Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2874–2883
Marseille, 20-25 June 2022

A Corpus of German Citizen Contributions in Mobility Planning: Supporting Evaluation Through Multidimensional Classification

Julia Romberg, Laura Mark, Tobias Escher
Heinrich Heine University Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf
{julia.romberg, laura.mark, tobias.escher}@hhu.de

Abstract

Political authorities in democratic countries regularly consult the public in order to allow citizens to voice their ideas and concerns on specific issues. When trying to evaluate the (often large number of) contributions by the public in order to inform decision-making, authorities regularly face challenges due to restricted resources. We identify several tasks whose automated support can help in the evaluation of public participation. These are i) the recognition of arguments, more precisely premises and their conclusions, ii) the assessment of the concreteness of arguments, iii) the detection of textual descriptions of locations in order to assign citizens' ideas to a spatial location, and iv) the thematic categorization of contributions. To enable future research efforts to develop techniques addressing these four tasks, we introduce the *CIMT PartEval Corpus*, a new publicly-available German-language corpus that includes several thousand citizen contributions from six mobility-related planning processes in five German municipalities. The corpus provides annotations for each of these tasks which have not been available in German for the domain of public participation before either at all or in this scope and variety.

Keywords: public participation, argument mining, thematic categorization, location detection, spatial planning, mobility

1. Introduction

Public participation is the "practice of consulting and involving members of the public in the agenda-setting, decision-making, and policy-forming activities" (Rowe and Frewer (2004), p. 512). By enabling citizens to communicate their preferences on specific issues, it is an important element of representative democracies to improve responsiveness between the electorate and their representatives. While there is a debate about what role such consultative procedures can or indeed should play (Parry and Moyser, 1994), here we focus on the more practical issue of how to process and evaluate the input of citizens once public authorities have chosen to engage in such consultations. This has become a more pressing issue because of concerns about declining public support for democratic actors and institutions (Norris, 2011) as well as the easy availability of online forms of participation which has led to widespread use of public participation, regularly resulting in large numbers of contributions from citizens.

Processing the contributions from citizens poses significant challenges for public authorities because norms of democratic equality and administrative justice demand that every single contribution is carefully evaluated. While it is desirable that people participate in large numbers for increasing the acceptance and possibly the usefulness of the output, public administration (or the private companies tasked with evaluation) often lack personnel and time to deal with large quantities of unstructured citizen input (Arana-Catania et al., 2021; Aitamurto et al., 2016; Simonofski et al., 2021). As a result, the evaluation process often takes a long time, which can lead to delays in the planning process and to discontinuities in public communication, with all the associated negative consequences for efficiency, transparency and public acceptance.

Given that evaluation usually means categorizing input from citizens into different dimensions (e.g. according to topic, urgency or responsibility) before taking a decision on the individual contribution, one opportunity to support this manual evaluation process to make it more efficient is pre-structuring citizens' input. While some approaches focus on user-generated structuring, i.e. by letting citizens classify their contributions themselves, these allow only to categorize a limited number of dimensions (in order not to overburden users), and are limited by the lack of expertise of the lay public. Instead, here we focus on utilizing Natural Language Processing (NLP) that has been suggested as an alternative (OECD, 2003). Despite the relevance for democratic participation as well as significant progress in NLP techniques, automated classification of citizen contributions has yet to be advanced to a level sufficient to offer reliable support for practice.

Therefore, in this paper we propose four classification tasks in order to support the evaluation process. We provide datasets from six public participation processes in five German cities that have been annotated according to all or a part of those four dimensions to enable the training of supervised models for these tasks. Table 1 gives an overview of the *CIMT PartEval Corpus*.

In dialogue with practitioners from public administrations, participation service providers and planning consultants, we identified four common tasks whose support through automation would benefit the evaluation of participation processes (Romberg and Escher, 2020).

			datasets						
task	unit level	total units	CD_B	CD_C	CD_M	CQ_B	MC_K	MC_O	language resource reference
i) argument components	sentences	17,852	10,442	1,704	2,193	1,505	2,008		(Romberg et al., 2022a)
ii) argument concreteness	sentence spans	1,127	679	92	110	55	191		(Romberg et al., 2022b)
iii) geographic location	token spans	4,830	4,087	743					(Romberg et al., 2022c)
iv) thematic categorization	documents	697						697	(Romberg et al., 2022d)

Table 1: Overview of the coded units for the different tasks and datasets included in the CIMT PartEval Corpus.

These are i) the detection of arguments, ii) the assessment of the concreteness of arguments, iii) the recognition of locations that contributions refer to, and iv) structuring according to topics.

Individually and in combination with each other, these tasks can help to structure the data and thus facilitate the analysis in the following ways: The distinction into different **argument components** is important because it allows practitioners to get a quick overview of the relevant parts of the contributions. The recognition of **concreteness** enables practitioners to filter the most specific contributions, e.g. as a possible starting point for evaluation. The **recognition of locations** is helpful for processes without user-generated geo-referencing because it allows clustering contributions in spatial entities, e.g. to detect hot spots or assign responsibilities based on geographical jurisdictions. Finally, the **thematic categorization** helps to obtain a content-related overview fast and makes it possible to analyse contributions with similar topics together and therefore find patterns and contradictions more easily. What is more, it is the basis for delegation to those administrative units responsible for dealing with the contributions.

