”, we capture conditional discourse relations triggered for instance by ‘if ... then’ constructions. In “Agreement”, we combine information contributed by discourse particles, for instance the agreement information triggered by sentence-initial ‘yes’, and speech act verbs signaling agreement (e.g. *bestürworten* ‘to support’). In the subdimension “Disagreement”, the information triggered by particles and speech act verbs (e.g. *bestreiten* ‘to deny’) is combined with conjunctions such as ‘instead of’ signaling contrastive discourse relations. In “Agreement vs. Disagreement”, we

set the two measures of “Agreement” (-2,3) and “Disagreement” (-3,4) in relation. On the one hand, we count the absolute number of these two measures (-1,2). On the other hand, we set them in relation (-2,2). In a similar manner, in “Arguing vs. Bargaining”, we subsume speech acts of arguing and bargaining (for instance units governed by ‘to justify and ‘to resign’, respectively). The measures “Negotiation Count” and “Negotiation Relation” describe on the one hand the absolute count (how much is on a scale), and on the other, the relation of “Arguing” vs. “Bargaining” (is the scale tipped to the one or the other side). Note that having count and relation measures together reveals a clearer picture of a phenomenon, i.e., a relation might show a 2:3 scale but only with the count can we distinguish between 20:30 vs. 2000:3000.

TABLE 4 Dimension “Accommodation”

Subdimension	Feature/Measure	Index	Type
Condition	Condition	-1, 1	NUM CONT
	Consequence	-2, 1	NUM CONT
Agreement vs. Disagreement	Arrangement Count	-1, 2	NUM CONT
	Arrangement Relation	-2, 2	NUM BIPOL
Agreement	Consensus	-1, 3	BINARY
	Agreement	-2, 3	BINARY
	Consensus Willing	-3, 3	BINARY
	Minimal Consensus	-4, 3	BINARY
	Concession	-5, 3	NUM CONT
Disagreement	Opposition	-1, 4	NUM CONT
	Dissent	-2, 4	BINARY
	Disagreement	-3, 4	BINARY
	Activate Opposition	-4, 4	BINARY
	Contrast	-5, 4	BINARY
Arguing vs. Bargaining	Negotiation Relation	-1, 5	NUM BIPOL
	Negotiation Count	-2, 5	NUM CONT
	Arguing	-3, 5	BINARY
	Bargaining	-4, 5	BINARY

After having laid out the linguistic groundwork on which Discourse Maps are based, we now discuss the visualization design in more detail, in particular regarding the modeling and visual representation of different data structures.

4.5 Visualization Design

In order to achieve a suitable visual representation of the data model at hand, we conducted several user studies and sketching sessions, going through a set of eight different prototypes for the Discourse Maps. A selection of five intermediate prototypes is discussed in Section [4.5.1](#). This section elaborates on the design rationale of the presented visual-

ization and the underlying data structure modeling.

Design Requirements The visual design of such a static, yet complex model has to fulfill a rigid set of requirements. First, this visualization is designed with a strict scheme of data relations in mind. The given hierarchy of dimensions, subdimensions, and measures defines a tight mold which the visual design has to follow. During our design process we experimented with different complexity levels of the visualization and came to realize that this visualization should not try to hide the complexity of the data model, as it is used by analysts as a “brain dump” to rid themselves from remembering the data model and rather focus on the arising patterns for analysis. Hence, a second design requirement was to construct a stable visual mapping that aims at representing all data model relations (taking into account that such a design comes with a learning curve that needs to be absolved). The third design requirement is that this visualization should enable a broad range of analysis tasks through allowing users to define the measures they are interested in focusing on analyzing. Lastly, and most importantly, the visual representation of a single turn should be comparable to the representation of a group of turns to enable comparability across levels of granularity, e.g., topics, speakers, parties, days, etc.

Analysis Tasks Based on the studies we conducted for our requirement analysis, we derived a set of analytical tasks that users intend to perform using the Discourse Maps. These include, most notably, the following tasks:

- Exploration of Measure Relations and Patterns
- Theory-Driven Hypothesis Generation for Expected Relations
- Interactive, Data-Driven Hypothesis Verification
- Comparative Analysis across Turns, Speakers, Parties, Topics
- Refinement of the Deliberation Theory based on Findings

Such an analysis of analytical tasks enabled an effective design of the Discourse Maps and, in turn, led to domain insight and a better understanding of discourse dynamics.

Data Structure Modeling In order to ensure that all features are mapped adequately, we subdivided them into three types.

1. *Numerical Continuous* features are normalized to a scale from 0 to 1 and usually represent relative counts, e.g., the amount of agreement and disagreement particles and speech acts (cf. Table 4, -1,2).
2. *Numerical Bipolar* features are mapped to a scale between -1 and 1 and are typically showing a diverging measure, e.g., the relation

between agreement and disagreement particles and speech acts (cf. Table 4 -2,2).

