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

Julia Romberg, Laura Mark, Tobias Escher
{julia.romberg, laura.mark, tobias.escher}@hhu.de

SPONSORED BY THE Federal Ministry of Education and Research
FONA Social-Ecological Research

Motivation

- 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 in order to inform decision-making, authorities often face challenges due to restricted resources [8, 1].
- Pre-structuring with supervised machine learning can make the evaluation process more efficient.
- We identified four common tasks in dialogue with practitioners whose support through automation would benefit the evaluation of participation processes:
  - **i) the detection of arguments**
  - **ii) assessing argument concreteness**
  - **iii) locating the contributions**
  - **iv) structuring according to topics**
- We introduce the **CIMT PartEval Corpus** including annotations for several thousand citizen contributions from six mobility-related planning processes in five municipalities in German language.

Overview of the CIMT PartEval Corpus

			datasets					
task	unit level	total units	CD_B	CD_C	CD_M	CQ_B	MC_K	MC_O
i) argument components [4]	sentences	17,852	10,442	1,704	2,193	1,505	2,008	
ii) argument concreteness [5]	sentence spans	1,127	679	92	110	55	191	
iii) geographic location [6]	token spans	4,830	4,087	743				
iv) thematic categorization [7]	documents	697						697

Information on the datasets, deriving from public participation processes in German municipalities:
CD_B, CD_E, CD_M: Map-based online dialogues for improving cycling in three cities, i.e. Bonn, Cologne (district Ehrenfeld), and Moers.
CQ_B: CD_B was supplemented with a representative survey of the population (questionnaires).
MC_K: The city of Krefeld held a map-based online dialogue for developing a mobility concept.
MC_O: The district office of Altona (city of Hamburg) implemented a map-based online dialogue as part of the transformation of the quarter Ottensen into a traffic-calmed neighborhood.

Subcorpus i) Argument Components

- Citizens communicate their ideas through arguments. In spatial planning processes, this usually involves describing a problem or condition, from which a proposition is derived.
- We focus on the identification of argumentative sentences and on the classification of two clausal properties using the following annotation scheme:

  Major positions (*mpos*): options for actions that are being proposed

  Premises: reasons that attack or support a major position or another premise

  Non-argumentative (*non-arg*): sentences without premise or major position

- Annotation process:
  - Ten percent of the dataset was coded by three coders, the remainder split equally.
  - Overall Fleiss' kappa agreement between 0.72 (CQ_B) and 0.80 (CD_C)
  - Coding inconsistencies were solved in a curation step by two process supervisors.
- Overview of the resulting annotation of argument component categories for the datasets:

		CD_B	CD_C	CD_M	MC_K	CQ_B	all
	total	10,442	1,704	2,193	2,008	1,505	17,852
arg	non-arg	1,153 (11.0%)	197 (11.6%)	382 (17.4%)	431 (21.5%)	172 (11.4%)	2,335
	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

- With multiple processes differing in format and process purpose, this subcorpus facilitates a comprehensive evaluation of machine learning methods including their potential for generalization [3].

Subcorpus ii) Argument Concreteness

- The less specific citizens' ideas are, the more difficult and time-consuming it will usually be for evaluators to derive measures for implementation.
- Automated evaluation of the concreteness of argument components can help practitioners filter out arguments that can be evaluated immediately.
- Our annotation scheme allows for a subdivision into three levels of concreteness:

  High: The *what*, *how* and *where* is specified (e.g. colour, surface, measurements).

  Intermediate: Some specifications are given (location, measures), but there is still room for interpretation.

  Low: No specific information (e.g. on location or description of measures) is contained.

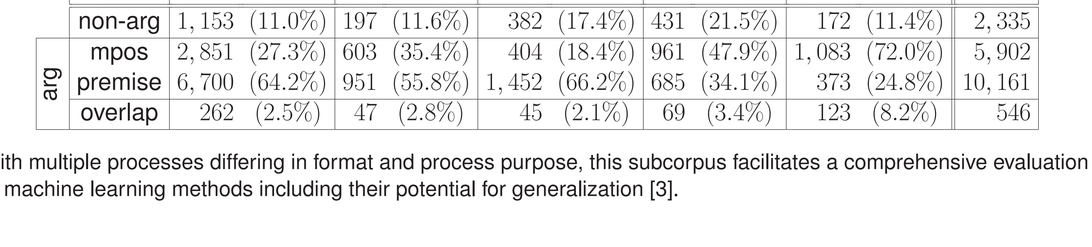
- Annotation process:
  - Coders first interrelated sentences of the same type (i.e. premises or major positions) to form units with coherent sense. Annotation supervisors resolved inconsistencies.

	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

Tab. 3: Units of interrelated argument components.