We have chosen to focus on one specific type of such participatory processes, namely those concerned with mobility such as the redesign of streets or the development of strategic mobility plans. Mobility planning is an important area within spatial planning in which consultations are regularly utilized. Structurally, these contributions are not different from participation processes on other issues but the focus on mobility allows us to provide a topic-specific categorization.

Our contributions are: We release a new annotated corpus (available under a Creative Commons License) for the development of supervised models to support the multidimensional evaluation of German-language public participation processes, consisting of six processes that differ in participation format and process focus. We provide annotations for the four described classification tasks. To the best of our knowledge, for some of the tasks, this is the first German-language (iii) or first-ever (ii, iv) annotated corpus from the domain of public participation. Particularly noteworthy are the new quality criterion for arguments (concreteness) and the thematic categorization scheme that is universally applicable to transport-related processes.

The remainder of this paper is as follows: In the next section, we review the existing language resources from the domain of public participation. We then present the public participation processes included in our corpus in Section 3. The four classification tasks are subsequently addressed in Section 4 (argument components), Section 5 (argument concreteness), Section 6 (geographic location), and Section 7 (thematic categorization). In each section, the task is introduced, followed by an overview of relevant work, a description of the annotation process and a presentation of the resulting dataset. Section 8 concludes with a summary and an outlook on future work.

2. Language Resources from Public Participation

In recent years, citizen contributions from different public participation processes have been annotated to support NLP research tasks, mainly the recognition of arguments and their properties as well as thematic categorization of citizen ideas. Most of these derived from rulemaking processes in the USA (Kwon et al., 2006; Arguello et al., 2008; Cardie et al., 2008; Park and Cardie, 2014; Konat et al., 2016; Aitamurto et al., 2016; Lawrence et al., 2017; Park and Cardie, 2018; Eidelman and Grom, 2019), some from processes in Chile (Fierro et al., 2017), Germany (Liebeck et al., 2016), Japan (Morio and Fujita, 2018) and Korea (Kim et al., 2021).

In the field of argument mining, the focus was especially in recognizing argumentation components and their supporting relations. Lawrence et al. (2017) and Konat et al. (2016) focused on the dialogical relation. Park and Cardie (2018) annotated comments with a more detailed scheme, in which propositions were subdivided into different types and then linked. A rather general argumentation scheme for informal online public participation processes was introduced by Liebeck et al. (2016). More specific is the adaptation to the thread structure of online platforms by Morio and Fujita (2018) who added intra-post and inter-post relationships. Probably the largest dataset was presented by Eidelman and Grom (2019), in which about 1.8 million sentences from various rulemaking efforts were semi-automatically assigned argument claim types.

Further work put the emphasis on the quality of citizens' arguments such as the verifiability of propositions (Park and Cardie, 2014). Arguello et al. (2008) proposed the recognition of citations in citizen comments to value them as factual evidence for claims and opinions.

Moreover, attention was paid to structuring citizens' ideas thematically. Cardie et al. (2008) and Aitamurto et al. (2016) focused on thematic categorization of transportation-related rulemaking processes by developing customized categorization schemes. A somewhat different approach to thematic categorization was taken by Kim et al. (2021) who assigned complaints that were submitted to a civic online participation platform to respective administrative fields.

Only a few datasets were coded according to multiple viewpoints. One is that of Kwon et al. (2006), whose multidimensional coding included thematic categorization and the analysis of argument structure. In Fierro et al. (2017), a large-scale dataset of citizen arguments collected during Chile's 2016 constitutional process was presented. Arguments were categorized according to their function and thematically organized into a hierarchy of constitutional concepts.

In summary, there exists only a single German dataset for the domain of public participation and this focuses only on argument mining within a single process (Liebeck et al., 2016). On thematic categorization of citizen ideas we find only a few corpora even for other languages (mainly English). None address concreteness or geographic location and few offer annotations representing multiple dimensions.

To address this gap, we present a collection of German-language datasets coded according to several dimensions, namely i) argument components, ii) concreteness of arguments, iii) location detection, and iv) thematic categorization, since there are no existing (German-) language resources in our application domain for the latter three tasks.

3. Datasets

We consider six different public participation processes in our data collection, namely three "Raddialoge" ("Cycling Dialogues") in the cities of Bonn, Cologne (district Ehrenfeld) and Moers as well as "Leben in Bonn" ("Living in Bonn"), "Krefeld bewegen" ("Moving Krefeld"), and Hamburg's "freiRaum Ottensen" ("Space for Ottensen"). While these are all related to urban mobility planning, they span different mobility-related issues and participation formats.