3. *Binary* features are either 0 or 1 and indicate whether an attribute exists or not, e.g., whether a turn contains an agreement particle or speech act or not (cf. Table 4 -2,3).

Note, that these scales are defined on a speaker turn level. When multiple turns are aggregated, the aggregated feature score is mapped back to the same range, with the exception of binary features that are mapped to a continuum from 0 to 1, instead of a discrete scale.

4.5.1 Design Iterations

As previously mentioned, the current visualization design is the result of an iterative process that incorporated the expert feedback into the evolution of the *Discourse Map* prototypes. Figure 2 depicts five out of eight prototypes from the different iterations. Some of the designs mapped a selection of the most important dimensions to the shape of a glyph (e.g., Iterations 1 & 3). Other designs positioned the feature dimensions into a defined structure (e.g., Iterations 4 & 5). Again others used a global visual layout to enhance comparability (e.g., Iterations 2 & 5). However, these designs did not conform to the mental model of the domain experts and, in turn, did not facilitate the externalization of their domain knowledge.

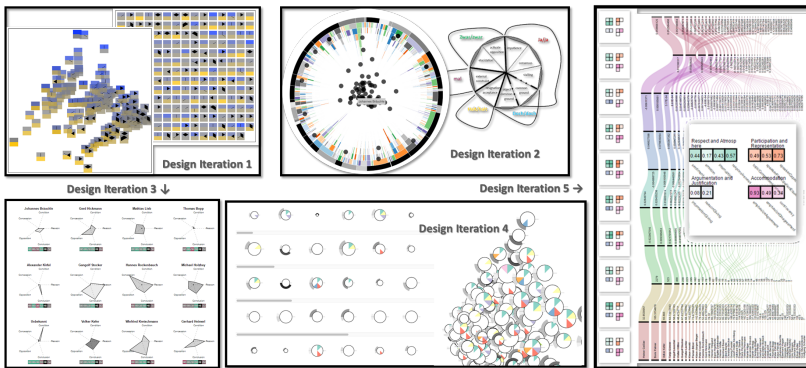


FIGURE 2 Five prototypes from previous design iterations.

Following our four design requirements, we attempted to find the most suitable visual mapping of the data scheme of our domain experts, while reducing visual complexity and hiding unnecessary information to a second detail level. Iterations 1, 3, 4 & 5, therefore, used a defined

glyph structure with up to 12 dimensions to encode the required information (requirement 1). However, our intermediate evaluations showed that these mappings were not sufficient to encode all relations our users care about for the high-level analysis (requirement 2). Increasing the level of visual complexity by encoding more dimensions into the glyph did not scale for some of the designs (e.g., Iterations 1, 3 & 4). In addition, these designs dictated a strict order of the dimensions and were not flexible to accommodate multiple analysis tasks, for example through selecting specific dimensions to focus on (requirement 3). Alternative mappings that used the whole screen space to show the relation between feature dimensions (e.g., Iterations 2 & 5) did not allow for comparability across text granularity levels (requirement 4).

Hence, the design of the Discourse Maps as presented here evolved through a long-term design process. After the creation of each prototype, we conducted an expert evaluation, followed by a discussion of design choices and a sketching session. These sketching sessions were usually the starting point for refining a given prototype or, alternatively, beginning a new design iteration. Involving the domain experts into the visualization design process allowed us to incorporate their understanding of deliberation theory into the visualization design and paved the way for creating a (sophisticated) visual encoding that truly externalized their domain understanding, as described in the following section.

4.5.2 Discourse Maps

Given the multimodality of the computed feature set and the derived requirements and tasks, the visual design of our approach has to consider three important principles. First, the visualization needs to preserve and represent the mental model of deliberative communication as defined by political science. Second, the hybrid set of features needs to be mapped onto visual variables that are intuitive and enhance the recognition of the information. Third, the comparability of different aggregation levels needs to be ensured, i.e., the user has to be able to compare the same information across different levels of detail, for instance for a single turn, for individual speakers, for individual topics or for speaker positions.

Accounting for all these principles, our approach allows the user to draw inferences regarding the progress of the debate, speaker behavior and argumentative strategies in large amounts of deliberative communication at a glance. Discourse Maps are designed as glyph-based, small multiples that encode all relevant information in an index map, as highlighted in Figure [1](#). These small multiples can be regarded as fin-

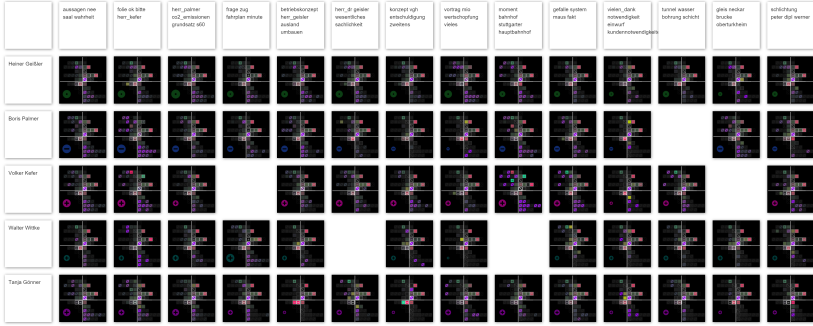


FIGURE 3 *Topics* (columns) by *Speakers* (rows) grid of Discourse Maps sorted to the highest number of turns in rows and columns.

gerprints of deliberative communication and are designed to enhance a recognition of the information, in compliance with the design guidelines of [Borgo et al. \(2013\)](#) (DG5: “Justify the choice of outcome measures in terms of their relevance to the objectives of the empirical study”). Hence, important dimensions for the data are highlighted using luminisence and position.

