

Data Statement for ding-01

1 HEADER

Dataset Title: ding-01

Dataset Curator(s):

- Jeongwoo Kang, CNRS, Laboratoire d’Informatique de Grenoble (Université Grenoble Alpes), Lead annotator
- Maria Boritchev, Télécom Paris, Co-annotator
- Maximin Coavoux, CNRS, Laboratoire d’Informatique de Grenoble (Université Grenoble Alpes), Co-annotator

Dataset Version: 0.1

Dataset Citation: [Kang et al. \(2025\)](#)

Data Statement Authors:

- Jeongwoo Kang, CNRS, Laboratoire d’Informatique de Grenoble (Université Grenoble Alpes), lead author
- Maria Boritchev, Télécom Paris,
- Maximin Coavoux, CNRS, Laboratoire d’Informatique de Grenoble (Université Grenoble Alpes),

Data Statement Version: 0.1

Data Statement DOI: <https://doi.org/10.5281/zenodo.15537426>

2 EXECUTIVE SUMMARY

The corpus is created to address the lack of French data for semantic analysis, which has been a significant barrier to developing French NLP applications.

We manually annotate a spontaneous spoken French corpus, DinG ([Boritchev and Amblard, 2022](#)), using the Abstract Meaning Representation formalism ([Banarescu et al., 2013](#), AMR). The current release of the dataset includes 1,830 utterances annotated in AMR.

3 CURATION RATIONALE

The lack of French data for semantic analysis motivated the creation of this corpus. Abstract Meaning Representation (AMR) is currently one of the most widely used semantic analysis frameworks. AMR represents the meaning of texts in a structured graph format. Such structured data is more machine-readable and therefore used in many NLP applications such as chatbot development, automatic text summarization, and human-robot interaction. However, large-scale data that is manually annotated is only available in English. This situation poses a significant barrier to the development of French NLP applications. We manually annotate a French corpus in AMR to bridge the gap.

We chose to annotate an existing French corpus *Dialogue in Games* ([Boritchev and Amblard, 2022](#), DinG) - a transcription of a multi-party dialogue of board game players. DinG was chosen for two reasons: DinG is available under a free license¹. As our goal is to make our data public, selecting open-source data is a crucial requirement. Second, DinG features natural spontaneous dialogues between speakers, capturing authentic conversational flow and a wide range of dialogic phenomena. More generally, the domain of spontaneous conversations is massively underrepresented in linguistic resources.

¹The *Attribution ShareAlike Creative Commons* (CC BY-SA 4.0) license.

4 DOCUMENTATION FOR SOURCE DATASETS

The original corpus was obtained from DinG (Boritchev and Amblard, 2022), which consists of manually transcribed multi-party dialogues among French-speaking players of the board game Catan.²

5 LANGUAGE VARIETIES

N/A. We refer the readers to the source dataset data-statement.

6 LANGUAGE USER DEMOGRAPHIC

N/A. We refer the readers to the source dataset data statement.

7 ANNOTATOR DEMOGRAPHIC

Lead annotator

- Gender: Female
- Socioeconomic status: PhD in Natural Language Processing (NLP)
- First language(s): Korean
- Proficiency in the language(s) of the data being annotated:
 - French: DALF C1³/ Full Professional Proficiency
- Relevant training: PhD research in AMR parsing

Co-annotator 1

- Socioeconomic status: tenured researcher
- First language(s): French
- Proficiency in the language(s) of the data being annotated: first language

²For more information about the game, we guide the readers to <https://catanuniverse.com/en>.

³DALF (*Diplôme approfondi de langue française*) is a diploma given by the French Ministry of Education to certify the level of French-language skills of non-French speakers.

Co-annotator 2

- Gender: Female
- Socioeconomic status: associate professor
- First language(s): Russian, French
- Proficiency in the language(s) of the data being annotated: first language
- Relevant training: research in AMR parsing

8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS

N/A. We refer the readers to the source dataset data statement.

9 PREPROCESSING AND DATA FORMATTING

The dataset format follows AMR 3.0 (Knight et al., 2020), meaning that example IDs and input texts are prefixed with corresponding tags (`# ::id` and `# ::sent`), and their corresponding AMR graph is written in PENMAN format (Kasper, 1989). As an annotation tool, we used metAMoRphosED, an open-source AMR editor (Heinecke, 2023).

10 CAPTURE QUALITY

N/A. We refer the readers to the source dataset data statement.

11 LIMITATIONS

Although two additional co-annotators cross-checked the quality of the dataset, the corpus was primarily annotated by a single lead annotator. Therefore, the co-annotators may have been influenced to agree with the existing annotation during the cross validation.

In addition, the original corpus includes some inaudible parts. When there are inaudible parts, its annotation in AMR can be incomplete due to a lack of connections between arguments.

Finally, ambiguity in the transcription and the lack of prosodic information (e.g., pauses indicating sentence boundaries) could have led to misinterpretations of the speaker’s intent. Without access to this additional data, we had to choose one interpretation over others, potentially misrepresenting the original meaning conveyed through tone, emphasis, or pacing.

12 METADATA

Metadata of `ding-01` (e.g., annotation process, inter-annotator agreement score) is provided by Kang et al. (2025).

13 DISCLOSURES AND ETHICAL REVIEW

This work was funded by the Carnot Cognition Institute (ANAGRAM project) and the French National Research Agency, via the SynPaX project (ANR-23-CE23-0017-01).

14 DISTRIBUTION

- This dataset is licensed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license. It is available for download at <https://doi.org/10.5281/zenodo.15537426>.
- DOI of the dataset: 10.5281/zenodo.15537426
- Date(s) of distribution of the dataset: 28/05/2025

15 MAINTENANCE

- This dataset is a work in progress. Errors are regularly reviewed and corrected by the author through July 2025. If users identify any issues, they are encouraged to report them to the dataset co-managers via email: maria.boritchev@telecom-paris.fr, maximin.coavoux@univ-grenoble-alpes.fr.
- While the dataset is under active development, updates are made frequently—either daily or

weekly. A summary of these updates will be published monthly on the publication URL.

- Previous versions of the dataset are archived and accessible on the publication URL.

16 OTHER

N/A

17 GLOSSARY

- **AMR**: Abstract Meaning Representation

About this document

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 3 Schema. The template was prepared by Angelina McMillan-Major and Emily M. Bender and can be found at <http://techpolicylab.uw.edu/data-statements>.

References

- Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for sembanking. In Pareja-Lora, A., Liakata, M., and Dipper, S., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Boritchev, M. and Amblard, M. (2022). A multi-party dialogue resource in French. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis,

- S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 814–823, Marseille, France. European Language Resources Association.
- Heinecke, J. (2023). metAMoRphosED: a graphical editor for Abstract Meaning Representation. In *19th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Nancy.
- Kang, J., Boritchev, M., and Coavoux, M. (2025). ding-01 :ARG0 un corpus AMR pour le français parlé spontané. In *Actes de la 32ème Conférence sur le Traitement Automatique des Langues Naturelles*, Marseille, France. ATALA.
- Kasper, R. T. (1989). A flexible interface for linking applications to Penman’s sentence generator. In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.
- Knight, K., Badarau, B., Baranescu, L., Bonial, C., Bardocz, M., Griffitt, K., Hermjakob, U., Marcu, D., Palmer, M., O’Gorman, T., and Schneider, N. (2020). Abstract meaning representation (amr) annotation release 3.0 ldc2020t02. Philadelphia: Linguistic Data Consortium.