In detail, the three "Raddialog" datasets derive from largely identical participation processes conducted in autumn 2017 in which the local authorities invited their citizens to propose measures to improve cycling in the city. A map-based online platform allowed citizens to locate their contributions on a map, resulting in $2,314$ unique contributions consisting on average of 4.83 sentences (standard deviation $\sigma = 2.63$) for **Raddialog Bonn** (henceforth CD_B), 366 contributions (4.66 sentences, $\sigma = 3.00$) for **Raddialog Ehrenfeld** (CD_C) and 459 contributions (4.78 sentences, $\sigma = 2.61$) for **Raddialog Moers** (CD_M). In addition, in Bonn the online platform was supplemented with a representative survey of the population. In total, 761 citizens expressed up to three suggestions for improvement either via the paper-based questionnaire or an online alternative, resulting in $1,386$ contributions (1.09 sentences, $\sigma = 0.37$) for **"Leben in Bonn"** (CQ_B).

Within **"Krefeld bewegen"** (MC_K) the city of Krefeld invited citizen comments on the development of a mobility concept. The first phase in 2020 focused on general aims of the new concept and the second phase invited suggestions for specific measures. This resulted in 337 contributions (5.96 sentences, $\sigma = 5.63$).

The most recent dataset included in the corpus derives from a public participation process by the district of Altona in Hamburg (**"freiRaum Ottensen"**, MC_O). As part of the transformation of its quarter Ottensen into a traffic-calmed neighborhood, the district office implemented a map-based online dialogue that took place in August 2021. In total, it received 697 contributions (4.95 sentences, $\sigma = 2.49$).

All datasets were separately examined by service providers as well as our team and any potentially identifying personal information was removed. The data in the corpus is available under a Creative Commons CC BY licence and may be distributed in accordance with the corresponding conditions. Users of the online participation platforms accepted these conditions via the terms of use of these platforms, while the data originating from the questionnaires was released under this licence by the principal investigator of the survey.

4. Sentence-level Argument Components

A central aspect through which citizens communicate their ideas are arguments. Automated analysis of arguments, known as argument mining, enables practitioners to get a quick overview of relevant text passages. We here focus on two common tasks in argument mining, namely the identification of argument components and the identification of clausal properties (Lawrence and Reed, 2019). Part of our corpus for argument component analysis (described in this section) has previously been introduced in Romberg and Conrad (2021).

4.1. Related Work

Previous work in our application domain either followed the classic claim-premise model (Liebeck et al., 2016; Morio and Fujita, 2018), or had a stronger focus on the intrinsic characteristics of claims (Fierro et al., 2017; Park and Cardie, 2018), e.g. if claims are factual, contain values or propose policies. For more detail on related work, please see Romberg and Conrad (2021).

Our work is closest to that of Liebeck et al. (2016) whose THF Airport ArgMining Corpus is the only German-language public participation dataset for argument mining. However, there are several differences between the corpora: First, we provide seven times more sentences coded with argument components. Second, our focus is not on the dialogue structure within each thread but on the detection of propositions within the initial contributions. Third, our corpus comprises several processes differing in format and

		CD_B		CD_C		CD_M		MC_K		CQ_B		all
	total	10,442		1,704		2,193		2,008		1,505		17,852
	non-arg	1,153	(11.0%)	197	(11.6%)	382	(17.4%)	431	(21.5%)	172	(11.4%)	2,335
arg	mpos	2,851	(27.3%)	603	(35.4%)	404	(18.4%)	961	(47.9%)	1,083	(72.0%)	5,902
	premise	6,700	(64.2%)	951	(55.8%)	1,452	(66.2%)	685	(34.1%)	373	(24.8%)	10,161
	overlap	262	(2.5%)	47	(2.8%)	45	(2.1%)	69	(3.4%)	123	(8.2%)	546

Table 2: Distribution of sentences among the different argument component categories per dataset.

		CD_B	CD_C	CD_M	CQ_B	MC_K	all
	sentences	1,251	191	230	188	376	2,236
kappa	non-arg	0.58	0.68	0.67	0.59	0.69	0.63
	mpos	0.82	0.81	0.81	0.73	0.78	0.81
	premise	0.82	0.87	0.83	0.83	0.77	0.84
	overall	0.76	0.80	0.77	0.72	0.73	0.77

Table 3: Number of sentences under consideration and kappa agreement for argument component annotation.

process subject but all coded with a uniform coding scheme. This enables a comprehensive evaluation of machine learning methods, also with respect to the transferability of trained models to new processes, allowing robust models to be developed. Such cross-dataset evaluation is important to better assess the practical applicability of models.

4.2. Definition of Argument Components

Public participation allows citizens to contribute to a decision-making process by proposing their ideas and voicing concerns. In spatial planning processes, this usually involves describing a problem or condition, from which a proposition is derived. We thus define two types of argument components: *Major positions* (short *mpos*) are options for actions that are being proposed. *Premises* are reasons that attack or support either a major position or another premise. With this, we adopt one part of the argumentation scheme of Liebeck et al. (2016). Sentences without premise or major position are considered as *non-argumentative* (*non-arg*).

4.3. Annotation Study

First, we developed annotation guidelines based on 151 contributions from the dataset CD_B. Subsequently, the remaining contributions, as well as all contributions from CD_C, CD_M, CQ_B and MC_K were coded. Three annotators were instructed to decide for each sentence (titles included) whether it has argumentative content and, if yes, if it is a *major position* or a *premise*. Since some sentences contain components of both types, multi-labeling was allowed.

To assess coder agreement on this task, about ten percent of each dataset was processed by all the coders. This sums up to 585 contributions with 2,236 sentences. The agreement on these sentences was measured using Fleiss (1971)' kappa.