The template for a Discourse Map shown in [Figure 1](#) mirrors the four dimensions of deliberation proposed by our political science partners, with each quadrant representing one dimension: NW (Accommodation), NE (Atmosphere & Respect), SE (Participation), SW (Argumentation & Justification). Each subdimension is represented as a row and each annotation within a subdimension is represented as a small rectangular box, the so-called *feature-glyph*. Each subdimension is dynamically positioned nearer to the center the more often its annotation occurs in the data. In addition, each feature-glyph is positioned nearer to the coordinates the more often it occurs within its subdimension. This dynamic layout generation allows the creation of adaptable Discourse Maps depending on the underlying data. However, the internal layout of a map is stable for a given corpus to avoid confusion and to enable analysts to memorize layouts corresponding to their data. Furthermore, in order to show the average length of each unit of analysis — a turn, we include a small circular icon on the bottom left of the Discourse Map. This is scaled to the length of the underlying turns and indicates the relative size of the underlying text.

A Discourse Map represents one or more speaker turns, depending on the segmentation of the underlying discourse. For example, [Figure 3](#) shows a *topics X speakers* grid of Discourse Maps, i.e., each Discourse Map represents all turns a certain speaker has said. Such a grid-based

segmentation enables the generation of multiple views, alternating the aggregation based on *speakers*, *parties* (speaker positions), as well as, *topics*. Hence, grids such as *topics X parties*, *turns X speakers*, etc., can be created dynamically.

In addition to the dynamic generation of grids for the Discourse Maps, users can interactively select the dimensions, subdimensions, and features they would like to focus their analysis on. This is done through an index map (cf. Figure 1) that allows users to turn individual elements of the map on or off. A feature-glyph that is disabled is rendered in black. Furthermore, other interaction techniques are designed for supporting the analysis process, as described in Section 4.5.4.

4.5.3 Feature-Glyph Design

As described in the previous section, Discourse Maps represent individual measures as feature-glyphs. Each feature-glyph is a small rectangular box that is mapped to certain attributes related to the features presented. Figure 4 illustrates the design of a feature-glyph, mapping three values to a rectangular box.

First, to facilitate the localization, comparison, and distinction of glyphs, we have to take into account different types of data, as described in Section 4.5. These are represented using a shape in the middle of each feature-glyph. The *Numerical Continuous* type is based on frequency counts represented by a simple rectangle \square with no additional lines or shapes. For *Binary* occurrences (e.g., reason phrase present or not), the rectangle includes a diagonal line \square . The last type is *Numerical Bipolar* features, where we range from positive to negative values, e.g., for sentiment or emotion words. Since the relation of positive and negative occurrences is relevant, we assume that the normal state is the neutral box \square and indicate a drift to the positive \square or negative \square sides with

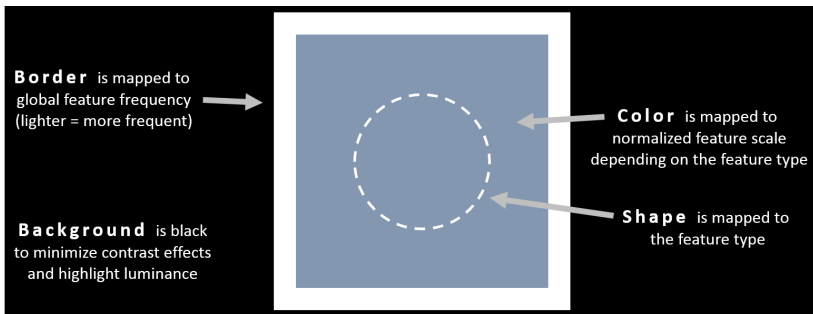





FIGURE 4 Feature-glyph design, utilizing border, color, and shape.


a plus or a minus sign, respectively.



FIGURE 5 Color mappings for the different data types.

Second, the normalized feature value of the particular measure at hand is represented using color. Here, the color scheme (as shown in Figure 5) differentiates between the three data types: *Binary* (a), *Numerical Continuous* (b), and *Numerical Bipolar* (c). These were chosen according to perceptive criteria that highlight the encoded values with the luminance of the color. All color scales were created using ColorCat (Mittelstädt et al. 2015). An example for the distinction of glyph types by the color scheme is shown here: . Here a sequence of three features is shown, consisting of two *Numerical Continuous* measures, followed by one *Numerical Bipolar* one. As noted above, the aggregation of *Binary* measures results in a continuous numerical scale, which is shown using an interpolation of the two ends of the the binary colormap (Figure 5a), resulting in different shades of purple, e.g., .