  - Coders then rated the resulting units' concreteness (five ratings per unit).

Example:



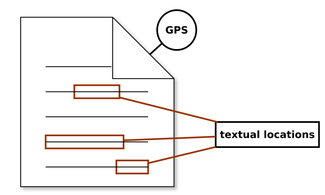
  - Due to the subjective nature, we dispense with a manual curation step in which an unambiguous assignment of concreteness to units is made.
  - Perception of concreteness is subjective but not arbitrary (Krippendorff's alpha$_w$ of 0.46).
- Analysis reveals that citizens clearly tend to write highly concrete arguments. Only twenty percent of the argument units have intermediate or low concreteness.

Subcorpus iii) Geographic Location

- The geographic location of citizens' contributions can be of great importance to the evaluation as it allows georeferencing of contributions and clustering of ideas by location.
- Depending on the data source, geo-references may not be available (e.g. paper questionnaires).
- To determine the geographic coordinates of a document's associated location by its textual content (*text-based geolocation*), we provide annotations of *textual locations* and *GPS coordinates*.

  Textual location: Sequence of words that refers to the spatial placement of a contribution (named entities or constructions with more fine-grained location information that can be unambiguously marked on a map).

  GPS coordinates: Spatial reference (point)



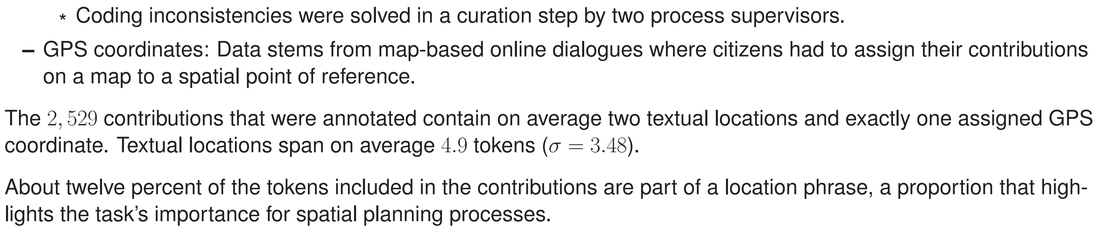
- Annotation process:
  - Textual locations:
    - Ten percent of the dataset was coded by three coders, the remainder split equally.
    - Krippendorff's alpha of 0.75
    - Coding inconsistencies were solved in a curation step by two process supervisors.
  - GPS coordinates: Data stems from map-based online dialogues where citizens had to assign their contributions on a map to a spatial point of reference.
- The 2,529 contributions that were annotated contain on average two textual locations and exactly one assigned GPS coordinate. Textual locations span on average 4.9 tokens ($\sigma = 3.48$).
- About twelve percent of the tokens included in the contributions are part of a location phrase, a proportion that highlights the task's importance for spatial planning processes.

Subcorpus iv) Thematic Categorization

- Thematic categorization of citizen contributions enables the joint analysis of contributions with similar topics and the delegation of contributions to the responsible administrative units.
- Most previous work fitted their categorization scheme to the participation processes under consideration. However, from-scratch development for new processes is impractical [2]:

  The time required for schema development and training data annotation may quickly exceed the effort of a purely manual analysis, especially for processes with fewer contributions.
- One solution is the development of universal categorization schemes.

  → We provide such a solution for transport-related processes.
- Annotation process:
  - All contributions were annotated by three coders.
  - Coding inconsistencies were solved in a curation step by two process supervisors, one an urban planner.
- Overview of the thematic categorization scheme for mobility (including category-wise Fleiss' kappa agreement and class support):



- Annotation efforts are ongoing to
  - provide a sufficient data basis for minority classes.
  - add further requirements for public space (e.g. noise, accessibility, quality of stay) to the scheme.

References and Acknowledgements

[1] M. Arana-Catania et al. "Citizen Participation and Machine Learning for a Better Democracy". In: *Digital Government: Research and Practice* 2.3 (2021), pp. 1–22.

[2] S. Purpura, C. Cardie, and J. Simons. "Active Learning for e-Rulemaking: Public Comment Categorization". In: *Proc. of the 9th Annual International Digital Government Research Conference*. Digital Government Society of North America, 2008, pp. 34–243.

[3] J. Romberg and S. Conrad. "Citizen Involvement in Urban Planning - How Can Municipalities Be Supported in Evaluating Public Participation Processes for Mobility Transitions?" In: *Proc. of the 8th WS on Argument Mining*. ACL, 2021, pp. 89–99.

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

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

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

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

[8] A. Simonofski, J. Fink, and C. Burnay. "Supporting policy-making with social media and e-participation platforms data: A policy analytics framework". In: *Government Information Quarterly* 38.3 (2021), p. 101590.

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.