With an overall agreement of 0.77[1], the coding can be considered reliable (see Table 3). However, there was a greater uncertainty in the selection of non-argumentative sentences, while the agreement between the two types of argument components was rather high. In a subsequent curation phase, the sentences with inconsistent coding were reviewed and resolved by two annotation process supervisors. This showed that there were regular misclassifications of whether a sentence was indeed argumentative, with coders being more inclined to classify argumentative sentences as non-argumentative than vice versa. Furthermore, it can be seen within the argumentative sentences that the assignment of premises was more accurate than that of major positions.

Due to the considerable time required for multiple coding and given the high reliability, we decided to have the remaining 4,126 contributions with 15,616 sentences coded only once, evenly distributed among the coders.

4.4. Corpus Statistics and Discussion

The resulting distribution of sentences among the annotation classes is given in Table 2. Overall, the share of sentences without argumentative content is small. Depending on the process, 80 to 90 percent of sentences are argumentative. However, the distribution of argument component types varies greatly between the different processes. Premises clearly predominate in the cycling dialogues, while the other two processes seem to be more conclusion-oriented and favor major positions. This is particularly evident in the survey data where participants had limited space for writing suggestions. For online platforms, few sentences contain both a major position and a premise (overlap). In contrast, in the survey data there is a greater overlap of argument components, which nonetheless affects less than one in ten sentences. The variety of the processes included in the corpus results in very different class distributions, supporting the development of robust machine learning models.

5. Argument Concreteness

We then focus on the concreteness of the argumentative components, the automated evaluation of which can help practitioners filter out arguments that can be evaluated immediately. The less specific citizens' ideas are, the more difficult and hence time-consuming it will be for evaluators to derive measures for implementation.

[1] In the overall calculation, sentences containing both major position and premise constitute an additional category.

5.1. Related Work

The evaluability of public participation contributions has previously been raised by Park and Cardie (2018), Park and Cardie (2014) and Arguello et al. (2008), who saw the lack of reasoning and evidence verifying citizen contributions as the main obstacle to evaluating propositions. However, in the evaluation of spatial planning processes, we consider the level of concreteness of the arguments (i.e. how detailed current conditions and proposed improvements are described) as the most important indicator for evaluability.

To the best of our knowledge, we are the first to provide a resource for this type of concreteness of arguments, while other aspects of the *quality of arguments* have received increasing attention in recent years (e.g. Habernal and Gurevych (2016a), Habernal and Gurevych (2016b), Toledo et al. (2019), Gretz et al. (2020)). A systematic taxonomy of dimensions for argument quality, regarding logic, rhetoric and dialectic aspects, can be viewed in Wachsmuth et al. (2017).

5.2. Definition of Degrees of Concreteness

We propose a distinction between high, intermediate and low concreteness. Argument components are *highly concrete* when they contain details that specify the **what**, **how** and **where**. Such specifications can be colour, surface, measurements, etc. (an example is "cycle paths often in poor condition, tarred surface torn up, bumpy due to roots, mostly only half width because overgrown"). Contributions with an *intermediate concreteness* contain some specifications like location or descriptions of what exactly should be done, but leave some room for interpretation (e.g.: "new cycle lanes without interruptions" - the measure is described and somewhat detailed, but it is not clear how it should look exactly and where it should be located). Contributions with *low concreteness* contain no information on location or specific measures, so that a variety of measures could be deducted (e.g.: "unfavorable traffic lights" - it does not become clear, what exactly the problem is, where it is and what should be done).

Distinction of concreteness was applied only to argumentative components, non-argumentative sentences were excluded. In order to support different use cases, such as searching either for (concrete) major positions or (concrete) premises, we consider the concreteness of the two types of argument components separately.

5.3. Annotation Study

We decided to use the curated documents of the previous annotation task in order to ensure the soundness of the annotation of sentence-level argument components. Determining the concreteness of solitary sentence-level argument components is hardly feasible. Therefore, the coders first interrelated argument components of the same types (i.e. premises or major positions) to form units with coherent sense, and the annotation supervisors resolved inconsistencies. In a second step, we asked coders to rate the resulting units' concreteness using guidelines that were developed on the same data as with argument components.

	CD_B	CD_C	CD_M	MC_K	CQ_B	all
mpos	265	40	40	126	42	513
premise	414	52	70	65	13	614
total	679	92	110	191	55	1,127

Table 4: Units of interrelated argument components.

It turned out that the perception of concreteness is rather subjective, which was also confirmed to us by those responsible for analyzing the contributions. We thus decided to include a total of five annotators to obtain a multitude of individual concreteness ratings. Due to the subjective nature, we dispense with a manual curation step in which an unambiguous assignment of concreteness to units is made, but instead release the five individual codings. While the assessment of concreteness exhibits some subjectivity, it is not arbitrary as is documented by Krippendorff (2013)'s weighted alpha[2], which shows an agreement score of 0.46.