Third, the global frequency of each feature is represented by the border color. A white color can be understood as a feature that can be measured for all turns, while a dark gray border indicates a feature that is only present in a few turns within the whole corpus. Note that, throughout the feature-glyph design, we use luminance to indicate noteworthy phenomena. Hence, when a glyph has a black border and a black filling it fades into the background and does not disturb the analysis, e.g., . However, if a glyph has a light border and a dark filling, it will be noticed as an important feature that has a near-zero value for this particular Discourse Map. To compensate for contrast effects potentially caused by a border gradient, we calculate the minimally required number of pixels per glyph based on the model proposed by Mittelstädt et al. (2014).

The overall design of the feature-glyphs is tailored to facilitate their identification and comparison to enable global, as well as, local pattern detection. By using the border to highlight the global feature frequency, we can distinguish important features (i.e., relevant for the discourse at hand) from not so prominent ones. For example, for an instance of the two features  *Maas Index* (-2,-1) and *avg. Sentence Complexity* (-1,-1), we can see that the two features are *Numerical Bipolar*, with a negative value for the MI and a positive value for the SC. Furthermore,

we can detect that the SC has been measured for more turns in this particular discourse (shown by the lighter border color) and is, thus, potentially more important for the analysis.

4.5.4 Interactivity

The overall visual workspace around the Discourse Maps is tailored to the exploration and analysis of deliberation patterns across different layers of the discourse. The most basic layer visualizes discourse turns over time. Each turn is ordered sequentially and represented as a feature-glyph. This visualization helps determining deliberative segments within a discourse. In a second layer of analysis, the sequential order is visualized not with respect to the complete discourse but for each speaker separately. With this layer, the deliberative behavior of speakers is compared over time. This visualization supports the identification of deviant behavior of speakers. With the third visualization, as illustrated in Figure 3, patterns of speakers are compared between different topics in a *topics X speakers* grid. The top row shows the different topics, generated with an incremental hierarchical topic modeling algorithm (El-Assady et al. 2018c). Each column represents one speaker, each speaker turn is assigned to one topic. Blank spaces without a Discourse Map show that a speaker did not contribute to a specific topic. This visualization facilitates the detection of topics that are characterized by a particular pattern of deliberation.

Finally, the glyphs can be summarized and aggregated with respect to some given metadata, for instance with respect to different parties, as shown in Figure 6. This level of analysis is used for the case study in Section 4.6, where the deliberative patterns are investigated for the different parties of speakers.

Overall, this interactive segmentation enables users to adjust the discourse granularity and the generation of Discourse Maps to their respective analysis question.

Furthermore, to enable a focused analysis of certain aspects of communication using these complex glyphs, we designed a number of selection and filtering techniques, as well as, details-on-demand (hover to read specific value or click to enlarge the map) interactions. Together with the interactive aggregation of glyphs, the analysis of communication dynamics using our system can be utilized to answer a variety of questions with respect to deliberative communication.

4.6 Use Case

In order to showcase that Discourse Maps can be used to analyze discourse where a controversial topic is discussed between multiple in-

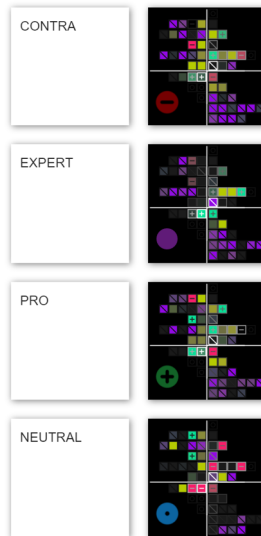


FIGURE 6 Feature-glyphs aggregated to speaker party.

terlocutors, we use the transcripts of Stuttgart 21 (henceforth S21)² a public arbitration process in the German city of Stuttgart, where a new railway and urban development plan caused a massive public conflict in 2010. The transcribed minutes consist of nine days of sessions, each lasting about seven hours with more than 70 participants. In total, the transcripts contain around 265,000 tokens in about 6,300 speaker turns. The aim of the use case is to show that the different speaker parties exhibit different discourse patterns, in particular regarding their argumentative patterns, their patterns regarding information giving and refusing and patterns of who leads or hinders the discourse.

The first entry point to the analysis of the S21 arbitration is through the analysis of the typical speaker patterns using the *Speaker Profiles*, as shown in Figure 7. This view gives a short biography for each speaker and displays a Discourse Map of their aggregated turns, as a summary to their contributions to the discourse. It also shows the party they belong to, i.e., the group to which their turns will be grouped in further aggregation steps.

There are four speaker parties in the S21 arbitration: the mediator Heiner Geißler (NEUTRAL), the proponents of the S21 project (PRO),

²Until October 2014 the transcripts were publicly available for download at <http://stuttgart21.wikiwam.de/Schlichtungsprotokolle>. A new, edited version of the minutes can be found here: <http://www.schlichtung-s21.de/dokumente.html>.

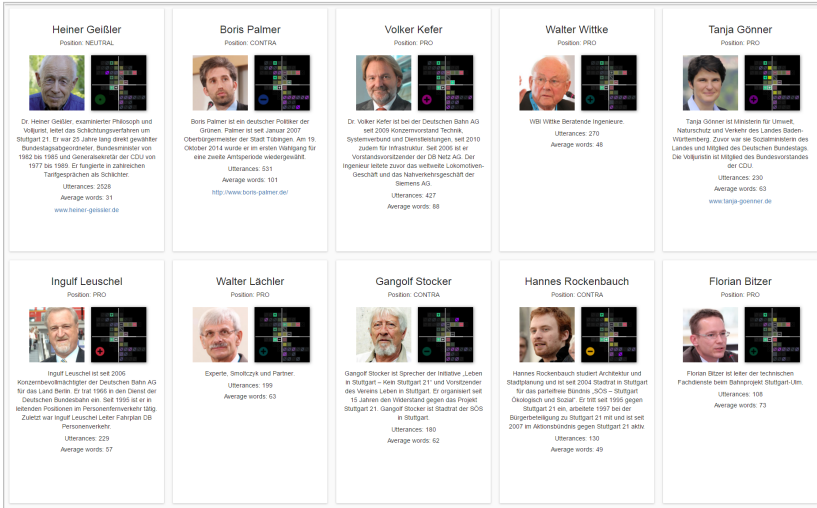


FIGURE 7 S21 *Speaker Profiles* of the speakers with the most turns.

the project opponents (CONTRA), and an independent group of experts (EXPERT). In order to allow the comparison of different speaker parties, we aggregate the Discourse Maps of all speakers based on their party affiliation, i.e., the more than 6,300 individual Discourse Maps (one for each speaker turn) are aggregated to only four Discourse Maps (see Figure 8). For comparing features in the discourse, individual glyphs in the Discourse Map are selected.

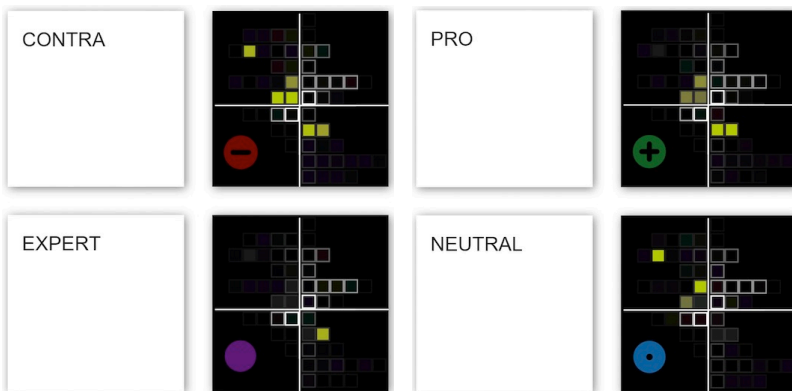


FIGURE 8 Argumentative patterns in S21.

We first investigate the argumentative structure of the different parties, which we assume to consist of the speaker party's usage of causal, contrastive, conditional and concessive discourse structures across the debate (positions of the individual feature in the Discourse Maps in Figure 8; lower right quadrant: premises (1,-2), conclusions (2,-2); upper left quadrant: consequence (-1,1), condition (-2,1), opposition (-1,2), concession (-4,4). All features are numerical, i.e., the feature-glyphs in the Discourse Map encode the frequency with which argumentative structures occur: The more frequent they occur, the brighter the glyph (for the color mapping see Figure 5).

Figure 8 shows that the patterns differ substantially: While the project proponents (PRO) and the project opponents (CONTRA) generally have a high frequency of premises and conclusions (brightness of (1,-2) and (2,-2), respectively), the experts (EXPERT) only employ conclusions (2,-2), the mediator (NEUTRAL) uses none of those patterns. However, he has the highest frequency of oppositions (brightest feature glyph in (-1,2) across the four speaker parties) and concessions in comparable frequency to the opposition (brightness of (-4,4)). Investigating the data more closely, it becomes clear that the mediator tries to come to a conclusion regarding individual points in the debate by either opposing information of individual speakers or conceding to them.

Another important aspect in the context of the S21 arbitration is the degree to which the speaker parties negotiate and accommodate. For the analysis we take into account four features, shown in Figure 9: consensus (-1,3), agreement (-2,3), the negotiation relation (-1,5), and the negotiation count (-2,5). The brighter the glyphs for consensus and the negotiation relation, the more frequent lexical items indicate consensus and negotiation. For instance, the opponents of S21 (CONTRA) have a high frequency of consensus-indicating lexical items and a comparatively lower frequency of negotiation-indicating items. The experts show neither, the proponents only show patterns of negotiation and the mediator shows a low frequency for both consensus and negotiation.