5.4. Corpus Statistics and Discussion

Overall, 513 units of interrelated sentences containing major positions and 614 units of interrelated sentences containing premises were formed and coded by concreteness (see Table 4). To each of the units belong five codings by the different annotators. There is complete agreement among coders in 478 cases, about 42 percent of the units. In the majority of disagreements, coders chose adjacent categories, so while subjective perception differs slightly, there is a consistent trend in whether the unit (460 in total) is rather concrete or vague. Within 189 units, however, a strongly subjective assessment is evident, in which all or the two opposing degrees of concreteness were assigned.

Analysis of the degrees of concreteness reveals that citizens clearly tend to write highly concrete arguments in the processes considered here. Nevertheless, on average about twenty percent of the argument units have intermediate or low concreteness, thus automated recognition will allow highlighting the most relevant (concrete) content.

6. Geographic Location

In spatial planning processes, the geographic location of citizens' contributions is of great importance to the evaluation as it allows geo-referencing of contributions and clustering of ideas by location. Map-based processes on online platforms offer a possibility in which citizens can locate their ideas on a map. However,

[2] We weight using the Euclidean distance to account for the level of deviation between the codings, i.e., whether they are adjacent (e.g., low/high and medium concreteness) or non-adjacent categories (low and high concreteness).

not all public participation in spatial planning is geo-referenced as exemplified by the survey-based data (CQ_B) in our corpus. To address this problem we propose the use of text-based geo-location and present a dataset of textual locations and GPS coordinates.

6.1. Related Work

Text-based document geo-location is the task of determining the geographic coordinates of a document's associated location by its textual content. Originally a task from information retrieval, it combines language modelling and geographical information science.

This task was initially approached through clustering. Much of these works relied on named entity recognition to narrow the feature space to geographical indications (e.g. Smith and Crane (2001)). Other approaches relied on more unsupervised vocabulary selection strategies (e.g. Adams and Janowicz (2012), Wing and Baldridge (2014)). Putting a stronger focus on natural language processing and supervised learning, the recognition of textual location phrases was supported by the development of specified annotated corpora. McNamee et al. (2020), for example, concentrated on fine-grained tagging of location phrases that complement named entity mentions with additional words which provide further information to specify a location (e.g. prepositions).

Further work directly combined the recognition of location information with a subsequent geo-coding step to associate the textual locations to GPS coordinates. Application domains were, inter alia, textual narratives from travel blogs (Skoumas et al., 2016) and news articles to map the local news coverage (Gupta and Nishu, 2020).

With public participation processes, we here introduce a new application domain that differs from previously targeted genres in document length, text quality, and prevalence of location, among other factors. Our use case requires a very precise mapping to pinpoint geo-coordinates, with location information as accurate as streets, intersections, and addresses.

6.2. Definition of Location Phrases

We define a textual *location* as a single word or a sequence of words included in a citizen's contribution that refers to the spatial placement of the respective contribution. These can be named entities, such as street names or city districts, but also, beyond that, constructions with more fine-grained location information that can be unambiguously marked on a map. Such phrases usually contain information that specifies the exact location, like the description of a specific angle (e.g. approaching some location from the right-hand side, or in the direction of the main station).

A known problem in determining the geo-positions from textual descriptions are ambiguous locations (e.g. Awamura et al. (2015), Smith and Crane (2001)). This includes, for example, street names, squares, or stations (like main station) without assignment to a city. For our use case, many of these cases are solved by the fact that the context in which the processes take place is usually known. Furthermore, we do not understand a word sequence as a location if it refers to several places in the city ("many/various/all parks in the city") or does not have a spatial reference point that specifies its geo-location (like "in the one-way street").

6.3. GPS Coordinates

The next step following the recognition of textual locations is the assignment to GPS coordinates based on the location phrases.

We chose the cycling dialogues (CD_B, CD_C) for the text-based document geo-location task because an assignment of GPS coordinates had already been part of the map-based online platforms, where each citizen was requested to explicitly indicate the location of their contribution as a point on the map. More complex shapes such as polygons were not allowed. We can assume that the textual location descriptions and the geo-locations given refer to the same entity, since citizens generally adhered to the requirements of point-wise referencing, and that the textual description should belong to the geo-referenced location. GPS coordinates are thus included in our annotated data corpus alongside the location phrases.

6.4. Annotation Study

Three trained annotators were instructed to identify the textual location spans within $2,529$ contributions from CD_B and CD_C. The coding guidelines were previously developed on additional 151 contributions from CD_B. Each location unit could consist of any number of consecutive words, but units could not cross sentence boundaries. 305 contributions, about ten percent of each dataset, served to determine the inter-annotator agreement and the remaining $2,224$ contributions were divided equally among the annotators. After calculating the inter-annotator agreement, documents with multiple annotations were reviewed by two supervisors and conflicts were resolved to obtain a unified coding.

We consider Krippendorff et al. (2016)'s alpha for unitizing textual continua[3] to evaluate the reliability of the coders. The alpha measure of 0.75 proves a high agreement between the coders. We assume that the coders worked as reliably on the contributions that were single-coded.

A look at the contributions with multiple codings shows that disagreements in the handling of prepositions (e.g. along, across, into, left/right of) occurred repeatedly. Another source of disagreement were nouns (e.g. bike lane, one-way street, sidewalk) at the beginning of location units. According to our guidelines, the coders had to decide whether additional words made

[3] We use the modified version of earlier definitions (Krippendorff, 1995; Krippendorff, 2013), which corrects shortcomings for studies with more than two annotators.

the location more precise. It turned out that perceptions did not always coincide on this.