Agreement and the negotiation are bipolar: The redder the glyph, the stronger disagreement and counter-negotiation, the greener the glyph, the stronger agreement and negotiation. The combination of numerical and bipolar features allows us to interpret the patterns for the four speaker parties: Whereas the mediator has a high degree of negotiation and accommodation moves, the experts exhibit a comparatively low degree, as is to be expected from their role in the discourse.

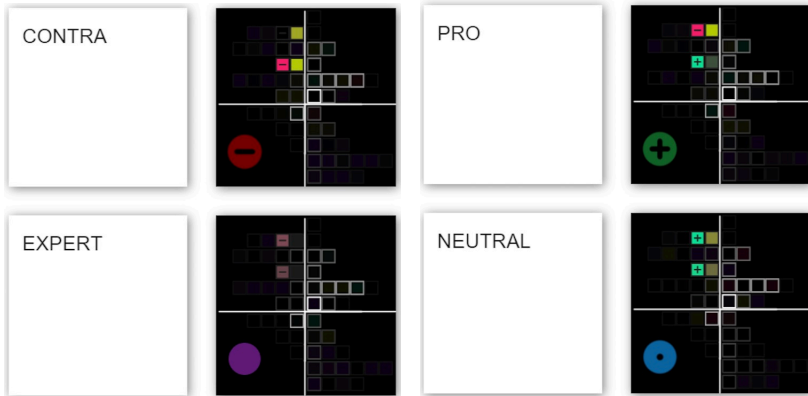


FIGURE 9 Negotiation and accommodation in S21.

4.7 Evaluation

The evaluation is intended to verify that Discourse Maps are a viable means to interpret patterns in large amounts of debate data. To that end we conducted a user study, where we presented the users with different Discourse Maps and a scale as to possible interpretations (e.g. ‘Given this Map, rate the following speakers regarding their degree of reason-giving’, with ‘1’ for the strongest manifestation of feature X, and ‘4’ for the weakest). The six study participants were Master and PhD students in linguistics or computational linguistics and they each encoded six features for four speakers, resulting in 144 measurements. In the first evaluation, we calculate the deviation from a gold standard rating of a domain expert. Table 5 shows that the average of the participants deviated by 0.041 points in their measurement from the gold standard with a standard deviation of 0.544 points.

TABLE 5 Summary result of the user study for the six measures.

	Average Error	Standard Deviation
Reason	-0.083	0.408
Conclusion	0.000	0.510
Concession	0.208	0.414
Contrast	0.000	0.978
Common Ground	0.000	0.510
Consensus Willing	0.125	0.448
Average	0.041	0.544

In addition to this quantitative feedback, we collected qualitative feedback through semi-structured interviews and expert testimonies. These showed a consensus that although the design of the Discourse Maps visualization is fairly complex, it enables the answering of challenging research questions and the investigation of complex phenomena and hypotheses. Hence, with Discourse Maps we created a specialized expert tool that is tailored to deliberation analysis. Through the strict integration of the theoretical dimensions and hierarchy, the experts stated that this visualization became a sort of “brain dump” — reducing the cognitive load of remembering feature relationships and focusing their analysis on the interesting patterns.

Overall, the process of designing such an approach has taught us that the trade-off between the simplicity of the design and the expressiveness of the visualization does not always have to go towards simplicity, especially for tools intended to analyze complex phenomena and targeting expert analysts. Throughout the design process, we were confronted with the feedback that: showing more details in the visualization is desirable, even if that would require certain training in reading the visual encoding. We refer to this as training experts to become literate in reading patterns from the Discourse Maps.

4.8 Discussion and Conclusion

To conclude this paper, we discuss the lessons learned from our collaboration, as well as the limitations of our approach. Lastly, we provide a brief summary and point to future work.

4.8.1 Lessons Learned

Through our iterative design process, as well as the tight collaboration with domain experts, we have learned multiple lessons that are of general interest beyond our concrete use case. The first and most important thing we realized through this process is that a targeted analysis of the users’ domain understanding might lead to complex visual encoding to become expressive enough for the problem at hand. According to the user feedback we received, the design of the Discourse Maps had to be tightly aligned to their mental models of the analyzed subject matter, taking into account that the resulting complexity of the visualization will require a training phase and memorization during analysis.

However, through the consistency in the representation on different text-granularity levels, Discourse Maps allowed users to compare single utterances with aggregates based on topics, speakers, etc. This enabled every Discourse Map to serve as a fingerprint of the underlying data. While the dynamic layout generation allows the creation of adaptable

Discourse Maps depending on the corpus characteristics, the internal layout of a map is stable for a given corpus to avoid confusion and enable analysts to memorize layouts corresponding to their data.

Lastly, a notable thing to highlight that was useful for designing such an information-dense visual representation is the two-level visual encoding of the data. As described in the paper, we can regard the feature-glyphs as multi-dimensional glyphs ordered in the grid of a Discourse Map. However, the Discourse Maps themselves could also be seen as glyphs, ordered in the grid of the overall layout-canvas as small-multiples. This allows for a data analysis and comparison on two levels of detail and, thus, has been praised by our domain experts as a useful “*overview first, details-on-demand*” technique.

4.8.2 Limitations

As highlighted in Section [4.5.1](#), this work was a constantly evolving endeavor to search for a suitable visual design, while ensuring an effective visual mapping for the domain problem complexity. We thus considered the most important attributes (i.e., data structure and feature values) to be mapped to the central visual attributes, while taking into account that less important attributes (e.g., size of the underlying text) are mapped to peripheral visual attributes with limited comparability ranges. Such trade-offs were at the heart of every design iterations and are subject to future work.

In particular, limitations include the high visual complexity (remedied by the double encoding of the feature-glyphs, as well as the interactivity, e.g., the ability to enlarge glyphs for a focused analysis); the potentially low visual dynamic range of color mapping (remedied by interactive mouse-over text-popups with the exact feature values, as well as the relative normalization of all color ranges for each feature; and the arrangement of objects to make use of the visual proximity to enhance comparison.

4.8.3 Summary and Future Work

We have presented Discourse Maps, a Visual Analytics approach to analyze conversation dynamics based on the theory of deliberative communication. Our approach is molded to a hierarchical frame of dimensions, subdimensions, and measures determined with respect to a framework informed by questions coming from political science. Discourse Maps are designed in conformity with the guidelines for glyph-based visualizations and enable an interactive, explorative analysis process that can be utilized to form new data-driven hypotheses and verify them. We have showcased the usefulness of our technique via a use case from the

S21 arbitration and evaluated the overall approach with quantitative and qualitative studies.

References

- Asher, Nick and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.
- Bergstrom, Tony and Karrie Karahalios. 2009. Conversation clusters: grouping conversation topics through human-computer dialog. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2349–2352.
- Bögel, Tina, Annette Hautli-Janisz, Sebastian Sulger, and Miriam Butt. 2014. Automatic detection of causal relations in german multilog. In *Proceedings of the EAACL 2014 Workshop on Computational Approaches to Causality in Language*, pages 20–27.
- Borgo, Rita, Johannes Kehrler, David H. Chung, Eamonn Maguire, Robert S. Laramee, Helwig Hauser, Matthew Ward, and Min Chen. 2013. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In M. Sbert and L. Szirmay-Kalos, eds., *Eurographics 2013 – State of the Art Reports*, pages 39–63.
- Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2548–2555.
- Ceriani, Lidia and Paolo Verme. 2012. The origins of the Gini index: Extracts from *Variabilità e Mutabilità* (1912) by Corrado Gini. *The Journal of Economic Inequality* 10(3):421–443.
- Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259.
- Donath, Judith and Fernanda B. Viégas. 2002. The chat circles series: explorations in designing abstract graphical communication interfaces. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, pages 359–369.
- El-Assady, Mennatallah, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. 2016. ConToVi: Multi-party conversation exploration using topic-space views. *Computer Graphics Forum* 35(3):431–440.
- El-Assady, Mennatallah, Annette Hautli-Janisz, Valentin Gold, Miriam Butt, Katharina Holzinger, and Daniel A. Keim. 2017a. Interactive visual analysis of transcribed multi-party discourse. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 49–54.