6.5. Corpus Statistics and Discussion

The corpus comprises $2,529$ contributions, each of which is assigned to a GPS coordinate, and these contributions contain $4,830$ location phrases. The length of the location phrases varies from a single to up to 36 tokens, with on average 4.9 tokens ($\sigma = 3.48$). Examples for very short locations are street names or districts (e.g. downtown), while longer units contain more precise descriptions.

Overall, about twelve percent of the tokens included in the contributions are part of a location phrase, a proportion that further illustrates the relevance of automated location of citizen ideas for spatial planning processes.

7. Thematic Categorization

Lastly, we address the thematic categorization of citizen contributions in our data corpus. This makes it possible to analyse contributions with similar topics together and detect patterns as well as to delegate contributions to the responsible administrative units.

7.1. Related Work

Content structuring by thematic categories has been addressed before, including by Kwon et al. (2006) for a mercury rulemaking process and by Fierro et al. (2017) in the context of a constitutional process. Cardie et al. (2008) and Aitamurto et al. (2016), like us, focused on transportation-related processes.

A problem shared by previous work is that the categories were fitted to the individual participation process. Such specification makes the development of supervised classification models for real-world use (i.e. beyond research purposes) impractical. If schema and training data have to be developed from scratch for each new process, the time required may quickly exceed the effort of a purely manual analysis, especially for processes with fewer contributions. This problem has previously been described by Purpura et al. (2008), who proposed active learning to reduce the amount of training data. Still, the amount of training data needed for an adequate prediction quality may remain high.

An alternative solution is to use categorization schemes that are universally applicable to multiple participation processes. These can be used to train models which can subsequently be applied to further processes without the need for additional training. An example is the work of Kim et al. (2021), in which contributions were assigned to the competent administrative fields (e.g. housing, culture, environment) based on a guideline for governments. We follow this example and define a universal scheme of transportation-related categories that is not limited to individual processes but can be used for structuring all kinds of mobility-related planning processes.

7.2. A Categorization Scheme for Mobility

We propose a category scheme that covers modes of transport as well as related aspects and allows multi-labeling.

The categorization scheme was developed based on a variety of sources including existing mobility concepts (e.g. Der Senator für Umwelt, Bau und Verkehr (2014)), categorizations proposed in documentations of participation processes (e.g. Zebralog (2020)), and topic choices currently available to users of online consultations[4]. This draft was then subjected to feedback from experts with practical experience in the evaluation of contributions, namely representatives of participation service providers, planning offices and administration, and subsequently improved.

Figure 1 provides an overview of **modes of transport**, almost always relevant in mobility-related processes, and their **specifications**. Please note that it is also possible for a contribution not to be assigned to any mode.

Regarding modes of transport, it is firstly specified if the contribution deals with *motorized* or *non-motorized transport* (or both). If the contribution explicitly refers to particular modes, these are then further specified: non-motorized modes are *cycling*, *walking* and *scooters*. Motorized modes encompass *local* and *long-distance public transport* as well as *commercial transport* which includes, e.g., delivery and waste disposal. Private cars are not included as a separate sub-category of motorized transport. Instead, relevant contributions will be subsumed under motorized modes because even when contributions refer specifically to "cars", the issues usually concern all motorized modes - even if this is not explicitly stated, e.g. when criticizing traffic signaling. As a matter of fact, there are hardly any issues that refer exclusively to private car traffic[5].

Only if the contribution concerns a mode of transport, it can then be assigned to one or more **specifications** such as the type of traffic (*moving traffic* or *stationary traffic*, i.e. parking). What is more, the categories of *new services* and *inter- and multimodality* can be added as supplementary information to the mode of transport, the first referring to technological advancements like e-mobility or app-based offers, the second referring to the connection of and between different modes of transport, like intermodal booking systems or the design of interchanges.

This nested system of categories allows both a general and a more specific classification of the data. The possibility to assign more than one topic to a contribution is an essential difference to most user-generated structuring approaches in online consultations. This multi-labeling is often necessary because contributions can deal with more than one topic.

[4] E.g., see the participation tool of service provider "tetraeder": www.buergerbeteiligung.de/ beispielhausen/

[5] An exception is residential parking, which can be identified through the specification "stationary traffic".

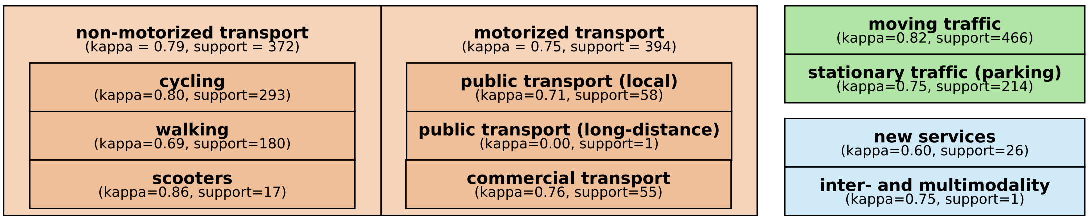

Figure 1: Overview of thematic categorization scheme for mobility. Numbers in parentheses denote inter-annotator agreement (Fleiss' kappa) and class support after solving disagreements.

7.3. Annotation Study

We started annotation with MC_O, a process that aims at a comprehensive mobility concept and therefore includes contributions on various modes of transport. The 697 contributions were coded by three coders according to our hierarchical scheme. Detailed coding guidelines were developed and the coders were trained on contributions from MC_K and further processes not part of the collection presented here. Since it became apparent during the coding process that some categories occurred much less frequently than others, we decided to have each document coded by all coders.

To analyze the reliability of the codings we calculated the Fleiss' kappa agreement for the categories reported in Figure 1. Most categories show a rather high level of agreement of 0.75 and above. Some categories with lower agreement such as long-distance public transport or inter- and multimodality suffer from very few contributions identified as belonging to this category (see next section), which is why the significance of kappa should be viewed with caution here. A subsequent screening and revision of the disagreements by two supervisors, one an urban planner, led to a final unique coding, which is the one presented in the following.

7.4. Corpus Statistics and Discussion

The class support of the final coding is depicted in Figure 1. About 82 percent of the contributions were about motorized or non-motorized transport, with moving traffic prevailing over stationary traffic. The optional categories new services and inter- and multimodality hardly occurred in the process under consideration, just as scooters and long-distance public transport.

These categories remain in the categorization schema as our aim is to provide a comprehensive scheme for all modes of transport, independent of this specific process. In other processes, we expect a different distribution of the classes. In order to provide a sufficient data basis for the development of generally valid classification models, including minority classes, the coding of further processes is scheduled.

18 percent of the documents (126) were assigned to none of the mobility-related categories; those mainly focused on other requirements for public space (e.g. noise, accessibility, quality of stay). Such requirements will be included as additional dimensions in further scheme development.

8. Conclusion and Future Work

When public authorities consult the public, they have to ensure that all contributions are properly considered. In order to support this process that is vital to democratic participation yet costly in terms of resources, we have identified four classification tasks and introduced a new publicly-available German-language corpus.

Our corpus is the first German-language corpus in the domain of public participation that provides annotations of textual and GPS locations, as well as a thematic categorization for modes of transport. Furthermore, it provides annotations to distinguish argument components and their concreteness. In contrast to the previous datasets on argument mining for public participation, this corpus contains six different datasets varying in participation format (online platform vs. questionnaire) and issue. This enables the training of more transferable and robust machine learning methods.

Efforts to develop NLP models to solve the practical application tasks can now rely on this corpus. While it consists of mobility-related processes, its application is not limited to such issues as with the exception of thematic categorization, the tasks are generic to participation processes. The thematic categorization scheme is universally applicable in the mobility section.

Currently we are extending the annotation of the present corpus, as well as adding new datasets in order to increase diversity and representation of minority classes. What is more, we are working on expanding the thematic categorization scheme with additional dimensions (e.g. quality of public space, traffic safety or noise pollution). We have started to develop classification models for these four tasks based on the annotated corpus. A first model for the detection and classification of argument component detection has been introduced in Romberg and Conrad (2021). Our ultimate goal is to provide an open source application that supports public authorities in the evaluation of public participation contributions.

9. Acknowledgements

This publication is based on research in the project CIMT/Partizipationsnutzen, which is funded by the Federal Ministry of Education and Research of Germany as part of its Social-Ecological Research funding priority, funding no. 01UU1904. Responsibility for the content of this publication lies with the author.

10. Bibliographical References

Adams, B. and Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM'12)*, pages 375–378. AAAI Press.

Aitamurto, T., Chen, K., Cherif, A., Galli, J. S., and Santana, L. (2016). Civic CrowdAnalytics: Making sense of crowdsourced civic input with big data tools. In *Proceedings of the 20th International Academic Mindtrek Conference*, pages 86–94. Association for Computing Machinery.

Arana-Catania, M., Lier, F.-A. V., Procter, R., Tkachenko, N., He, Y., Zubiaga, A., and Liakata, M. (2021). Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3):1–22.

Arguello, J., Callan, J., and Shulman, S. (2008). Recognizing citations in public comments. *Journal of Information Technology & Politics*, 5(1):49–71.

Awamura, T., Kawahara, D., Aramaki, E., Shibata, T., and Kurohashi, S. (2015). Location name disambiguation exploiting spatial proximity and temporal consistency. In *Proceedings of the Third International Workshop on Natural Language Processing for Social Media*, pages 1–9. Association for Computational Linguistics.

Cardie, C., Farina, C., Rawding, M., and Aijaz, A. (2008). An eRulemaking corpus: Identifying substantive issues in public comments. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2757–2763. European Language Resources Association.

Der Senator für Umwelt, Bau und Verkehr. (2014). *Verkehrsentwicklungsplan Bremen 2025*. Bremen.

Eidelman, V. and Grom, B. (2019). Argument identification in public comments from eRulemaking. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 199–203. Association for Computing Machinery.