- El-Assady, Mennatallah, Rita Sevastjanova, Bela Gipp, Daniel Keim, and Christopher Collins. 2017b. NEREx: Named-entity relationship exploration in multi-party conversations. *Computer Graphics Forum* 36(3):213–225.
- El-Assady, Mennatallah, Rita Sevastjanova, Daniel Keim, and Christopher Collins. 2018a. ThreadReconstructor: Modeling reply-chains to untangle conversational text through visual analytics. *Computer Graphics Forum* 37(3):351–365.
- El-Assady, Mennatallah, Rita Sevastjanova, Fabian Sperrle, Daniel A. Keim, and Christopher Collins. 2018b. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE Transactions on Visualization and Computer Graphics* 24(1):382–391.
- El-Assady, Mennatallah, Fabian Sperrle, Oliver Deussen, Daniel A. Keim, and Christopher Collins. 2018c. Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE Transactions on Visualization and Computer Graphics* 25(1):374–384.
- Fishkin, James S. and Robert C. Luskin. 2005. Experimenting with a democratic ideal: Deliberative polling and public opinion. *Acta Politica* 40:284–298.
- Fuchs, Johannes, Petra Isenberg, Anastasia Bezerianos, and Daniel A. Keim. 2017. A systematic review of experimental studies on data glyphs. *IEEE Transactions on Visualization and Computer Graphics* 23(7):1863–1879.
- Gerhards, Jürgen. 1997. Diskursive versus liberale Öffentlichkeit. Eine empirische Auseinandersetzung mit Jürgen Habermas. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 49:1–47.
- Gold, Valentin, Annette Hautli-Janisz, Katharina Holzinger, and Mennatallah El-Assady. 2016. VisArgue: Analysis and visualization of deliberative political communication. *Political Communication Report* 26(1–2).
- Gold, Valentin and Katharina Holzinger. 2015. An automated text-analysis approach to measuring deliberative quality. Paper presented at the 73th Annual Meeting of the Midwest Political Science Association, San Francisco.
- Gold, Valentin, Christian Rohrdantz, and Mennatallah El-Assady. 2015. Exploratory text analysis using lexical episode plots. In E. Bertini, J. Kennedy, and E. Puppo, eds., *Eurographics Conference on Visualization: Short Papers*, pages 85–90.
- Hautli-Janisz, Annette and Miriam Butt. 2016. On the role of discourse particles for mining arguments in German dialogs. In *Proceedings of the COMMA 2016 FLA workshop*, pages 10–17.
- Hoque, Enamul and Giuseppe Carenini. 2016. MultiConVis: A visual text analytics system for exploring a collection of online conversations. In *Proceedings of Intelligent User Interfaces*, pages 96–107.
- Jekat, Susanne, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. 1995. Dialogue acts in verbmobil. Tech. rep., Saarländische Universitäts- und Landesbibliothek.

- Jentner, Wolfgang, Mennatallah El-Assady, Bela Gipp, and Daniel A. Keim. 2017. Feature alignment for the analysis of verbatim text transcripts. In *Proceedings of the EuroVis Workshop on Visual Analytics*, pages 13–17.
- Keim, Daniel A. and Daniela Oelke. 2007. Literature fingerprinting: A new method for visual literary analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122.
- Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1*, pages 423–430.
- Landwehr, Katharina and Katharina Holzinger. 2010. Institutional determinants of deliberative interaction. *European Political Science Review* 2:373–400.
- Lassiter, Daniel. 2010. Gradable epistemic modals, probability, and scale structure. *Semantics and Linguistic Theory* 20:197–215.
- Leshed, Gilly, Diego Perez, Jeffrey T. Hancock, Dan Cosley, Jeremy Birnholtz, Soyoung Lee, Poppy L. McLeod, and Geri Gay. 2009. Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 537–546.
- Lin, Ziheng, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering* 20:151–184.
- Mairesse, Francois, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* 30:457–500.
- Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a theory of text organization. *Text* 8(3):243–281.
- Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: MIT Press.
- McCarthy, Philip M. and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing* 24(4):459–488.
- Mittelstädt, Sebastian, Dominik Jäckle, Florian Stoffel, and Daniel A. Keim. 2015. ColorCAT: Guided Design of Colormaps for Combined Analysis Tasks. In E. Bertini, J. Kennedy, and E. Puppo, eds., *Eurographics Conference on Visualization: Short Papers*, pages 115–119.
- Mittelstädt, Sebastian, Andreas Stoffel, and Daniel A. Keim. 2014. Methods for compensating contrast effects in information visualization. *Computer Graphics Forum* 33(3):231–240.
- Mohammad, Saif M. and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create and emotion lexicon. In *Proceedings of the NAACL 2015 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.
- Polanyi, Livia, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. Sentential structure and discourse parsing. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 80–87.

- Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation 2008*, pages 2961–2968.
- Rafferty, Anne N. and Christopher D. Manning. 2008. Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the ACL 2008 Workshop on Parsing German*, pages 40–46.
- Schiller, Anne. 1994. Dmor - user's guide. Tech. rep., Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.
- Shahaf, Dafna, Carlos Guestrin, and Eric Horvitz. 2012. Trains of thought: Generating information maps. In *Proceedings of the 21st International Conference on World Wide Web*, pages 899–908.
- Sridhar, Dhanya, James Foulds, Marilyn Walker, Bert Huang, and Lise Getoor. 2015. Joint models of disagreement and stance in online debate. In *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 116–125.
- Stalnaker, Robert. 2002. Common Ground. *Linguistics and Philosophy* 25(5-6):701–721.
- Stede, Manfred and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation 2014*, pages 925–929.
- Versley, Yannick and Anna Gastel. 2013. Linguistic tests for discourse relations in the Tüba-D/Z corpus of written German. *Dialogue and Discourse* 4(2):142–173.
- Zarisheva, Elina and Tatjana Scheffler. 2015. Dialogue act annotation for twitter data. In *Proceedings of the SIGDIAL 2015 Conference*, pages 114–123.
- Zimmermann, Malte. 2011. Discourse Particles. In P. Portner, C. Maienborn, and K. von Heusinger, eds., *Semantics (Handbücher zur Sprach- und Kommunikationswissenschaft)*, pages 2011–2038. Mouton de Gruyter.