Fierro, C., Fuentes, C., Pérez, J., and Quezada, M. (2017). 200K+ crowdsourced political arguments for a new Chilean constitution. In *Proceedings of the 4th Workshop on Argument Mining*, pages 1–10. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. (2020). A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Gupta, S. and Nishu, K. (2020). Mapping local news coverage: Precise location extraction in textual news content using fine-tuned BERT based language model. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 155–162. Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2016a). What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223. Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2016b). Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics.

Kim, B., Yoo, M., Park, K. C., Lee, K. R., and Kim, J. H. (2021). A value of civic voices for smart city: A big data analysis of civic queries posed by Seoul citizens. *Cities*, 108:102941.

Konat, B., Lawrence, J., Park, J., Budzynska, K., and Reed, C. (2016). A corpus of argument networks: Using graph properties to analyse divisive issues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3899–3906. European Language Resources Association.

Krippendorff, K., Mathet, Y., Bouvry, S., and Widlöcher, A. (2016). On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Marsden, P.V. (ed.) Sociological Methodology*, 25:47–76.

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology*. Sage publications.

Kwon, N., Shulman, S. W., and Hovy, E. (2006). Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 international conference on Digital government research*, pages 157–166. Digital Government Society of North America.

Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Lawrence, J., Park, J., Budzynska, K., Cardie, C., Konat, B., and Reed, C. (2017). Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–22.

Liebeck, M., Esau, K., and Conrad, S. (2016). What to do with an airport? Mining arguments in the German online participation project Tempelhofer Feld. In *Proceedings of the Third Workshop on Argument Mining*, pages 144–153. Association for Computational Linguistics.

McNamee, P., Mayfield, J., Costello, C., Bishop, C., and Anderson, S. (2020). Tagging location phrases in text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4521–4528. European Language Resources Association.

Morio, G. and Fujita, K. (2018). Annotating online civic discussion threads for argument mining. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 546–553. IEEE.

Norris, P. (2011). *Democratic Deficit : Critical Citizens Revisited*. Cambridge University Press, Cambridge, GBR.

OECD. (2003). *Promise and Problems of E-Democracy*. OECD.

Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38. Association for Computational Linguistics.

Park, J. and Cardie, C. (2018). A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association.

Parry, G. and Moyser, G. (1994). More participation, more democracy? In David Beetham, editor, *Defining and Measuring Democracy*. Sage, London.

Purpura, S., Cardie, C., and Simons, J. (2008). Active learning for e-rulemaking: Public comment categorization. In *Proceedings of the 9th Annual International Digital Government Research Conference*, pages 34–243. Digital Government Society of North America.

Romberg, J. and Conrad, S. (2021). Citizen involvement in urban planning - how can municipalities be supported in evaluating public participation processes for mobility transitions? In *Proceedings of the 8th Workshop on Argument Mining*, pages 89–99. Association for Computational Linguistics.

Romberg, J. and Escher, T. (2020). Analyse der Anforderungen an eine Software zur (teil-) automatisierten Unterstützung bei der Auswertung von Beteiligungsverfahren. Working Paper 1, CIMT Research Group, Institute for Social Sciences, Heinrich Heine University Düsseldorf.

Rowe, G. and Frewer, L. J. (2004). Evaluating public-participation exercises: A research agenda. *Science, Technology, & Human Values*, 29(4):512–556.

Simonofski, A., Fink, J., and Burnay, C. (2021). Supporting policy-making with social media and e-participation platforms data: A policy analytics framework. *Government Information Quarterly*, 38(3):101590.

Skoumas, G., Pfoser, D., Kyrillidis, A., and Sellis, T. (2016). Location estimation using crowdsourced spatial relations. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 2(2):1–23.

Smith, D. A. and Crane, G. (2001). Disambiguating geographic names in a historical digital library. In Panos Constantopoulos et al., editors, *Research and Advanced Technology for Digital Libraries*, pages 127–136. Springer Berlin Heidelberg.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., and Slonim, N. (2019). Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5625–5635. Association for Computational Linguistics.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Wing, B. and Baldridge, J. (2014). Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 336–348. Association for Computational Linguistics.

Zebralog, (2020). *Integriertes Mobilitätskonzept der Stadt Krefeld: Dokumentation der zweiten Online-Beteiligungsphase (24.01. - 21.02.2020).* Berlin, Bonn.

11. Language Resource References

Romberg, J., Mark, L., and Escher, T., (2022a). *CIMT PartEval Corpus - Argument Components (Subcorpus).* ISLRN 484-558-142-596-7. https://github.com/juliaromberg/cimt-argument-mining-dataset.

Romberg, J., Mark, L., and Escher, T., (2022b). *CIMT PartEval Corpus - Argument Concreteness (Subcorpus).* ISLRN 776-577-161-062-9. https://github.com/juliaromberg/cimt-argument-concreteness-dataset.

Romberg, J., Mark, L., and Escher, T., (2022c). *CIMT PartEval Corpus - Geographic Location (Subcorpus).* ISLRN 951-974-499-316-4. https://github.com/juliaromberg/cimt-geographic-location-dataset.

Romberg, J., Mark, L., and Escher, T., (2022d). *CIMT PartEval Corpus - Thematic Categorization (Subcorpus).* ISLRN 441-856-914-941-8. https://github.com/juliaromberg/cimt-thematic-categorization-dataset.